

# Introduction à l'analyse numérique

## Notes du Cours LM 334

Albert Cohen

2 novembre 2009

Bien avant l'introduction des ordinateurs, mathématiciens et ingénieurs se sont posés la question du *calcul approché* dans différents contextes. Il peut s'agir d'évaluer la solution d'un système d'équations, d'approcher le graphe d'une fonction connue à partir de ses valeurs en des points, ou d'une fonction inconnue qui est solution d'une équation différentielle.

*L'analyse numérique* est la branche des mathématiques qui étudie les méthodes permettant de résoudre ces problèmes et analyse leur performance. Ses développements récents sont intimement liés à ceux des moyens de calcul offerts par l'informatique.

L'objectif de ce cours est de familiariser les étudiants avec quelques notions simples d'analyse numérique, et de les préparer à des cours plus avancés dans ce domaine. En particulier, les cours de deuxième semestre "Calcul numérique matriciel" et "Méthodes numériques pour les équations différentielles" développent de façon très détaillée les notions qui sont introduites dans les parties 1 et 5 de ce cours.

Ces notes contiennent la totalité des résultats du cours exposés en Amphi. Les démonstrations de certains résultats - les plus simples - sont omises, et c'est un excellent exercice que d'essayer de les refaire uniquement à partir des indications données dans les notes.

**Avertissement :** ces notes sont régulièrement mises à jour et corrigées. Toutes les remarques ou questions permettant d'en améliorer la rédaction peuvent être envoyées à l'adresse [cohen@ann.jussieu.fr](mailto:cohen@ann.jussieu.fr).

# 1 Notions de calcul matriciel

## 1.1 Quelques rappels

A toutes fins utiles, nous rappelons dans cette section quelques notions d'algèbre linéaire qu'il est indispensable de connaître pour l'étudiant de L3. Dans toute la suite, on désigne par  $\mathbb{K}$  un corps commutatif qui est soit celui des nombres réels noté  $\mathbb{R}$ , soit celui des nombres complexes noté  $\mathbb{C}$ . On rappelle qu'un *espace vectoriel*  $E$  sur  $\mathbb{K}$  est muni d'une loi d'addition interne

$$(x, y) \in E \times E \mapsto x + y \in E$$

tel que  $(E, +)$  est un groupe commutatif et d'une loi de multiplication externe

$$(\lambda, x) \in \mathbb{K} \times E \mapsto \lambda x \in E,$$

qui vérifie les propriétés

$$(\lambda + \mu)x = \lambda x + \mu x, \lambda(x + y) = \lambda x + \lambda y, \lambda(\mu x) = (\lambda\mu)x \text{ et } 1x = x,$$

pour tout  $\lambda, \mu \in \mathbb{K}$  et  $x, y \in E$ .

Les éléments de  $E$  et de  $\mathbb{K}$  sont respectivement appelés *vecteurs* et *scalaires*. Un sous ensemble  $F \subset E$  est un sous-espace vectoriel de  $E$  si et seulement si il est stable par les lois d'addition et de multiplication externe, c'est à dire que pour tout  $\lambda, \mu \in \mathbb{K}$  et  $x, y \in F$  on a  $\lambda x + \mu y \in F$ .

Une famille de vecteurs  $(e_1, \dots, e_n)$  d'un espace vectoriel  $E$  est dite *génératrice* si et seulement si tout vecteur  $x \in E$  est une combinaison linéaire des vecteurs de cette famille : il existe  $(x_1, \dots, x_n) \in \mathbb{K}^n$  tels que

$$x = \sum_{i=1}^n x_i e_i.$$

La famille  $(e_1, \dots, e_n)$  est dite *libre* ou linéairement indépendante si et seulement si

$$\sum_{i=1}^n x_i e_i = 0 \Rightarrow x_1 = \dots = x_n = 0.$$

Une famille non-libre est dite liée ou linéairement dépendante. La famille  $(e_1, \dots, e_n)$  est une *base* si et seulement si elle est libre et génératrice. Dans ce cas pour tout  $x \in E$ , il existe un unique  $n$ -uplet  $(x_1, \dots, x_n) \in \mathbb{K}^n$  tels que  $x = \sum_{i=1}^n x_i e_i$ . Les  $x_i$  sont les coordonnées de  $x$  dans la base  $(e_1, \dots, e_n)$ .

L'espace  $E$  est de dimension finie si et seulement si il existe une base  $(e_1, \dots, e_n)$  de  $E$ . On montre alors que toute base de  $E$  comporte exactement  $n$  vecteurs et on dit que  $n = \dim(E)$  est la dimension de  $E$ . Par exemple les espaces  $\mathbb{R}^n$  et  $\mathbb{C}^n$  sont des espaces vectoriels de dimension  $n$  sur  $\mathbb{R}$  et  $\mathbb{C}$  respectivement. La base dite *canonique* pour ces espaces est  $(e_1, \dots, e_n)$  où  $e_i$  est le vecteur  $(0, \dots, 0, 1, 0, \dots, 0)$  avec 1 en  $i$ -ème position. L'ensemble des polynômes de degré  $n$  à coefficients réels est un espace de dimension  $n + 1$  dont une base est donnée par les polynômes  $x \mapsto x^k$  pour  $k = 0, \dots, n$ . Il existe des espaces de dimension infinie, par exemple l'espace des polynômes de degré quelconques.

Les espaces  $\mathbb{K}^n$  sont munis d'un produit scalaire (appelé aussi produit hermitien si  $\mathbb{K} = \mathbb{C}$ ) : pour  $u = (u_1, \dots, u_n)$  et  $v = (v_1, \dots, v_n)$ , on pose

$$u \cdot v = \langle u, v \rangle := \sum_{i=1}^n u_i \bar{v}_i.$$

A ce produit scalaire est associée la norme dite euclidienne si  $\mathbb{K} = \mathbb{R}$  ou hermitienne si  $\mathbb{K} = \mathbb{C}$  :

$$\|u\| := \sqrt{\langle u, u \rangle} = \sqrt{\sum_{i=1}^n |u_i|^2},$$

où  $|x|$  désigne le module de  $x$  si  $x \in \mathbb{C}$  et sa valeur absolue si  $x \in \mathbb{R}$ . L'espace  $\mathbb{K}^n$  est complet pour cette norme. On parle d'espace euclidien si  $\mathbb{K} = \mathbb{R}$  et hermitien si  $\mathbb{K} = \mathbb{C}$ .

Si  $E$  et  $F$  sont des espaces vectoriels sur le même corps  $\mathbb{K}$ , une application  $L : E \rightarrow F$  est dite linéaire si et seulement si elle vérifie

$$L(x + y) = L(x) + L(y) \text{ et } L(\lambda x) = \lambda L(x),$$

pour tout  $x, y \in E$  et  $\lambda \in \mathbb{K}$ . L'ensemble des applications linéaires de  $E$  dans  $F$  est noté  $\mathcal{L}(E, F)$  et constitue lui-même un espace vectoriel. Lorsque  $E = F$  on note cet espace  $\mathcal{L}(E)$  et on dit que  $L \in \mathcal{L}(E)$  est un endomorphisme de  $E$ . Le *noyau* et l'*image* de  $L \in \mathcal{L}(E, F)$  sont les sous espaces vectoriels de  $E$  et  $F$  définis par

$$\text{Ker}(L) := \{x \in E ; L(x) = 0\} \text{ et } \text{Im}(L) := \{y = L(x) ; x \in E\},$$

et le rang de  $L$  est défini par  $\text{rg}(L) := \dim(\text{Im}(L))$ . Le théorème du rang affirme que si  $E$  est de dimension finie, on a

$$\text{rg}(L) + \dim(\text{Ker}(L)) = \dim(E),$$

On rappelle que  $L$  est surjective si et seulement si  $\text{Im}(L) = F$ , ce qui équivaut à  $\text{rg}(L) = \dim(F)$  lorsque  $F$  est de dimension finie, et que  $L$  injective si et seulement si  $\text{Ker}(L) = \{0\}$ , ce qui équivaut à  $\text{rg}(L) = \dim(E)$  lorsque  $E$  est de dimension finie. Une application linéaire bijective est appelée isomorphisme, on a dans ce cas nécessairement  $\dim(E) = \dim(F)$ . L'ensemble des isomorphismes de  $E$  dans lui-même muni de la relation de composition est un groupe. L'élément neutre de ce groupe est l'application identité.

Si  $L \in \mathcal{L}(E, F)$  et si  $(e_1, \dots, e_n)$  et  $(f_1, \dots, f_m)$  sont des bases de  $E$  et de  $F$ , on peut représenter  $L$  par la *matrice*  $m \times n$

$$A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{pmatrix},$$

dont la  $j$ -ème colonne est le vecteur des coordonnées de  $L(e_j)$  dans la base  $(f_1, \dots, f_m)$  :

$$L(e_j) = \sum_{i=1}^m a_{i,j} f_i.$$

Pour tout  $x = \sum_{j=1}^n x_j e_j \in E$ , l'image  $y = L(x)$  s'écrit alors  $y = \sum_{i=1}^m y_i f_i$  avec  $y_i = \sum_{j=1}^n a_{i,j} x_j$ , c'est à dire

$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = A \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

On dit que  $A$  est la matrice - ou la représentation matricielle - de  $L$  dans les bases  $(e_1, \dots, e_n)$  et  $(f_1, \dots, f_m)$ . Dans le cas où  $E = F$  et  $e_i = f_i$  on dit que  $A$  est la matrice de  $L$  dans la base  $(e_1, \dots, e_n)$ . Notons que  $A$  est en particulier la matrice de l'application  $x \mapsto Ax$  dans la base canonique de  $\mathbb{C}^n$ . On rappelle que  $\text{Im}(A)$ ,  $\text{Ker}(A)$  et  $\text{rg}(A)$  désignent l'image, le noyau et le rang de l'application linéaire  $x \mapsto Ax$ .

L'ensemble des matrices  $m \times n$  à coefficients dans  $\mathbb{K}$  est noté  $\mathcal{M}_{m,n}(\mathbb{K})$ . C'est un espace vectoriel sur  $\mathbb{K}$  de dimension  $mn$ . L'ensemble des *matrices carrées*  $n \times n$  est noté  $\mathcal{M}_n(\mathbb{K})$ . Une matrice carrée  $A = (a_{i,j})$  est dite triangulaire supérieure si et seulement si  $a_{i,j} = 0$  si  $j < i$ , triangulaire inférieure si et seulement si  $a_{i,j} = 0$  si  $i < j$ , diagonale si et seulement si  $a_{i,j} = 0$  si  $i \neq j$  auquel cas on la note parfois  $A = \text{diag}(a_{1,1}, \dots, a_{n,n})$ . La matrice identité est  $I = \text{diag}(1, \dots, 1)$ . La *trace* d'une matrice carrée  $A = (a_{i,j})$  est la quantité

$$\text{tr}(A) = \sum_{i=1}^n a_{i,i}.$$

La notion de produit matriciel est liée à celle composition des applications linéaires : si  $L \in \mathcal{L}(E, F)$  a pour matrice  $A = (a_{i,j}) \in \mathcal{M}_{m,n}$  dans les bases  $(e_1, \dots, e_n)$  et  $(f_1, \dots, f_m)$  et si  $U \in \mathcal{L}(F, G)$  a pour matrice  $B = (b_{i,j}) \in \mathcal{M}_{p,m}$  dans les bases  $(f_1, \dots, f_m)$  et  $(g_1, \dots, g_p)$ , alors l'application  $U \circ L \in \mathcal{L}(E, G)$

a pour matrice  $C = (c_{i,j}) \in \mathcal{M}_{p,n}$  dans les bases  $(e_1, \dots, e_n)$  et  $(g_1, \dots, g_p)$ , où  $C = BA$  est le produit matriciel défini par

$$c_{i,j} = \sum_{k=1, \dots, m} b_{i,k} a_{k,j}.$$

Les matrices transposée et adjointe de  $A = (a_{i,j}) \in \mathcal{M}_{m,n}(\mathbb{K})$  sont les matrices  $A^t = (a_{i,j}^t)$  et  $A^* = (a_{i,j}^*)$  de  $\mathcal{M}_{n,m}(\mathbb{K})$  définies par  $a_{i,j}^t = a_{j,i}$  et  $a_{i,j}^* = \overline{a_{j,i}}$ . Ces deux notions sont les mêmes lorsque  $\mathbb{K} = \mathbb{R}$ . On a pour tout  $x \in \mathbb{K}^m$  et  $y \in \mathbb{K}^n$  la relation

$$\langle A^* x, y \rangle = \langle x, Ay \rangle.$$

Une propriété importante est que  $\text{Ker}(A^*)$  est le supplémentaire orthogonal de  $\text{Im}(A)$  pour les produit scalaires défini ci-dessus.

Une matrice carrée  $A \in \mathcal{M}_n(\mathbb{K})$  est inversible si et seulement si il existe  $B \in \mathcal{M}_n(\mathbb{K})$  tel que  $AB = BA = I$ . La matrice  $B$  est l'inverse de  $A$  notée  $A^{-1}$ . Ceci équivaut à  $\text{Im}(A) = \mathbb{K}^n$  c'est à dire  $\text{rg}(A) = n$ , ainsi qu'à  $\text{Ker}(A) = \{0\}$ . Les matrices inversibles sont les matrices qui représentent les isomorphismes. L'ensemble des matrices inversibles  $n \times n$  muni de la loi du produit est un groupe dont l'élément neutre est  $I$ . Il est appelé groupe linéaire et noté  $GL_n(\mathbb{K})$ .

On peut étudier l'inversibilité d'une matrice carrée  $A$  par son déterminant noté  $\det(A)$ . On rappelle que l'application de  $\mathcal{M}_n(\mathbb{K})$  dans  $\mathbb{K}$  qui associe  $\det(A)$  à  $A$  est caractérisée par les trois propriétés suivantes : (i) multilinéaire par rapports aux vecteurs colonnes  $(a_1, \dots, a_n)$  de  $A$ , c'est à dire linéaire par rapport à la colonne  $a_j$  lorsque l'on fixe les autres, (ii) antisymétrique c'est à dire change de signe par échange de deux colonnes, et (iii)  $\det(I) = 1$ . Ces propriétés permettent de calculer le déterminant de matrices quelconques en se ramenant à celui d'une matrice triangulaire qui vaut le produit de ses éléments diagonaux. On peut aussi utiliser le développement du déterminant par rapport aux éléments d'une ligne ou d'une colonne. On rappelle que  $A$  est inversible si et seulement si  $\det(A) \neq 0$  ainsi que les propriétés

$$\det(AB) = \det A \det B, \det(A^{-1}) = (\det(A))^{-1} \text{ et } \det(A^*) = \overline{\det(A)}.$$

Si  $(e_1, \dots, e_n)$  et  $(f_1, \dots, f_n)$  sont deux bases d'un même espace  $E$ , on leur associe la matrice de passage ou de changement de base  $P = (p_{i,j})$  de la première vers la deuxième base, dont les vecteurs colonnes sont les coordonnées des vecteurs de la deuxième base dans la première :

$$f_j = \sum_{i=1}^n p_{i,j} e_i.$$

C'est une matrice inversible et réciproquement toute matrice inversible peut-être vue comme une matrice de changement de base. Si on applique  $P$  au vecteur de coordonnées de  $x \in E$  dans la base  $(f_1, \dots, f_n)$  on obtient le vecteur de coordonnées de  $x$  dans la base  $(e_1, \dots, e_n)$ . Si  $A$  est la matrice de  $L \in \mathcal{L}(E)$  dans la base  $(e_1, \dots, e_n)$ , alors

$$B = P^{-1}AP$$

est la matrice de  $L$  dans la base  $(f_1, \dots, f_n)$ . Deux matrices vérifiant une telle identité pour une matrice inversible  $P$  sont dites *semblables*. On montre que deux matrices semblables ont même déterminant et même trace.

## 1.2 Réduction des matrices

Dans tout ce qui suit, on considère uniquement des matrices carrées. La réduction d'une matrice  $A$  consiste à rechercher une matrice  $B$  semblable à  $A$  et qui est diagonale ou triangulaire. Rappelons tout d'abord la notion de *valeur propre* d'une matrice.

**Définition 1.2.1** Soit  $A \in \mathcal{M}_n(\mathbb{K})$ . On dit que  $\lambda \in \mathbb{K}$  est une valeur propre de  $A$  si et seulement si il existe un vecteur  $x \in \mathbb{K}^n$  non-nul tel que  $Ax = \lambda x$ . On dit que  $x$  est vecteur propre de  $A$  pour la valeur propre  $\lambda$ .

L'ensemble des vecteurs propres de  $A$  pour la valeur propre  $\lambda$  est l'espace vectoriel

$$E_\lambda := \text{Ker}(A - \lambda I).$$

Il est appelé espace propre pour la valeur  $\lambda$ . Afin d'identifier les valeurs propres d'une matrice  $A \in \mathcal{M}_n(\mathbb{K})$ , on introduit son *polynôme caractéristique* défini par

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

En développant le déterminant par rapport aux éléments d'une colonne, il est aisé de montrer que  $P$  est un polynôme de degré  $n$ , à coefficients dans  $\mathbb{K}$ . Les racines de  $P$  sont les  $\lambda$  tels que  $A - \lambda I$  n'est pas inversible, c'est à dire précisément les valeurs propres de  $A$ .

On en déduit que  $A$  admet au plus  $n$  valeurs propres. On remarque que si  $A \in \mathcal{M}_n(\mathbb{R})$ , le polynôme  $P$  à coefficients réels peut admettre des racines complexes, ce qui signifie que  $A$  vue comme une matrice de  $\mathcal{M}_n(\mathbb{C})$  - qui contient  $\mathcal{M}_n(\mathbb{R})$  - admet des valeurs propres et vecteurs propres complexes. Lorsque  $\lambda_0$  est une racine multiple de  $P$  c'est à dire  $(\lambda - \lambda_0)^k$  se factorise dans  $P(\lambda)$ , on dit que  $\lambda_0$  est une valeur propre de multiplicité  $k$ .

Une propriété importante est l'invariance du polynôme caractéristique - et donc des valeurs propres - par changement de base :

$$\det(P^{-1}AP - \lambda I) = \det(P^{-1}(A - \lambda I)P) = \det(A - \lambda I).$$

Si  $\{\lambda_1, \dots, \lambda_p\}$  sont les valeurs propres distinctes de  $A$ , les espaces propres  $E_{\lambda_i}$  ont la propriété de somme directe : pour tout vecteurs  $u_i \in E_{\lambda_i}$

$$\sum_{i=1}^p u_i = 0 \Rightarrow u_1 = \dots = u_p = 0.$$

Cette propriété se démontre aisément par récurrence sur  $p$  : elle est triviale pour  $p = 1$  et pour  $p > 1$  l'égalité  $\sum_{i=1}^p u_i = 0$  entraîne  $\sum_{i=1}^p \lambda_i u_i = 0$ . En multipliant la première identité par  $\lambda_p$  et en faisant la différence avec la seconde, on obtient ainsi

$$\sum_{i=1}^{p-1} (\lambda_i - \lambda_p) u_i = 0,$$

et l'hypothèse de récurrence permet de conclure.

**Définition 1.2.2** Une matrice  $A \in \mathcal{M}_n(\mathbb{K})$  est dite *triangulable* (respectivement *diagonalisable*) si et seulement si il existe une matrice inversible  $P \in GL_n(\mathbb{K})$  et une matrice triangulaire supérieure  $T$  (respectivement diagonale  $D$ ) de  $\mathcal{M}_n(\mathbb{K})$  telles que

$$A = PTP^{-1} \text{ (respectivement } A = PDP^{-1}\text{)}.$$

Autrement dit, la matrice représentant l'application  $x \mapsto Ax$  dans la base des vecteurs colonnes de  $P$  est triangulaire ou diagonale. On remarque que si  $(\lambda_1, \dots, \lambda_n)$  sont les éléments diagonaux de  $T$  ou de  $D$ , le polynôme caractéristique de  $A$  qui est le même que celui de  $T$  ou  $D$  est alors donné par

$$P(\lambda) = \prod_{i=1}^n (\lambda - \lambda_i).$$

Les valeurs propres de  $A$  sont donc exactement les  $(\lambda_1, \dots, \lambda_n)$ . On remarque aussi que dans le cas où  $A$  est diagonalisable les colonnes de  $P$  forment alors une base de vecteurs propres de  $A$ .

**Théorème 1.2.1** Toute matrice  $A \in \mathcal{M}_n(\mathbb{C})$  est triangulable.

**Preuve :** on effectue une récurrence sur  $n$ , le résultat étant trivial en dimension  $n = 1$ . On le suppose vrai à l'ordre  $n - 1$ . Si  $A \in \mathcal{M}_n(\mathbb{C})$ , son polynôme caractéristique admet au moins une racine  $\lambda_1 \in \mathbb{C}$ , et il existe donc un vecteur  $e_1 \neq 0$  tel que  $Ae_1 = \lambda_1 e_1$ . On complète  $e_1$  par des vecteurs  $(e_2, \dots, e_n)$  pour obtenir une base de  $\mathbb{C}^n$ . Si on introduit la matrice de passage  $P_1$  de la base canonique dans la base  $(e_1, \dots, e_n)$ , la représentation de  $A$  dans cette base est donc de la forme

$$P_1^{-1}AP_1 = \begin{pmatrix} \lambda_1 & \alpha_2 & \dots & \alpha_n \\ 0 & & & \\ \vdots & & B & \\ 0 & & & \end{pmatrix},$$

où  $B \in \mathcal{M}_{n-1}(\mathbb{C})$ . Par application de l'hypothèse de récurrence, il existe une matrice inversible  $P_2$  de taille  $n - 1$  telle que  $P_2^{-1}BP_2 = T_2$  où  $T_2$  est une matrice triangulaire supérieure d'ordre  $n - 1$ . La matrice  $P_2$  est une matrice de changement de base dans  $\mathbb{C}^{n-1}$ . On construit ainsi une matrice de passage  $P_3$  entre la base  $(e_1, e_2, \dots, e_n)$  et une nouvelle base  $(e_1, f_2, \dots, f_n)$

$$P_3 = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & P_2 & \\ 0 & & & \end{pmatrix}.$$

La matrice de passage de la base canonique dans la base  $(e_1, f_2, \dots, f_n)$  est  $P = P_1P_3$  et la représentation de  $A$  dans cette base est donc de la forme

$$P^{-1}AP = \begin{pmatrix} \lambda_1 & \beta_2 & \dots & \beta_n \\ 0 & & & \\ \vdots & & P_2^{-1}BP_2 & \\ 0 & & & \end{pmatrix} = \begin{pmatrix} \lambda_1 & \beta_2 & \dots & \beta_n \\ 0 & & & \\ \vdots & & T_2 & \\ 0 & & & \end{pmatrix} = T,$$

où  $T$  est une matrice triangulaire supérieure. □

**Remarque 1.2.1** Si  $A$  est réelle, le résultat s'applique, mais  $T$  et  $P$  peuvent être complexes. Par exemple, pour la matrice

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

les valeurs propres et les vecteurs propres sont complexes.

**Définition 1.2.3** Une matrice  $U \in \mathcal{M}_n(\mathbb{K})$  est dite unitaire si et seulement si  $U^*U = I$  c'est à dire  $U^* = U^{-1}$ .

De cette définition, il découle qu'une matrice unitaire préserve le produit scalaire : pour tout  $x, y \in \mathbb{K}^n$  on a  $\langle Ux, Uy \rangle = \langle x, y \rangle$ . On en déduit que les vecteurs colonnes de  $U$  forment une base orthonormale de  $\mathbb{K}^n$  puisqu'ils sont les images de la base canonique qui est orthonormale. Le résultat suivant affirme qu'on peut trianguler une matrice complexe dans une base orthonormale.

**Proposition 1.2.1** Pour tout  $A \in \mathcal{M}_n(\mathbb{C})$  il existe une matrice unitaire  $U$  et une matrice triangulaire supérieure  $S$  telle que  $A = USU^*$ , c'est à dire  $S = U^{-1}AU$ .

**Preuve :** D'après le théorème 1.2.1 on sait déjà qu'il existe  $P$  inversible et  $T$  triangulaire supérieure telle que  $T = P^{-1}AP$ . En notant  $t_{i,j}$  les coefficients de  $T$  et  $(e_1, \dots, e_n)$  la base constituée par les vecteurs colonnes de  $P$  cela signifie que

$$Ae_j = \sum_{i \leq j} t_{i,j}e_i.$$

Appliquons à présent le procédé d'orthogonalisation de Gram-Schmidt à  $(e_1, \dots, e_n)$  en définissant par récurrence

$$f_1 := \frac{e_1}{\|e_1\|} \quad f_j := \frac{e_j - P_{j-1}e_j}{\|e_j - P_{j-1}e_j\|},$$

où  $P_{j-1}x := \sum_{k=1}^{j-1} \langle x, f_k \rangle f_k$  est la projection orthogonale de  $x$  sur l'espace engendré par  $(f_1, \dots, f_{j-1})$ , qui est aussi celui engendré par  $(e_1, \dots, e_{j-1})$ . Par ce procédé  $f_i$  est une combinaison linéaire des  $e_j$  pour  $j \leq i$  et  $e_i$  est une combinaison linéaire des  $f_j$  pour  $j \leq i$ . On en déduit que  $Af_j$  est une combinaison linéaire des  $f_i$  pour  $i \leq j$ , c'est à dire

$$Af_j = \sum_{i \leq j} s_{i,j} f_i.$$

En posant  $s_{i,j} = 0$  pour  $j < i$  on a donc  $U^{-1}AU = S$  où  $S$  est triangulaire supérieure et  $U$  est la matrice de passage de la base canonique à la base  $(f_1, \dots, f_n)$ , qui est unitaire puisque cette base est orthonormale.  $\square$

Intéressons nous à présent à la diagonalisation des matrices. Contrairement à la triangulation, toutes les matrices ne sont pas diagonalisables, et il n'existe pas de caractérisation simple des matrices qui le sont. Nous introduisons ci-dessous une classe importante de matrices diagonalisables.

**Définition 1.2.4** Une matrice  $A \in \mathcal{M}_n(\mathbb{C})$  est dite normale si et seulement si elle commute avec son adjoint, c'est à dire  $A^*A = AA^*$ .

En particulier, les matrices unitaires sont normales. Un autre cas particulier de matrices normales sont celles qui vérifient  $A^* = A$  et qui sont dites *auto-adjointes* ou *hermitiennes*. Il s'agit des matrices réelles symétriques dans le cas  $\mathbb{K} = \mathbb{R}$ .

**Théorème 1.2.2** Une matrice  $A \in \mathcal{M}_n(\mathbb{C})$  est normale si et seulement s'il elle est diagonalisable dans une base orthonormale de vecteur propres.

**Preuve :** Il est clair qu'une matrice  $A = UDU^*$ , avec  $U$  unitaire et  $D$  diagonale est normale. Réciproquement, on sait déjà par la Proposition 1.2.1 que  $A$  est triangulable dans une base orthonormée. Il existe donc  $U$  unitaire telle que  $A = UTU^*$ . Or  $AA^* = A^*A$  implique que  $TT^* = T^*T$ , ce qui montre que  $T$  est normale. On termine la démonstration en montrant que toute matrice, à la fois triangulaire et normale est diagonale. Soit donc  $T$  une matrice triangulaire (supérieure) et normale. Puisque  $T = (t_{i,j})_{1 \leq i,j \leq n}$  est triangulaire supérieure, on a  $t_{i,j} = 0$  si  $i > j$ . On en déduit, en identifiant l'élément en première ligne et première colonne du produit  $T^*T = TT^*$ , que

$$|t_{1,1}|^2 = \sum_{k=1}^n |t_{1,k}|^2,$$

et donc  $t_{1k} = 0$  pour tout  $1 < k \leq n$ , c'est-à-dire que la première ligne de  $T$  ne contient que des zéros, excepté le coefficient diagonal. Par récurrence, on suppose que les  $(i-1)$  premières lignes de  $T$  n'ont que des zéros, exceptés les coefficients diagonaux. En identifiant l'élément en  $i$ -ème ligne et  $i$ -ème colonne du produit  $T^*T = TT^*$ , on obtient

$$|t_{i,i}|^2 = \sum_{k=i}^n |t_{i,k}|^2,$$

et donc  $t_{i,k} = 0$  pour tout  $i < k \leq n$ , c'est-à-dire que la  $i$ -ème ligne de  $T$  n'a aussi que des zéros hors la diagonale. Donc  $T$  est diagonale.  $\square$

**Théorème 1.2.3** Une matrice  $A$  est auto-adjointe si et seulement si, elle est diagonalisable dans une base orthonormée avec des valeurs propres réelles

**Preuve :** Si  $A = UDU^{-1}$  avec  $D$  diagonale et réelle et  $U$  est unitaire, il est évident que  $A = A^*$ . Réciproquement, si  $A = A^*$ , elle est normale et on sait déjà qu'elle est diagonalisable dans une base orthonormée de vecteurs propres. Si  $\lambda$  est une de ces valeurs propre et  $x \neq 0$  un vecteur tel que  $Ax = \lambda x$ , on a

$$\lambda \|x\|^2 = \langle Ax, x \rangle = \langle A^*x, x \rangle = \langle x, Ax \rangle = \bar{\lambda} \|x\|^2,$$

ce qui montre que  $\lambda \in \mathbb{R}$ .  $\square$

**Remarque 1.2.2** On peut améliorer le théorème précédent dans le cas d'une matrice symétrique réelle (cas particulier de matrice auto-adjointe) en affirmant que la matrice  $U$  est aussi réelle : il suffit pour cela de reprendre le raisonnement de la proposition 1.2.1, en montrant d'abord que puisque les valeurs propres sont réelles on peut partir d'une base  $e_n$  de vecteurs propres réels, puis on orthonormalise cette base par le procédé de Gram-Schmidt et on aboutit ainsi à une matrice  $U$  unitaire et réelle.

**Remarque 1.2.3** A toute matrice réelle symétrique  $A$  est associée la forme quadratique sur  $\mathbb{R}^n$  :

$$q(x) = \langle Ax, x \rangle = \sum_{i=1}^n \sum_{j=1}^n a_{i,j} x_i x_j.$$

En décomposant  $x$  suivant une base orthonormée  $(e_1, \dots, e_n)$  de vecteurs propres de  $A$  suivant  $x = \sum_{i=1}^n y_i e_i$  on peut ainsi écrire

$$q(x) = \sum_{i=1}^n \lambda_i y_i^2,$$

où  $\lambda_i$  est la valeur propre associée à  $e_i$ . Ceci montre en particulier que  $A$  est positive (respectivement définie positive) si et seulement si  $\lambda_i \geq 0$  (respectivement  $\lambda_i > 0$ ) pour  $i = 1, \dots, n$ . Cette remarque s'étend aux matrices auto-adjointes complexes avec  $q(x) = \langle Ax, x \rangle = \sum_{i=1}^n \sum_{j=1}^n a_{i,j} x_i \bar{x}_j$ .

### 1.3 Normes matricielles

Rappelons la définition d'une norme.

**Définition 1.3.1** Une norme sur un espace vectoriel  $E$  est une application  $x \mapsto \|x\|$  de  $E$  dans  $\mathbb{R}_+$  qui vérifie les propriétés suivantes

1.  $\|x\| = 0 \Rightarrow x = 0$ ,
2.  $\|\lambda x\| = |\lambda| \|x\|$  pour tout  $x \in E$  et  $\lambda \in \mathbb{K}$
3.  $\|x + y\| \leq \|x\| + \|y\|$  pour tout  $x, y \in E$ .

Nous avons déjà mentionné la norme euclidienne sur  $\mathbb{R}^n$  ou  $\mathbb{C}^n$  définie par

$$\|x\| := (\langle x, x \rangle)^{\frac{1}{2}} = \left( \sum_{i=1}^n |x_i|^2 \right)^{\frac{1}{2}}.$$

Il existe d'autres norme sur  $\mathbb{K}^n$  en particulier les normes  $\ell^p$

$$\|x\|_p := \left( \sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}},$$

dont la norme euclidienne est un cas particulier ( $p = 2$ ) et la norme "sup" ou  $\ell^\infty$

$$\|x\|_\infty := \max_{i=1, \dots, n} |x_i|.$$

Rappelons que deux normes  $\|\cdot\|_a$  et  $\|\cdot\|_b$  sont équivalentes si il existe des constantes  $0 < c \leq C$  telles que

$$c\|x\|_a \leq \|x\|_b \leq C\|x\|_a,$$

et que l'on a le résultat fondamental suivant.

**Théorème 1.3.1** Si  $E$  est un espace vectoriel de dimension finie, réel ou complexe, toutes les normes sur  $E$  sont équivalentes.

On s'intéresse à présent aux normes sur les espaces vectoriels de matrices. L'espace  $\mathcal{M}_{m,n}(\mathbb{K})$  est de dimension finie  $mn$ . On peut introduire comme pour  $\mathbb{K}^n$  les normes  $\ell^p$  sur les coefficients matriciels définies pour  $A = (a_{i,j})$  par

$$\|A\| := \left( \sum_{i=1}^m \sum_{j=1}^n |a_{i,j}|^p \right)^{\frac{1}{p}},$$

et la norme  $\ell^\infty$  définie comme le max des  $|a_{i,j}|$ . Notons que la norme  $\ell^2$  (aussi appelée norme de Hilbert-Schmidt) peut-être définie comme  $\text{tr}(A^*A)$ . En nous restreignant à présent au cadre des matrices carrées, on peut de manière naturelle une norme sur  $\mathcal{M}_n(\mathbb{K})$  à une norme vectorielle sur  $\mathbb{K}^n$ .

**Définition 1.3.2** Soit  $\|\cdot\|$  une norme sur  $\mathbb{K}^n$ . On lui associe une norme matricielle - dite "subordonnée" - sur  $\mathcal{M}_n(\mathbb{K})$  définie par

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

Il est très facile de vérifier que la quantité  $\|A\|$  définie ci-dessus vérifie bien les propriétés d'une norme. On remarque que  $\|I\| = 1$  pour toute norme de ce type.

**Remarque 1.3.1** Par linéarité, on peut aussi écrire

$$\|A\| = \sup_{\|x\|=1} \|Ax\|.$$

Comme  $x \mapsto \|Ax\|$  est continue et que la sphère unité  $\{\|x\| = 1\}$  est un ensemble compact, on voit que le sup est atteint (et peut donc être remplacé par un max).

Si  $\|\cdot\|$  est une norme vectorielle sur  $\mathbb{K}^n$ , on a par la définition de la norme subordonnée

$$\|Ax\| \leq \|A\| \|x\|,$$

pour tout  $x \in \mathbb{K}^n$  et  $A \in \mathcal{M}_n(\mathbb{K})$ . D'après la remarque précédente, il existe  $x_A \neq 0$  tel que

$$\|Ax_A\| = \|A\| \|x_A\|.$$

En appliquant deux fois de suite l'inégalité  $\|Ax\| \leq \|A\| \|x\|$ , on obtient que pour tout  $A, B \in \mathcal{M}_n(\mathbb{K})$  et  $x \in \mathbb{K}^n$ ,

$$\|ABx\| \leq \|A\| \|B\| \|x\|,$$

ce qui entraîne la propriété fondamentale suivante.

**Proposition 1.3.1** Pour tout  $A, B \in \mathcal{M}_n(\mathbb{K})$ , on a pour toute norme subordonnée à une norme vectorielle

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

En particulier  $\|A^n\| \leq \|A\|^n$  et  $\|A\| \|A^{-1}\| \geq 1$ .

Par abus de notation, on utilisera la même notation pour la norme subordonnée que pour la norme sur  $\mathbb{K}^n$ . Par exemple la norme subordonnée à la norme  $\ell^p$  est notée

$$\|A\|_p := \sup_{x \in \mathbb{K}^n, x \neq 0} \frac{\|Ax\|_p}{\|x\|_p},$$

et parfois appelée norme  $\ell^p$  de  $A$ . Elle ne doit pas être confondue avec la norme  $\ell^p$  des coefficients matriciels qui n'est pas une norme subordonnée à une norme vectorielle.

En l'absence de précision  $\|x\|$  désigne systématiquement la norme euclidienne et  $\|A\|$  la norme subordonnée à celle-ci.

Le résultat suivant donne les moyens de calculer explicitement les normes subordonnées aux normes vectorielles  $\ell^1$  et  $\ell^\infty$ .

**Proposition 1.3.2** Soit  $\|A\|_1$  et  $\|A\|_\infty$  les normes matricielles subordonnées aux normes  $\ell^1$  et  $\ell^\infty$  sur  $\mathbb{K}^n$ . On a

$$\|A\|_1 = \sup_{1 \leq j \leq n} \left( \sum_{i=1}^n |a_{i,j}| \right) \text{ et } \|A\|_\infty = \sup_{1 \leq i \leq n} \left( \sum_{j=1}^n |a_{i,j}| \right)$$

**Preuve :** Pour la norme  $\ell^1$  on écrit

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

On en déduit l'inégalité

$$\|A\|_1 \leq \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{i,j}|.$$

Soit  $j_0$  un indice tel que

$$\max_{1 \leq j \leq n} \sum_{i=1}^n |a_{i,j}| = \sum_{i=1}^n |a_{i,j_0}|.$$

Soit  $u \in \mathbb{K}^n$  défini par  $u_j = 0$  si  $j \neq j_0$ , et  $u_{j_0} = 1$ . On a

$$\|u\|_1 = 1 \text{ et } \|Au\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{i,j}|,$$

d'où le résultat. Pour la norme  $\ell^\infty$  on écrit

$$\|Ax\|_\infty = \max_{1 \leq i \leq n} \left| \sum_{j=1}^n a_{i,j} x_j \right| \leq \|x\|_\infty \left( \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{i,j}| \right),$$

d'où l'on déduit

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

Soit  $i_0$  un indice tel que

$$\max_{1 \leq i \leq n} \sum_{j=1}^n |a_{i,j}| = \sum_{j=1}^n |a_{i_0,j}|.$$

Soit  $u \in \mathbb{K}^n$  défini par  $u_j = 0$  si  $a_{i_0,j} = 0$ , et  $u_j = \frac{\bar{a}_{i_0,j}}{|a_{i_0,j}|}$  si  $a_{i_0,j} \neq 0$  (c'est à dire le signe de  $a_{i_0,j}$  dans le cas d'une matrice réelle). Si  $A$  est non nulle, on vérifie aisément que  $u$  est aussi non nul et que  $\|u\|_\infty = 1$  (si  $A = 0$ , il n'y a rien à démontrer). De plus,

$$\|Au\|_\infty \geq \left| \sum_{j=1}^n a_{i_0,j} u_j \right| = \sum_{j=1}^n |a_{i_0,j}| = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{i,j}|,$$

d'où le résultat. □

Introduisons à présent une quantité très importante en calcul numérique matriciel.

**Définition 1.3.3** Le rayon spectral de  $A \in \mathcal{M}_n(\mathbb{K})$  est défini par

$$\varrho(A) := \max_{i=1, \dots, p} |\lambda_i|,$$

où  $(\lambda_1, \dots, \lambda_p)$  sont les valeurs propres de  $A$  (dans le cas d'une matrice réelles on considère aussi ses valeurs propres complexes).

Le rayon spectral permet de calculer la norme subordonnée à la norme euclidienne.

**Proposition 1.3.3** Si  $A \in \mathcal{M}_n(\mathbb{K})$  est quelconque, on a

$$\|A\| = \sqrt{\varrho(A^*A)}.$$

Dans le cas où  $A$  est auto-adjointe, on a de plus

$$\|A\| = \max_{\|x\|=1} |\langle Ax, x \rangle| = \max_{x \in \mathbb{K}, x \neq 0} \frac{|\langle Ax, x \rangle|}{\|x\|^2} = \varrho(A).$$

**Preuve :** Prouvons tout d'abord la deuxième assertion. Si  $A$  est auto-adjointe, le Théorème 1.2.3 affirme qu'il existe une base orthonormée de vecteurs propres  $(e_1, \dots, e_n)$  de  $A$ . Pour  $x \in \mathbb{K}^n$  décomposé suivant  $x = \sum_{i=1}^n x_i e_i$  dans cette base on a

$$|\langle Ax, x \rangle| = \left| \sum_{i=1}^n \lambda_i |x_i|^2 \right| \leq \varrho(A) \|x\|^2,$$

et

$$\|Ax\|^2 = \sum_{i=1}^n \lambda_i^2 |x_i|^2 \leq \varrho(A)^2 \|x\|^2.$$

D'autre part si  $i_0$  est tel que  $\varrho(A) = |\lambda_{i_0}|$  en prenant  $x$  tel que  $x_{i_0} = 1$  et  $x_i = 0$  si  $i \neq i_0$  on trouve  $|\langle Ax, x \rangle| = \varrho(A) \|x\|^2$  et  $\|Ax\|^2 = \varrho(A)^2 \|x\|^2$ . On en déduit l'égalité annoncée. On en déduit ensuite la première assertion en écrivant, pour toute matrice  $A \in \mathcal{M}_n(\mathbb{C})$ ,

$$\|A\|^2 = \max_{\|x\|=1} \|Ax\|^2 = \max_{\|x\|=1} \langle A^*Ax, x \rangle,$$

et en remarquant alors que  $A^*A$  est autoadjointe et positive. □

La proposition suivante montre que le rayon spectral est un minorant des normes subordonnées.

**Proposition 1.3.4** Pour toute norme subordonnée à une norme vectorielle sur  $\mathbb{K}^n$  on a pour tout  $A \in \mathcal{M}_n(\mathbb{K})$

$$\varrho(A) \leq \|A\|$$

**Preuve :** Supposons toute d'abord  $\mathbb{K} = \mathbb{C}$ . Soit  $\lambda$  une valeur propre de  $A \in \mathcal{M}_n(\mathbb{C})$  telle que  $\varrho(A) = |\lambda|$  et soit  $x \neq 0$  tel que  $Ax = \lambda x$ . On a pour ce vecteur

$$\|Ax\| = |\lambda| \|x\| = \varrho(A) \|x\|,$$

ce qui entraîne  $\|A\| \geq \varrho(A)$ . Supposons à présent  $\mathbb{K} = \mathbb{R}$ . Dans ce cas, on ne peut pas raisonner de la même manière car on n'est pas assuré d'avoir un vecteur propre à coordonnées réelles. Si  $\|\cdot\|$  est une norme sur  $\mathbb{R}^n$  on lui associe la norme sur  $\mathbb{C}^n$  : pour tout  $x \in \mathbb{C}^n$

$$\|x\|_* := \max\{\|\Re(x)\|, \|\Im(x)\|\},$$

où les coordonnées des vecteurs  $\Re(x)$  et  $\Im(x)$  sont les parties réelles et imaginaires des coordonnées de  $x$ . Toute matrice  $A \in \mathcal{M}_n(\mathbb{R})$  peut-être vue comme une matrice complexe et on a d'après ce qui précède

$$\varrho(A) \leq \|A\|_* = \max_{x \in \mathbb{C}^n, x \neq 0} \frac{\|Ax\|_*}{\|x\|_*}.$$

On montre alors que  $\|A\|$  et  $\|A\|_*$  sont égales. En effet, on a d'une part

$$\begin{aligned} \|A\|_* &= \max_{x \in \mathbb{C}^n, \|x\|_* = 1} \max\{\|\Re(Ax)\|, \|\Im(Ax)\|\} \\ &= \max_{x \in \mathbb{C}^n, \|x\|_* = 1} \max\{\|A\Re(x)\|, \|A\Im(x)\|\} \\ &\leq \max_{x \in \mathbb{C}^n, \|x\|_* = 1} \|A\| \max\{\|\Re(x)\|, \|\Im(x)\|\} = \|A\|. \end{aligned}$$

et d'autre part

$$\begin{aligned}\|A\| &= \max_{x \in \mathbb{R}^n, \|x\|=1} \|Ax\| \\ &= \max_{x \in \mathbb{R}^n, \|x\|_* = 1} \|Ax\|_* \\ &\leq \max_{x \in \mathbb{C}^n, \|x\|_* = 1} \|Ax\|_* = \|A\|_*.\end{aligned}$$

Ceci permet de conclure dans le cas  $\mathbb{K} = \mathbb{R}$ . □

Le résultat suivant joue un rôle important dans l'étude des puissances  $A^k$  d'une matrice, comme on le verra dans la section suivante.

**Proposition 1.3.5** *Pour toute matrice  $A \in \mathcal{M}_n(\mathbb{K})$  et pour tout réel  $\varepsilon > 0$ , il existe une norme subordonnée à une norme sur  $\mathbb{C}^n$  (qui dépend de  $A$  et  $\varepsilon$ ) telle que*

$$\|A\| \leq \varrho(A) + \varepsilon.$$

**Preuve :** D'après la Proposition 1.2.1, il existe une matrice  $U$  unitaire telle que  $T = U^{-1}AU$  est triangulaire supérieure et les éléments diagonaux  $t_{i,i}$  sont les valeurs propres de  $A$ . Pour tout  $\delta > 0$  on définit une matrice diagonale  $D_\delta = \text{diag}(1, \delta, \delta^2, \dots, \delta^{n-1})$  de telle sorte qu'en posant  $U_\delta := UD_\delta$ , la matrice  $T_\delta$  définie par

$$T_\delta = U_\delta^{-1}AU_\delta = (UD_\delta)^{-1}A(UD_\delta) = D_\delta^{-1}TD_\delta$$

vérifie

$$T_\delta = \begin{pmatrix} t_{1,1} & \delta t_{1,2} & \cdots & \delta^{n-1}t_{1,n} \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \delta t_{n-1,n} \\ 0 & \cdots & 0 & t_{n,n} \end{pmatrix}.$$

Etant donné  $\varepsilon > 0$ , on peut choisir  $\delta$  suffisamment petit pour que les éléments extra-diagonaux de  $T_\delta$  soient aussi très petits, par exemple pour que, pour tout  $1 \leq i \leq n-1$ ,

$$\sum_{j=i+1}^n \delta^{j-i} |t_{i,j}| \leq \varepsilon.$$

Comme les  $t_{i,i}$  sont les valeurs propres de  $T_\delta$  qui est semblable à  $A$ , on en déduit que  $\|T_\delta\|_\infty \leq \varrho(A) + \varepsilon$ . L'application

$$B \mapsto \|B\| = \|U_\delta^{-1}BU_\delta\|_\infty = \max_{x \in \mathbb{C}^n, x \neq 0} \frac{\|U_\delta^{-1}BU_\delta x\|_\infty}{\|x\|_\infty} = \max_{y \in \mathbb{C}^n, y \neq 0} \frac{\|U_\delta^{-1}By\|_\infty}{\|U_\delta^{-1}y\|_\infty}$$

est la norme subordonnée à la norme vectorielle  $x \mapsto \|U_\delta^{-1}x\|_\infty$  qui dépend de  $A$  et  $\varepsilon$ , et on a

$$\|A\| \leq \varrho(A) + \varepsilon,$$

d'où le résultat. □

## 1.4 Séries de matrices

Soit  $P(z) = a_0 + a_1z + a_2z^2 + \cdots + a_mz^m$  un polynôme à coefficients dans  $\mathbb{K}$ . Pour tout  $A \in \mathcal{M}_n(\mathbb{K})$ , il est naturel de définir  $P(A) \in \mathcal{M}_n(\mathbb{K})$  par la formule

$$P(A) := a_0I + a_1A + a_2A^2 + \cdots + a_mA^m.$$

On s'intéresse à l'extension de ce procédé à des fonctions  $f$  qui ont un développement en *série entière* autour de l'origine

$$f(z) = \sum_{k \geq 0} a_k z^k$$

et on étudie dont la convergence de la série

$$f(A) := \sum_{k \geq 0} a_k A^k,$$

où l'on a posé  $A^0 = I$ . De façon générale, si  $(A_k)_{k \geq 0}$  est une suite de matrice de  $\mathcal{M}_n(\mathbb{K})$  la série  $\sum_{k \geq 0} A_k$  converge si et seulement si il existe une matrice  $A$  telle que pour tout  $\varepsilon > 0$  il existe  $K$  tel que

$$\left\| \sum_{k=0}^K A_k - A \right\| \leq \varepsilon.$$

Notons que le choix de la norme n'a pas d'importance puisque toutes sont équivalentes. Comme  $\mathcal{M}_n(\mathbb{K})$  est complet, la convergence est équivalente au *critère de Cauchy* : pour tout  $\varepsilon > 0$ , il existe  $K$  tel que

$$\left\| \sum_{k=K}^L A_k \right\| \leq \varepsilon,$$

pour tout  $L > K$ . La série est dite normalement convergente si et seulement si la série à termes réels positifs  $\sum_{k \geq 0} \|A_k\|$  converge, ce qui est une propriété plus forte que la convergence.

Afin d'étudier la série  $\sum_{k \geq 0} a_k A^k$ , il est important de comprendre le comportement des puissances  $A^k$  lorsque  $k \rightarrow +\infty$ . Ce comportement est décrit par le théorème important suivant.

**Théorème 1.4.1** *Soit  $A \in \mathcal{M}_n(\mathbb{K})$  et  $\|\cdot\|$  une norme sur  $\mathcal{M}_n(\mathbb{K})$ . On a*

$$\lim_{k \rightarrow +\infty} \|A^k\|^{\frac{1}{k}} = \varrho(A),$$

où  $\varrho(A)$  est le rayon spectral de  $A$ .

**Preuve :** Soit  $\varepsilon > 0$ . D'après la Proposition 1.3.5, pour tout réel  $\tilde{\varepsilon} > 0$ , il existe une norme subordonnée  $\|\cdot\|_*$  sur  $\mathbb{C}^n$  telle que

$$\|A\|_* \leq \varrho(A) + \tilde{\varepsilon}.$$

Pour cette norme on a pour tout  $k \geq 0$ ,

$$\|A^k\|_* \leq (\varrho(A) + \tilde{\varepsilon})^k.$$

Comme cette norme est équivalente à la norme  $\|\cdot\|$ , il existe une constante  $C > 0$  telle que

$$\|A^k\| \leq C(\varrho(A) + \tilde{\varepsilon})^k,$$

soit

$$\|A^k\|^{\frac{1}{k}} \leq C^{\frac{1}{k}}(\varrho(A) + \tilde{\varepsilon}).$$

Puisque  $C^{\frac{1}{k}}$  tend vers 1 quand  $k \rightarrow +\infty$ , on peut choisir  $\tilde{\varepsilon} > 0$  et  $K$  tel que

$$\|A^k\|^{\frac{1}{k}} \leq \varrho(A) + \varepsilon.$$

pour tout  $k > K$ . D'autre part, d'après la Proposition 1.3.4, on sait que pour toute norme subordonnée  $\|\cdot\|_*$ , on a aussi bien pour les matrices réelles ou complexe  $\varrho(A) \leq \|A\|_*$ , d'où pour tout  $k \geq 0$ ,

$$\varrho(A)^k = \varrho(A^k) \leq \|A^k\|_* \leq C\|A^k\|,$$

où on a encore utilisé l'équivalence des normes, ce qui entraîne

$$\|A^k\|^{\frac{1}{k}} \geq C^{-\frac{1}{k}} \varrho(A).$$

On peut donc choisir  $K$  tel que

$$\varrho(A) - \varepsilon \leq \|A^k\|^{\frac{1}{k}} \leq \varrho(A) + \varepsilon,$$

pour tout  $k > K$ . □

Une conséquence immédiate de ce théorème est la dichotomie suivante :

$$\varrho(A) < 1 \Rightarrow \lim_{k \rightarrow +\infty} \|A^k\| = 0,$$

et

$$\varrho(A) > 1 \Rightarrow \lim_{k \rightarrow +\infty} \|A^k\| = +\infty.$$

Ce théorème nous permet d'étudier la convergence de la série  $\sum_{k \geq 0} a_k A^k$  au moyen du rayon spectral de  $A$ . Rappelons que le rayon de convergence de la série entière  $f(z) = \sum_{k \geq 0} a_k z^k$  est défini par

$$R := [\limsup_{k \rightarrow +\infty} |a_k|^{\frac{1}{k}}]^{-1},$$

et que la série  $\sum_{k \geq 0} |a_k| r^k$  converge pour tout  $r \in [0, R[$ .

**Corollaire 1.4.1** *Soit  $f(z) = \sum_{k \geq 0} a_k z^k$  une série entière de rayon de convergence  $R > 0$ . Pour tout  $A \in \mathcal{M}_n(\mathbb{K})$  tel que  $\varrho(A) < R$ , la série définissant  $f(A) := \sum_{k \geq 0} a_k A^k$  est normalement convergente.*

**Preuve :** Soit  $r > 0$  tel que  $\varrho(A) < r < R$ . D'après le Théorème 1.4.1, il existe  $K$  tel que pour  $k > K$  on a

$$\|A^k\|^{\frac{1}{k}} \leq r,$$

d'où  $\|A^k\| \leq r^k$ , ce qui entraîne la convergence normale de  $\sum_{k \geq 0} a_k A^k$ . □

**Remarque 1.4.1** *Il est aussi facile de vérifier que la série est divergente lorsque  $\varrho(A) > R$  car le Théorème 1.4.1 permet alors de montrer que la norme du terme général  $\|a_k A^k\|$  ne tend pas vers 0.*

**Remarque 1.4.2** *Un résultat plus faible qui ne nécessite pas l'utilisation du rayon spectral est que la série converge normalement si on a  $\|A\| < R$  pour une norme subordonnée.*

Examinons à présent deux exemples importants de séries entières de matrices. Le premier est la *série de Neumann*, associée à la série entière  $f(z) = \sum_{k \geq 0} z^k$  dont le rayon de convergence est  $R = 1$  et qui vaut  $\frac{1}{1-z}$  pour tout  $|z| < 1$ . Pour  $A \in \mathcal{M}_n(\mathbb{K})$  telle que  $\varrho(A) < 1$  la série

$$f(A) = I + A + A^2 + \dots = \sum_{k \geq 0} A^k,$$

est normalement convergente. On voit d'autre part que

$$(I - A)f(A) = \lim_{K \rightarrow +\infty} \sum_{k=0}^K (I - A)A^k = \lim_{K \rightarrow +\infty} I - A^{K+1} = I,$$

ce qui prouve le résultat suivant.

**Proposition 1.4.1** *Si  $A \in \mathcal{M}_n(\mathbb{K})$  est telle que  $\varrho(A) < 1$ , alors  $I - A$  est inversible et son inverse  $(I - A)^{-1}$  est égal à la série de Neumann  $\sum_{k \geq 0} A^k$ .*

Le deuxième exemple est l'*exponentielle matricielle* associée à la série entière  $\exp(z) = \sum_{k \geq 0} \frac{z^k}{k!}$  dont le rayon de convergence est  $R = +\infty$  et converge pour tout  $z \in \mathbb{C}$ . Pour tout  $A \in \mathcal{M}_n(\mathbb{K})$  on peut donc définir la matrice exponentielle de  $A$  par

$$\exp(A) := I + A + \frac{1}{2}A^2 + \frac{1}{6}A^3 \dots = \sum_{k \geq 0} \frac{1}{k!} A^k,$$

qui est aussi notée  $e^A$ . Une propriété fondamentale de l'exponentielle réelle et complexe est  $\exp(a + b) = \exp(a)\exp(b)$ . Le résultat suivant indique que cette propriété est vérifiée par l'exponentielle matricielle sous la contrainte que les matrices commutent.

**Proposition 1.4.2** Si  $A, B \in \mathcal{M}_n(\mathbb{K})$  sont telles que  $AB = BA$ , on a alors

$$\exp(A + B) = \exp(A) \exp(B) = \exp(B) \exp(A).$$

**Preuve :** On développe le produit  $\exp(A) \exp(B)$  et on effectue une sommation “par paquets” suivant

$$\begin{aligned} \exp(A) \exp(B) &= \left( \sum_{k \geq 0} \frac{1}{k!} A^k \right) \left( \sum_{l \geq 0} \frac{1}{l!} B^l \right) \\ &= \sum_{k, l \geq 0} \frac{1}{k! l!} A^k B^l \\ &= \sum_{m \geq 0} \sum_{k=0}^m \frac{1}{k! (m-k)!} A^k B^{m-k} \\ &= \sum_{m \geq 0} \frac{1}{m!} \sum_{k=0}^m \frac{m!}{k! (m-k)!} A^k B^{m-k}. \end{aligned}$$

Comme les matrices  $A$  et  $B$  commutent, on peut appliquer la formule du binôme de Newton

$$(A + B)^m = \sum_{k=0}^m \frac{m!}{k! (m-k)!} A^k B^{m-k},$$

ce qui prouve  $\exp(A) \exp(B) = \sum_{m \geq 0} \frac{1}{m!} (A + B)^m = \exp(A + B)$ . □

**Remarque 1.4.3** On déduit en particulier de ce résultat que  $\exp(A)$  est toujours inversible d'inverse  $\exp(-A)$ .

**Remarque 1.4.4** Si  $AB = BA$  si deux séries entières  $f(A)$  et  $g(B)$  convergent, alors  $f(A)g(B) = g(B)f(A)$ .

**Remarque 1.4.5** Si  $f(z) = \sum_{k \geq 0} a_k z^k$  est une série entière de rayon de convergence  $R > 0$ , et si  $A$  et  $B = P^{-1}AP$  sont deux matrices semblables telles que  $\varrho(A) = \varrho(B) < R$ , on a alors

$$f(B) = \sum_{k \geq 0} a_k (P^{-1}AP)^k = \sum_{k \geq 0} a_k P^{-1} A^k P = P^{-1} f(A) P$$

En particulier si  $A$  est diagonalisable, et que l'on a donc  $P^{-1}AP = D$  où  $D = \text{diag}(\lambda_1, \dots, \lambda_n)$  est diagonale, on a alors

$$f(A) = P f(D) P^{-1} = P \text{diag}(f(\lambda_1), \dots, f(\lambda_n)) P^{-1},$$

ce qui permet de simplifier le calcul de  $f(A)$ . On a par exemple

$$\exp(A) = P \text{diag}(e^{\lambda_1}, \dots, e^{\lambda_n}) P^{-1}.$$

## 1.5 Résolution de systèmes linéaires

Un système linéaire  $m \times n$  est un système d'équation de la forme

$$\begin{cases} a_{1,1}x_1 + \dots + a_{1,n}x_n = b_1, \\ \vdots \\ a_{m,1}x_1 + \dots + a_{m,n}x_n = b_m. \end{cases}$$

où les inconnues sont  $(x_1, \dots, x_n)$  et les coefficients  $a_{i,j}$  du système et  $b_i$  du second membre sont donnés. Ce système s'écrit sous forme matricielle

$$Ax = b,$$

où  $A = (a_{i,j}) \in \mathcal{M}_{m,n}(\mathbb{K})$  et  $x$  et  $b$  sont des vecteurs colonnes de  $\mathbb{K}^n$  et  $\mathbb{K}^m$  dont les composantes sont  $(x_1, \dots, x_n)$  et  $(b_1, \dots, b_m)$ . Il existe au moins une solution si  $b \in \text{Im}(A)$  et il en existe au plus une si  $\text{Ker}(A) = \{0\}$ . Si  $x$  est une solution, l'ensemble de toutes les solution est l'espace affine

$$S = \{x + y ; y \in \text{Ker}(A)\}.$$

Le cas  $m = n$  des systèmes carrés  $n \times n$  est important car c'est le plus souvent rencontré. Dans ce cas, l'inversibilité de  $A$  est équivalente à l'existence et l'unicité d'une solution quelque soit le second membre  $b$ . On rappelle les *formules de Cramer* qui donnent une expression explicite de la solution lorsque  $A$  est inversible : en notant  $a_1, \dots, a_n$  les vecteurs colonnes de  $A$  on a

$$x_i = \frac{\det(a_1, \dots, a_{i-1}, b, a_{i+1}, \dots, a_n)}{\det(A)}.$$

Du point de vue du calcul numérique, ces formules sont inutilisables pour des systèmes de grande taille car le calcul du déterminant d'une matrice par son développement par ligne ou colonne nécessite de l'ordre de  $n!$  opérations. A titre indicatif, si  $n = 20$  le calcul de la solution par les formules de Cramer nécessite plus de  $5 \times 10^{19}$  opérations élémentaires (additions et multiplications) soit 50 milliards de milliards. On a donc recours à d'autres méthodes de résolution, parmi lesquelles il convient de distinguer deux catégories : les méthodes *directes* et les méthodes *itératives*

Les méthodes dites directes consistent en une série d'opérations qui conduisent à la solution exacte du système. La méthode directe la plus couramment utilisée est la méthode d'élimination (ou du pivot) de Gauss, dont nous rappelons le principe dans le cas d'un système  $n \times n$  que l'on peut exprimer au moyen de la *matrice augmentée*

$$\begin{pmatrix} a_{1,1} & \cdots & a_{1,n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{m,1} & \cdots & a_{m,n} & b_m \end{pmatrix},$$

la ligne  $i$  représentant l'équation  $\sum_{j=1}^n a_{i,j}x_j = b_i$ . On peut effectuer 3 types d'opérations sur le système et sa matrice augmentée qui laissent inchangé l'ensemble des solutions : (i) *échange* de deux lignes (ii) *multiplication* d'une ligne par un nombre arbitraire non-nul (iii) *substitution* c'est à dire ajout à la ligne  $i$  d'une autre ligne multipliée par un nombre arbitraire. On pratique d'abord un échange assurant que le premier coefficient de la première ligne est non nul et au moyen de  $n$  substitution on aboutit à un système du type

$$\begin{pmatrix} a'_{1,1} & a'_{1,2} & \cdots & a'_{1,n} & b'_1 \\ 0 & a'_{2,2} & \cdots & a'_{2,n} & b'_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & a'_{m,2} & \cdots & a'_{m,n} & b'_m \end{pmatrix},$$

avec  $a'_{1,1} = 0$  dans le cas où la première colonne du système initial était nulle. En itérant cette procédure, on aboutit à un système dont la matrice augmentée est de type *échelonnée* : le premier coefficient non-nul de la ligne  $i$  apparaît à une position horizontale  $j(i)$  strictement plus grande que celle du premier coefficient non-nul de la ligne  $i-1$  (sauf si la ligne  $i$  est entièrement nulle auquel cas toutes les lignes suivantes le sont aussi). Les positions  $(i, j(i))$  sont appelées "pivot". Voici un exemple d'une telle transformation dans le cas du système constitué des trois équations  $x_3 + 2x_4 = 1$ ,  $x_1 + 4x_2 - 2x_3 - x_4 = 3$  et  $2x_1 + 8x_2 + x_3 + 2x_4 = -1$  :

$$\begin{aligned} & \begin{pmatrix} 0 & 0 & 1 & 2 & 1 \\ 1 & 4 & -2 & -1 & 3 \\ 2 & 8 & 1 & 2 & -1 \end{pmatrix} \Leftrightarrow \begin{pmatrix} 1 & 4 & -2 & -1 & 3 \\ 0 & 0 & 1 & 2 & 1 \\ 2 & 8 & 1 & 2 & -1 \end{pmatrix} \\ & \Leftrightarrow \begin{pmatrix} 1 & 4 & -2 & -1 & 3 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 5 & 4 & -7 \end{pmatrix} \Leftrightarrow \begin{pmatrix} 1 & 4 & -2 & -1 & 3 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & -6 & -12 \end{pmatrix} \end{aligned}$$

Dans ce cas les positions de pivots sont  $(1, 1)$ ,  $(2, 3)$  et  $(3, 4)$ . En effectuant une série de multiplication et de substitutions à partir de la dernière ligne, on peut mettre la valeur 1 sur tous les pivots et des coefficients nuls au dessus de ceux-ci. Dans l'exemple précédent on obtient :

$$\begin{pmatrix} 1 & 4 & -2 & -1 & 3 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix} \Leftrightarrow \begin{pmatrix} 1 & 4 & -2 & 0 & 5 \\ 0 & 0 & 1 & 0 & -3 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix} \Leftrightarrow \begin{pmatrix} 1 & 4 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -3 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix}$$

Le système est alors complètement résolu : les variables  $x_j$  correspondant à une position de pivot, dites “essentiels” s’expriment en fonctions du second membre et des éventuelles autres variables dites “libres” qui apparaissent en cas de non-unicité de la solution. Dans l’exemple ci-dessus,  $x_2$  est la variable libre et l’ensemble des solutions est  $\{(-1 - 4t, t, -3, 2) ; t \in \mathbb{R}\}$ . Le système n’admet pas de solution lorsqu’un pivot apparaît sur la dernière colonne de la matrice augmentée.

Dans le cas d’un système  $n \times n$ , le nombre d’opérations élémentaires effectuées dans la méthode de Gauss est de l’ordre de  $n^3$ , soit 8000 dans le cas  $n = 20$  ce qui est nettement inférieur aux formules de Cramer. Ce nombre devient néanmoins élevé pour un ordinateur usuel lorsque  $n$  est de l’ordre du millier et il faut alors envisager d’autres types de méthodes.

Les méthodes itératives visent à construire de proche en proche une suite de vecteurs qui tend vers la solution exacte du système. Nous présentons ici un exemple élémentaire, la méthode de Richardson. On considère ici un système  $n \times n$  et on fait l’hypothèse que  $A$  est inversible. On pose  $x^0 = 0$  et on construit itérativement les vecteurs  $x^k \in \mathbb{K}^n$  par

$$x^{k+1} = x^k + \tau(b - Ax^k),$$

où  $\tau > 0$  est un nombre fixé. On note que la solution  $x$  de  $Ax = b$  est un point fixe de cette itération :

$$x = x + \tau(b - Ax).$$

On remarque aussi que si la suite  $x^k$  converge, sa limite  $x$  vérifie l’égalité ci-dessus et est donc solution de  $Ax = b$ . Afin d’étudier la convergence on introduit le vecteur d’erreur  $e^k := x - x^k$ . En faisant la différence entre les deux égalités ci-dessus, on obtient

$$e^{k+1} = (I - \tau A)e^k,$$

soit par itération

$$e^k = (I - \tau A)^k e^0 = (I - \tau A)^k x.$$

Si  $\|\cdot\|$  est une norme vectorielle donnée, on a donc

$$\|x - x^k\| = \|e^k\| \leq \|(I - \tau A)^k\| \|x\|,$$

où la norme matricielle est la norme subordonnée. En utilisant le Théorème 1.4.1 on obtient donc le résultat suivant qui donne une condition pour la convergence de la méthode de Richardson.

**Proposition 1.5.1** *Si  $\varrho(I - \tau A) < 1$  alors  $\lim_{k \rightarrow +\infty} x^k = x$ .*

Il n’est pas toujours possible d’assurer la condition  $\varrho(I - \tau A) < 1$ . Un cas important où cela est possible est celui où  $A$  est auto-adjointe, définie et positive. En notant  $0 < \lambda_{\min} < \lambda_{\max}$  les valeurs propres de  $A$ , les valeurs propres de  $I - \tau A$  sont comprises entre  $1 - \tau\lambda_{\max}$  et  $1 - \tau\lambda_{\min}$ , et par conséquent

$$0 < \tau < \frac{2}{\lambda_{\max}} \Rightarrow \varrho(I - \tau A) = \max\{|1 - \tau\lambda_{\max}|, |1 - \tau\lambda_{\min}|\} < 1.$$

On peut choisir la valeur  $\tau = \frac{1}{\lambda_{\max} + \lambda_{\min}}$  pour minimiser la valeur de  $\varrho(I - \tau A)$ . En introduisant le *nombre de conditionnement* de la matrice défini par  $A$

$$\kappa := \frac{\lambda_{\max}}{\lambda_{\min}},$$

on a alors

$$\varrho(I - \tau A) = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \frac{\kappa - 1}{\kappa + 1} = 1 - \frac{2}{\kappa + 1}.$$

Comme  $(I - \tau A)^k$  est autoadjointe on a  $\|(I - \tau A)^k\| = [\varrho(I - \tau A)]^k$  pour la norme subordonnée à la norme euclidienne, et donc

$$\|x - x^k\| \leq \left(1 - \frac{2}{\kappa + 1}\right)^k \|x\|.$$

Cette inégalité indique que l’erreur décroît exponentiellement mais avec un taux de décroissance proche de 1 lorsque  $\kappa$  est très grand par rapport à 1 (on dit alors que  $A$  est *mal conditionnée*).

**Remarque 1.5.1** Le nombre de conditionnement a aussi une influence sur la propagation des erreurs d'arrondis, quelque soit la méthode de résolution du système. Illustrons le dans le cas d'un système  $Ax = b$  où  $A$  est autoadjointe définie et positive. On a alors  $\|A\| = \lambda_{\max}$  et  $\|A^{-1}\| = \lambda_{\min}^{-1}$  et donc  $\kappa = \|A\|\|A^{-1}\|$  pour la norme subordonnée à la norme euclidienne. En supposant que l'on commet une erreur  $\delta b$  sur  $b$  et en notant  $\delta x$  l'erreur résultante sur  $x$ , telle que

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

c'est à dire  $\delta x = A^{-1}\delta b$ , on obtient alors pour les erreurs relatives

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

On voit ainsi par exemple qu'une erreur relative de 1% sur la donnée  $b$  peut se traduire par une erreur relative de 100% sur la solution  $x$  si le nombre de conditionnement est supérieur à 100.

Nous terminons ce chapitre en présentant la méthode des *moindres carrés* pour les systèmes réels. Cette méthode est utile pour traiter numériquement des systèmes qui ne sont pas nécessairement carrés et n'admettent pas nécessairement de solution unique.

**Définition 1.5.1** Soit  $A \in \mathcal{M}_{m,n}(\mathbb{R})$  et  $b \in \mathbb{R}^m$ . On dit que  $x^* \in \mathbb{R}^n$  est solution de  $Ax = b$  au sens des moindres carrés si et seulement si pour tout  $x \in \mathbb{R}^n$  on a

$$\|Ax^* - b\| \leq \|Ax - b\|,$$

où  $\|\cdot\|$  est la norme euclidienne.

On cherche donc  $x^*$  qui minimise la fonction  $E(x) = \|Ax - b\|^2$ . En écrivant pour tout  $h \in \mathbb{R}^n$

$$\begin{aligned} E(x+h) &= \|Ax + Ah - b\|^2 \\ &= \|Ax - b\|^2 + 2\langle Ax - b, Ah \rangle + \|Ah\|^2 \\ &= E(x) + \langle 2A^t(Ax - b), h \rangle + \|Ah\|^2, \end{aligned}$$

on voit que  $\nabla E(x) = 2A^t(Ax - b)$ . Si  $x^*$  est un minimiseur de  $E$ , il est par conséquent solution de l'équation

$$A^t Ax = A^t b,$$

appelée *équation normale* et qui est un système carré  $n \times n$ . Réciproquement le développement de  $E(x+h)$  nous montre que toute solution de l'équation normale est un minimiseur de  $E$ . On remarque aussi que si  $x^*$  est solution de  $Ax = b$ , il est aussi solution de l'équation normale et du problème des moindres carrés.

Contrairement à l'équation  $Ax = b$ , l'équation normale, admet toujours une solution. En effet, en remarquant que  $\langle A^t Ax, x \rangle = \|Ax\|^2$ , on obtient facilement que

$$\text{Ker}(A^t A) = \text{Ker}(A),$$

ce qui entraîne  $\text{Im}(A^t A) = \text{Im}(A^t)$ . Cette solution est unique dans le cas où  $\text{Ker}(A) = 0$ , auquel cas  $A^t A$  est autoadjointe, définie et positive.

## 2 Approximation de solutions d'équations

Une équation scalaire a la forme générale  $f(x) = 0$  où  $f$  est une fonction de  $\mathbb{R}$  dans  $\mathbb{R}$ . Un système de  $n$  équations à  $n$  inconnues peut aussi se mettre sous une telle forme avec  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$  représentant les  $n$  inconnues et  $f$  une fonction vectorielle de  $\mathbb{R}^n$  dans  $\mathbb{R}^n$ . La résolution exacte de telles équations est possible dans des cas simples, par exemple lorsque  $f$  est une fonction affine ou quadratique de  $\mathbb{R}$  dans  $\mathbb{R}$ , mais elle est souvent hors de portée pour des fonctions plus générales. On a alors recours à des méthodes numériques itératives pour la résolution approchée de l'équation. Ces méthodes sont basées sur une reformulation de l'équation par un problème équivalent de point fixe : trouver  $x$  tel que  $g(x) = x$  où  $g$  est une fonction liée à  $f$ . Un exemple d'une telle méthode a été donné dans la section §1.5 pour la résolution du système  $Ax = b$  en posant  $f(x) = b - Ax$  et  $g(x) = x + \tau(b - Ax)$ .

### 2.1 Le théorème du point fixe

On donne d'abord un résultat général pour une fonction  $g$  allant d'un espace vectoriel  $E$  dans lui-même.

**Définition 2.1.1** Soit  $E$  un espace vectoriel muni d'une norme  $\|\cdot\|$  et soit  $F$  un sous-ensemble de  $E$ . Une fonction  $g : E \rightarrow E$  définie sur  $F$  est dite "contractante" sur  $F$  si et seulement si il existe une constante  $0 < a < 1$  telle que

$$\|g(x) - g(y)\| \leq a\|x - y\|,$$

pour tout  $x, y \in F$ .

Une fonction contractante est donc une fonction  $a$ -lipschitzienne avec  $a < 1$ . On rappelle que les fonctions lipschitziennes sont toujours continues.

**Théorème 2.1.1** (du point fixe de Picard) Soit  $E$  un espace vectoriel de dimension finie muni d'une norme  $\|\cdot\|$  et  $F$  un sous-ensemble fermé de  $E$ . Soit  $g$  une fonction contractante sur  $F$  et telle que  $g(F) \subset F$ . Alors il existe un unique  $x^* \in F$  tel que  $g(x^*) = x^*$ .

**Preuve :** Soit  $x^0$  un point quelconque de  $F$ . On définit par récurrence la suite

$$x^{n+1} = g(x^n).$$

Cette suite est contenue dans  $F$  puisque  $g(F) \subset F$ . La propriété de contraction de  $g$  entraîne

$$\|x^{n+1} - x^n\| = \|g(x^n) - g(x^{n-1})\| \leq a\|x^n - x^{n-1}\|,$$

et par récurrence on obtient donc

$$\|x^{n+1} - x^n\| \leq a^n \|x^1 - x^0\|.$$

Comme  $a < 1$  ceci montre que la série de terme général  $\|x^{n+1} - x^n\|$  converge, ce qui entraîne la convergence de la série  $\sum (x^{n+1} - x^n)$  qui équivaut à la convergence de la suite  $x^n$ . Sa limite  $x^*$  appartient à  $F$  puisque cet ensemble est fermé, et par continuité de  $g$  l'égalité  $x^{n+1} = g(x^n)$  entraîne

$$x^* = g(x^*).$$

Si  $y^* \in F$  est un autre point fixe de  $g$ , on a

$$\|x^* - y^*\| = \|g(x^*) - g(y^*)\| \leq a\|x^* - y^*\|,$$

ce qui entraîne  $x^* = y^*$  puisque  $a < 1$ , d'où l'unicité du point fixe. □

**Remarque 2.1.1** Ce théorème s'étend au cas où  $E$  est un espace métrique muni d'une distance  $(x, y) \mapsto d(x, y)$  pour laquelle il est complet. La propriété de contraction s'écrit alors  $d(g(x), g(y)) \leq ad(x, y)$ .

**Remarque 2.1.2** Lorsque  $E$  est un espace de dimension finie tel que  $\mathbb{R}^n$ , on sait que toutes les normes sont équivalentes. Cependant, la propriété de contraction peut être vérifiée par une norme et non par une autre. Par exemple si  $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  est définie par  $g(x) = \frac{9}{10}R(x)$  où  $R$  est la rotation d'angle  $\pi/4$  autour de l'origine, on a la propriété de contraction

$$\|g(x) - g(y)\| = \frac{9}{10}\|x - y\|,$$

pour la norme euclidienne, mais pour la norme  $\ell^1$  on a avec  $x = (0, 0)$  et  $y = (1, 0)$ ,

$$\|g(x) - g(y)\|_1 = \frac{9\sqrt{2}}{10} > 1 = \|x - y\|_1.$$

Ceci montre que le choix de la norme est important pour établir la propriété de contraction.

**Remarque 2.1.3** Le théorème de Picard est "constructif" au sens où sa preuve donne une méthode itérative qui permet d'approcher le point fixe de  $g$ . Il existe d'autres théorèmes de point fixe qui ne sont pas constructifs. En particulier le théorème de Brouwer affirme l'existence d'un point fixe lorsque  $g$  est une fonction continue d'un compact  $F$  dans lui-même sans l'hypothèse de contraction, mais en ajoutant des hypothèses de nature topologique sur l'ensemble  $F$ . En dimension 1, il est facile de vérifier que ce résultat est vrai si le compact  $F \subset \mathbb{R}$  est connexe, c'est à dire un intervalle fermé : on remarque que la fonction  $x \mapsto g(x) - x$  admet au moins une valeur positive et une valeur négative, et par conséquent s'annule en un point de  $F$ . En dimension  $n \geq 1$ , ce résultat reste vrai si l'on suppose par exemple que  $F \subset \mathbb{R}^n$  est un convexe fermé, mais sa preuve est nettement plus difficile. Il existe des ensembles connexes pour lequel ce résultat est faux : considérer par exemple l'application  $g(x) = -x$  sur le cercle unité de  $\mathbb{R}^2$ .

## 2.2 La méthode du point fixe pour les fonctions réelles

Dans cette section comme dans la suivante, on suppose que  $g$  est une fonction à variable et à valeur réelles. Nous allons examiner plus en détail le comportement de l'algorithme du point fixe  $x^{n+1} = g(x^n)$  dans le cas où la fonction  $g$  est de classe  $\mathcal{C}^1$ . La proposition suivante est une conséquence immédiate du théorème des accroissement finis.

**Proposition 2.2.1** Soit  $g$  une fonction de classe  $\mathcal{C}^1$  sur un intervalle ouvert  $I$ , alors  $g$  est  $a$ -lipschitzienne sur un intervalle  $J \subset I$  si et seulement si  $|g'(t)| \leq a$  pour tout  $t \in J$ .

Considérons à présent une fonction  $g$  de classe  $\mathcal{C}^1$  sur un intervalle ouvert  $I$ , et supposons que  $x^* \in I$  soit un point fixe de  $g$ . Nous pouvons distinguer plusieurs cas en fonction des valeurs de la dérivée de  $g$  en  $x^*$ .

**Cas 1.**  $|g'(x^*)| < 1$ . Dans ce cas, puisque  $g'$  est continue, il existe  $r > 0$  tel que

$$|x - x^*| \leq r \Rightarrow |g'(x) - g'(x^*)| \leq \frac{1 - |g'(x^*)|}{2} \Rightarrow |g'(x)| \leq a := \frac{1 + |g'(x^*)|}{2} < 1.$$

D'après la proposition précédente  $g$  est  $a$ -lipschitzienne et donc contractante sur l'intervalle  $F = [x^* - r, x^* + r]$ . Pour tout  $x \in F$  on a

$$|g(x) - x^*| = |g(x) - g(x^*)| \leq a|x - x^*| \leq ar \leq r,$$

et par conséquent  $g(F) \subset F$ . Le Théorème 2.1.1 permet donc d'affirmer que pour tout  $x^0 \in F$ , la suite  $x^{n+1} = g(x^n)$  converge vers  $x^*$  avec une vitesse de convergence géométrique :

$$\|x^n - x^*\| = \|g(x^{n-1}) - g(x^*)\| \leq a\|x^{n-1} - x^*\| \leq \dots \leq a^n\|x^0 - x^*\| \leq ra^n.$$

On dit que le point fixe  $x^*$  est *attractif* : l'algorithme converge pour tout  $x^0$  suffisamment proche de  $x^*$ .

**Cas 2.**  $|g'(x^*)| > 1$ . Supposons par exemple  $g'(x^*) > 1$ . En utilisant la continuité de  $g'$  de la même manière que dans le cas précédent, on obtient qu'il existe  $\varepsilon > 0$  tel que

$$|x - x^*| < \varepsilon \Rightarrow g'(x) \geq a := \frac{1 + g'(x^*)}{2} > 1.$$

Par conséquent si  $x^n \in [x^* - \varepsilon, x^* + \varepsilon]$ , on a

$$|x^{n+1} - x^*| = |g(x^n) - g(x^*)| \geq a|x^n - x^*|,$$

ce qui montre que la suite  $x^n$  tend à s'éloigner de  $x^*$  lorsque qu'elle en est suffisamment proche, et que l'algorithme du point fixe ne converge pas en général vers  $x^*$ . On aboutit à la même conclusion si  $g'(x^*) < -1$ . On dit que le point fixe  $x^*$  est *répulsif*.

**Cas 3.**  $|g'(x^*)| = 1$ . Ce cas est ambigu et il n'est pas possible de conclure sur la nature du point fixe sans examen plus détaillé. Considérons par exemple  $g(x) = \sin(x)$  dont l'unique point fixe est  $x^* = 0$  pour lequel on a  $g'(x^*) = 1$ . C'est un point fixe attractif : puisque  $|\sin(x)| < |x|$  pour tout  $x \neq 0$ , la suite  $|x^n|$  est décroissante et minorée par 0. Par conséquent elle converge, et sa limite est 0 puisque c'est le seul point fixe. Considérons d'autre part  $g(x) = \operatorname{sh}(x) = (e^x - e^{-x})/2$  dont l'unique point fixe est aussi  $x^* = 0$  et pour lequel on a aussi  $g'(x^*) = 1$ . C'est un point fixe répulsif : puisque  $|\operatorname{sh}(x)| > |x|$  pour tout  $x \neq 0$ , la suite  $x^n$  s'éloigne de 0.

**Cas 4.**  $g'(x^*) = 0$ . On sait déjà d'après le cas 1 que le point  $x^*$  est attractif et que le théorème du point fixe s'applique dans un intervalle  $F = [x^* - r, x^* + r]$ . Dans le cas où  $g$  est de classe  $\mathcal{C}^2$  sur  $I$ , on peut améliorer l'estimation de convergence géométrique. En effet, en utilisant la formule de Taylor-Lagrange au deuxième ordre on écrit, pour tout  $x \in F$ ,

$$g(x) = g(x^*) + (x - x^*)g'(x^*) + \frac{1}{2}(x - x^*)^2 g''(t) = g(x^*) + \frac{1}{2}(x - x^*)^2 g''(t),$$

avec  $t$  compris entre  $x$  et  $x^*$ . En notant  $M_2 = \max_{t \in F} |g''(t)|$ , on a donc

$$|g(x) - x^*| = |g(x) - g(x^*)| \leq \frac{M_2}{2} |x - x^*|^2.$$

et donc si  $x^0 \in F$ ,

$$\frac{M_2}{2} |x^n - x^*| \leq \left( \frac{M_2}{2} |x^{n-1} - x^*| \right)^2 \leq \left( \frac{M_2}{2} |x^{n-2} - x^*| \right)^4 \leq \dots \leq \left( \frac{M_2}{2} |x^0 - x^*| \right)^{2^n},$$

soit finalement en posant  $b := \frac{M_2 r}{2}$  et en supposante  $M_2 \neq 0$ ,

$$|x^n - x^*| \leq \frac{2}{M_2} b^{2^n}.$$

Cette estimation de convergence est dite *quadratique*. Dans le cas où  $M_2 = 0$ , on a directement  $x^n = x^*$  pour tout  $n > 0$ . La convergence quadratique est plus beaucoup rapide que la convergence géométrique, mais il faut bien sûr supposer  $b < 1$  ce qu'il est toujours possible en choisissant  $r$  suffisamment petit, où en réinitialisant la suite  $x^n$  à partir d'un indice  $k$  pour laquelle  $\frac{M_2}{2} |x^k - x^*| < 1$ . On dit dans ce cas que le point fixe  $x^*$  est *super-attractif*.

Expliquons à présent comment on peut transformer une équation  $f(x) = 0$  en un problème équivalent de point fixe  $g(x) = x$ . Le choix de  $g$  n'est évidemment pas unique : par exemple si on considère  $f(x) = x^2 - a$  avec  $a > 0$ , l'équation  $f(x) = 0$  dont la solution dans  $\mathbb{R}_+$  est  $x^* = \sqrt{a}$  est équivalente aux problèmes de point fixe  $g(x) = x$  avec

$$g(x) = x + 2(x^2 - a) \quad \text{ou} \quad g(x) = x - 3(x^2 - a) \quad \text{ou} \quad g(x) = \frac{x}{2} + \frac{a}{2x}.$$

Les deux premiers choix sont des cas particulier de la formule générale

$$g(x) := x - \tau f(x),$$

avec  $\tau \neq 0$ , dont les points fixes sont exactement les solutions de l'équation  $f(x) = 0$  quelque soit la fonction  $f$ . Soit  $x^*$  l'une de ces solutions. Afin de comprendre si l'algorithme du point fixe appliqué à la fonction  $g$  peut converger vers  $x^*$  on remarque, en supposant  $f$  de classe  $\mathcal{C}^1$ , que l'on a

$$g'(x^*) = 1 - \tau f'(x^*).$$

Si  $f'(x^*) > 0$ , un choix du paramètre  $\tau$  dans l'intervalle  $]0, \frac{2}{f'(x^*)}[$  assure donc que  $|g'(x^*)| < 1$  ce qui correspond à un point fixe attractif : l'algorithme  $x^{n+1} = g(x^n)$  converge vers  $x^*$  si le point de départ  $x^0$  en est suffisamment proche. Si  $f'(x^*) < 0$ , il faut choisir  $\tau$  dans l'intervalle  $] \frac{2}{f'(x^*)}, 0[$ . Si  $f'(x^*) = 0$  on est pas assuré de la convergence de la méthode du point fixe appliquée à la fonction  $g$ , quelque soit la valeur de  $\tau$ . Dans le cas particulier de l'exemple précédent, on a  $f'(x^*) = 2\sqrt{a} > 0$ , ce qui montre que l'algorithme du point fixe converge avec le choix  $g(x) = x - \tau(x^2 - a)$  si  $0 < \tau < \frac{1}{\sqrt{a}}$ , c'est à dire  $\tau > 0$  et  $a\tau^2 < 1$ . On peut ainsi approcher la valeur de la racine carré d'un nombre avec une machine à calculer ne possédant que l'addition et la multiplication. Le troisième choix  $g(x) = \frac{x}{2} + \frac{a}{2x}$  ne rentre pas dans le cadre ci-dessus et est particulièrement intéressant puisque l'on a alors

$$g'(x^*) = \frac{1}{2} - \frac{a}{2x^{*2}} = 0,$$

ce qui montre que le point fixe est super-attractif. On peut ainsi approcher très rapidement la valeur de la racine carrée d'un nombre avec une machine à calculer ne possédant que l'addition, la multiplication et la division.

Donnons un autre exemple pour lequel on ne connaît pas à l'avance la solution de l'équation : on cherche  $x^*$  solution de  $x^2 = e^x$ , c'est à dire tel que  $f(x) = x^2 - e^x = 0$ . Un rapide examen des variations de  $f$  indique que  $x^*$  est unique et se trouve nécessairement dans l'intervalle  $] -1, 0[$ . Sur cet intervalle, la fonction  $f'(x) = 2x - e^x$  est strictement négative et satisfait  $f'(x) \geq -3$ . Par conséquent, l'algorithme du point fixe appliqué à la fonction  $g(x) = x + \tau(x^2 - e^x)$  avec  $-\frac{2}{3} < \tau < 0$  converge vers  $x^*$  si le point de départ  $x^0$  en est suffisamment proche.

## 2.3 La méthode de Newton

La méthode de Newton est une approche systématique pour résoudre numériquement une équation  $f(x) = 0$  dans le cas où  $f$  est dérivable. On part de la remarque qu'au voisinage d'un point  $x$ , la courbe de  $f$  est proche de sa tangente d'équation

$$\tilde{f}(y) = f(x) + (y - x)f'(x),$$

et on peut tenter d'approcher un point où  $f$  s'annule par celui où la tangente s'annule, c'est à dire  $y = x - \frac{f(x)}{f'(x)}$ . En partant d'un point  $x^0$  et en itérant ce procédé, on définit la suite

$$x^{n+1} = x^n - \frac{f(x^n)}{f'(x^n)}.$$

Cette méthode itérative peut donc s'interpréter comme une méthode de point fixe appliquée à la fonction  $g(x) := x - \frac{f(x)}{f'(x)}$ . Remarquons que cette fonction n'est définie que pour les  $x$  tels que  $f'(x) \neq 0$ , et que les points fixes de  $g$  sont exactement les solutions de l'équation  $f(x) = 0$  qui vérifient  $f'(x) \neq 0$ .

**Théorème 2.3.1** *Soit  $f$  une fonction de classe  $\mathcal{C}^2$ . Si  $x^*$  est solution de l'équation  $f(x) = 0$  et est tel que  $f'(x^*) \neq 0$ , alors c'est un point fixe super-attractif de  $g$  : la méthode de Newton converge quadratiquement si  $x^0$  est choisi suffisamment proche de  $x^*$ .*

**Preuve :** Puisque  $f'$  est continue et  $f'(x^*) \neq 0$ , il existe un intervalle ouvert  $I$  contenant  $x^*$  et tel que  $|f'(x)| > a := \frac{|f'(x^*)|}{2} > 0$  pour tout  $x \in I$ . Sur cet intervalle,  $g$  est de classe  $\mathcal{C}^1$  et on a

$$g'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2},$$

d'où  $g'(x^*) = 0$ . Si  $f$  est de classe  $\mathcal{C}^3$  et donc  $g$  de classe  $\mathcal{C}^2$ , cela suffit pour montrer que  $x^*$  est un point fixe super-attractif. Si l'on suppose seulement  $f$  de classe  $\mathcal{C}^2$ , on peut montrer directement le caractère super-attractif du point  $x^*$  à l'aide de la formule de Taylor-Lagrange, en écrivant pour  $x \in I$

$$\begin{aligned} g(x) - g(x^*) &= x - x^* - \frac{f(x)}{f'(x)} \\ &= x - x^* - \frac{f(x) - f(x^*)}{f'(x)} \\ &= x - x^* - \frac{(x-x^*)f'(x) - \frac{1}{2}(x-x^*)^2 f''(\alpha)}{f'(x)} \\ &= \frac{1}{2}(x-x^*)^2 \frac{f''(\alpha)}{f'(x)}, \end{aligned}$$

avec  $\alpha \in I$ . En posant  $M_2 := \sup_{x \in I} |f''(x)|$ , on obtient pour tout  $x \in I$  l'estimation

$$|g(x) - g(x^*)| \leq \frac{M_2}{2a} |x - x^*|^2.$$

Cette estimation permet d'affirmer que  $x^*$  est super-attractif, en suivant le raisonnement du cas 4 exposé dans la section précédente. La suite  $x^n$  converge donc quadratiquement vers  $x^*$  si  $x^0$  en est suffisamment proche.  $\square$

La méthode de Newton exige de pouvoir calculer la dérivée de la fonction  $f$ , ce qui n'est pas toujours possible. Si on a uniquement accès aux valeurs de  $f$  mais pas de  $f'$ , une variante consiste à remplacer  $f'(x^n)$  par le quotient  $\frac{f(x^n) - f(x^{n-1})}{x^n - x^{n-1}}$ . C'est la méthode de la *sécante* qui s'écrit

$$x^{n+1} = x^n - \frac{f(x^n)(x^n - x^{n-1})}{f(x^n) - f(x^{n-1})}.$$

Il faut dans ce cas se donner deux points d'initialisation  $x^0$  et  $x^1$ . Remarquons que cette méthode n'est définie que si on a toujours  $f(x^n) \neq f(x^{n-1})$ .

L'analyse de cette méthode est plus délicate que celle de la méthode de Newton. On suppose à nouveau  $f$  de classe  $\mathcal{C}^2$  et on fait l'hypothèse que  $f'(x^*) \neq 0$ . Comme dans la preuve de la méthode de Newton, on remarque qu'il existe  $\delta > 0$  tel que  $|f'(x)| \geq a := |f'(x^*)|/2 > 0$  pour tout  $x$  dans l'intervalle  $F = [x^* - \delta, x^* + \delta]$ .

**Lemme 2.3.1** *En posant pour  $x, y \in F$  tel que  $x \neq y$ ,*

$$z = y - \frac{f(y)(y-x)}{f(y) - f(x)},$$

*On a l'estimation*

$$|z - x^*| \leq K |y - x^*| \max\{|x - x^*|, |y - x^*|\},$$

*où  $K = \frac{3M_2}{2a}$  avec  $M_2 := \max_{t \in F} |f''(t)|$ .*

**Preuve :** En utilisant la formule de Taylor-Lagrange à l'ordre 2, on écrit

$$\begin{aligned} z - x^* &= y - x^* - \frac{f(y)(y-x)}{f(y) - f(x)} \\ &= y - x^* - (f(y) - f(x^*)) \frac{y-x}{f(y) - f(x)} \\ &= y - x^* - \left( f'(y)(y-x^*) + \frac{1}{2} f''(s)(y-x^*)^2 \right) \frac{y-x}{f(y) - f(x)} \\ &= \left( \frac{f(y) - f(x)}{y-x} - f'(y) \right) (y-x^*) - \frac{1}{2} f''(s)(y-x^*)^2 \frac{y-x}{f(y) - f(x)} \\ &= \left( \frac{1}{2} f''(t)(x-y)(y-x^*) - \frac{1}{2} f''(s)(y-x^*)^2 \right) \frac{y-x}{f(y) - f(x)} \end{aligned}$$

avec  $s, t \in F$ . En remarquant que  $\left| \frac{y-x}{f(y)-f(x)} \right| \leq a^{-1}$ , on en déduit

$$|z - x^*| \leq \frac{M_2}{2a} |y - x^*| (|x - y| + |y - x^*|) \leq \frac{M_2}{2a} |y - x^*| (|x - x^*| + 2|y - x^*|),$$

ce qui entraîne l'estimation annoncée.  $\square$

Ce résultat entraîne immédiatement que si  $\delta$  est suffisamment petit on a aussi  $z \in F$ , et que si  $y$  est différent de  $x^*$ , alors  $|z - x^*| < |y - x^*|$  et donc  $z \neq y$ . Par conséquent, si  $x^0$  et  $x^1$  appartiennent à  $F$ , il en est de même pour toute la suite  $x^n$  et celle-ci est bien définie pour tout  $n$  (sauf si elle atteint  $x^*$  pour un  $n$  fini auquel cas il n'y a plus lieu de continuer l'algorithme). On a de plus

$$K|x^{n+1} - x^*| \leq K|x^n - x^*| \max\{K|x^n - x^*|, K|x^{n-1} - x^*|\}.$$

Par récurrence, on en déduit

$$|x^n - x^*| \leq \frac{1}{K} (K \max\{|x^0 - x^*|, |x^1 - x^*|\})^{s_n} \leq \frac{1}{K} (K\delta)^{s_n},$$

où  $s_n$  est la suite de Fibonacci définie par  $s_0 = s_1 = 1$  et  $s_{n+1} = s_n + s_{n-1}$ . La suite de Fibonacci est asymptotiquement proportionnelle à  $c^n$  où  $c := \frac{1+\sqrt{5}}{2}$  est le "nombre d'or". Par conséquent la vitesse de convergence est très rapide (mais moins que celle de la méthode de Newton).

## 2.4 Le cas des fonctions vectorielles

Dans cette section, nous généralisons les résultats des deux sections précédentes aux fonctions vectorielles.

Soit  $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$  une telle fonction et  $(g_1, \dots, g_n)$  ses  $n$  composantes qui sont chacune des fonctions de  $\mathbb{R}^n$  dans  $\mathbb{R}$ . On rappelle que  $g$  est de classe  $\mathcal{C}^1$  sur un ouvert  $U \subset \mathbb{R}^n$  si et seulement si tous les  $g_i$  le sont, ce qui signifie que les dérivées partielles  $\frac{\partial g_i}{\partial x_j}$  sont définies et continues sur  $U$ . La différentielle de  $g$  au point  $f$  est une application linéaire  $dg_x \in \mathcal{L}(\mathbb{R}^n)$  dont la matrice dans la base canonique de  $\mathbb{R}^n$  est

$$dg_x = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \dots & \frac{\partial g_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial g_n}{\partial x_1} & \dots & \frac{\partial g_n}{\partial x_n} \end{pmatrix},$$

On rappelle que développement limité de  $g$  au premier ordre en  $x \in U$  s'écrit pour tout  $y \in U$

$$g(y) = g(x) + dg_x(y - x) + \|x - y\|\varepsilon(x - y),$$

où  $\varepsilon$  est une fonction vectorielle telle que  $\lim_{h \rightarrow 0} \varepsilon(h) = 0$ . Le résultat suivant est une généralisation de la Proposition 2.4.1.

**Proposition 2.4.1** *Soit  $\|\cdot\|$  une norme sur  $\mathbb{R}^n$  et  $g$  une fonction de classe  $\mathcal{C}^1$  sur un ouvert  $U \subset \mathbb{R}^n$  et à valeur dans  $\mathbb{R}^n$ . Si  $g$  est  $a$ -lipschitzienne sur  $U$  pour cette norme - c'est à dire  $\|g(x) - g(y)\| \leq a\|x - y\|$  pour tout  $x, y \in U$  - on a alors  $\|dg_x\| \leq a$  pour la norme matricielle subordonnée à la norme  $\|\cdot\|$ , en tout  $x \in U$ . Réciproquement, si  $\|dg_x\| \leq a$  pour tout  $x \in V$  où  $V \subset U$  est convexe, alors  $g$  est  $a$ -lipschitzienne sur  $V$  pour la norme  $\|\cdot\|$ .*

**Preuve :** Supposons que  $g$  est  $a$ -lipschitzienne sur  $U$  pour la norme  $\|\cdot\|$ . Soit  $x \in U$  et  $v \in \mathbb{R}^n$ . Pour  $t \in \mathbb{R}$  suffisamment petit  $x + tv$  appartient à  $U$  et on a

$$g(x + tv) = g(x) + tdg_x v + |t|\|v\|\varepsilon(tv),$$

ce qui entraîne

$$dg_x v = \lim_{t \rightarrow 0} \frac{g(x + tv) - g(x)}{t}.$$

Puisque  $\|g(x+tv) - g(x)\| \leq a|t|\|v\|$ , on en déduit  $\|dg_x v\| \leq a\|v\|$ . Comme ceci est vrai pour tout  $v \in \mathbb{R}^n$ , on en déduit

$$\|dg_x\| = \max_{v \in \mathbb{R}^n, v \neq 0} \frac{\|dg_x v\|}{\|v\|} \leq a.$$

Réciproquement supposons  $\|dg_x\| \leq a$  pour tout  $x \in V \subset U$  et  $V$  est convexe. Pour tout  $y \in V$  et  $t \in [0, 1]$ , on a  $x + t(y-x) \in V$  et on peut ainsi définir l'application  $h_{x,y}(t) := g(x + t(y-x))$  qui va de  $[0, 1]$  dans  $\mathbb{R}^n$ . On remarque que  $h_{x,y}(0) = g(x)$  et  $h_{x,y}(1) = g(y)$  et par conséquent

$$g(y) - g(x) = h(1) - h(0) = \int_0^1 h'_{x,y}(t) dt.$$

Par composition des différentielles on a  $h'_{x,y}(t) = dg_{x+t(y-x)}(y-x)$ , et on en déduit

$$\|g(y) - g(x)\| = \left\| \int_0^1 dg_{x+t(y-x)}(y-x) dt \right\| \leq \int_0^1 \|dg_{x+t(y-x)}(y-x)\| dt,$$

où on a utilisé l'inégalité  $\left\| \int_a^b h(t) dt \right\| \leq \int_a^b \|h(t)\| dt$  qui est valable pour toute norme (on peut la démontrer d'abord sur les sommes de Riemann et passer à la limite). En remarquant que

$$\|dg_{x+t(y-x)}(y-x)\| \leq \|dg_{x+t(y-x)}\| \|y-x\| \leq a\|y-x\|,$$

on en déduit que  $\|g(y) - g(x)\| \leq a\|y-x\|$ . □

Nous allons utiliser ce résultat pour étudier la méthode du point fixe. Soit  $g$  une fonction de classe  $\mathcal{C}^1$  sur un ouvert  $U \subset \mathbb{R}^n$  et à valeur dans  $\mathbb{R}^n$  et soit  $x^* \in U$  un point fixe de  $g$ . Comme dans le cas des fonctions réelles, on peut distinguer plusieurs cas selon les propriétés de  $dg_{x^*}$ , et plus précisément de son rayon spectral.

**Cas 1.**  $\rho(dg_{x^*}) < 1$ . D'après la proposition 1.3.5, pour tout  $\varepsilon > 0$  il existe une norme subordonnée à une norme  $\|\cdot\|$  sur  $\mathbb{C}^n$  telle que

$$\|dg_{x^*}\| \leq \rho(dg_{x^*}) + \varepsilon,$$

et en choisissant  $\varepsilon$  suffisamment petit on a donc

$$\|dg_{x^*}\| < 1.$$

Par continuité de  $dg_{x^*}$ , il existe  $r > 0$  tel que pour tout  $x$  tel que  $\|x - x^*\| \leq r$ , on a  $x \in U$  et

$$\|dg_x\| \leq a := \frac{1 + \|dg_{x^*}\|}{2} < 1.$$

D'après la proposition 2.4.1, ceci entraîne que  $g$  est  $a$ -lipschitzienne (donc contractante) sur la boule fermée  $F = B(x^*, r)$ . On a d'autre part

$$\|x - x^*\| \leq r \Rightarrow \|g(x) - x^*\| \leq a\|x - x^*\| \leq ar \leq r,$$

ce qui montre que  $g(F) \subset F$ . Le théorème du point fixe s'applique sur  $F$  : pour tout  $x^0 \in F$ , la suite  $x^{k+1} = g(x^k)$  converge vers  $x^*$  avec la vitesse géométrique

$$\|x^k - x^*\| \leq ra^k$$

On dit que le point fixe  $x^*$  est attractif.

**Remarque 2.4.1** Comme toutes les normes sont équivalentes, la suite  $x^n$  converge vers  $x^*$  dans n'importe quelle norme. La condition  $\|dg_{x^*}\| < 1$  pour une norme subordonnée est évidemment suffisante pour que  $x^*$  soit attractif, puisqu'elle entraîne  $\rho(dg_{x^*}) < 1$  d'après la proposition 1.3.4, mais elle n'est pas nécessaire : on peut avoir  $\|dg_{x^*}\| > 1$  pour certaines normes subordonnées et néanmoins  $\rho(dg_{x^*}) < 1$  ce qui signifie que le point fixe est attractif.

**Cas 2.**  $\rho(dg_{x^*}) > 1$ . Dans ce cas, il n'est pas possible de montrer la convergence de la méthode du point fixe et  $x^*$  est dit répulsif. Notons cependant que la méthode du point fixe peut converger pour certains choix particuliers de  $x^0$ . Par exemple si  $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$  est définie par  $g(u, v) = (2u, \frac{v}{2})$ , l'unique point fixe est  $x^* = 0$  et on a  $\rho(dg_0) = 2 > 1$ . Cependant on voit que l'algorithme du point fixe converge vers  $x^*$  si l'on part d'un point du type  $x^0 = (0, v)$  mais diverge pour tout autre point de départ.

**Cas 3.**  $\rho(dg_{x^*}) = 1$ . Comme pour les fonctions réelles, ce cas est ambigu et on ne peut pas conclure sur la convergence de la méthode du point fixe sans une étude plus spécifique de la fonction  $g$ .

**Cas 4.**  $\rho(dg_{x^*}) = 0$ . On peut dans ce cas établir le caractère super-attractif du point  $x^*$  si  $g$  est de classe  $\mathcal{C}^2$ . On sait déjà d'après le cas 1 que le théorème du point fixe s'applique dans une boule  $F = B(x^*, r)$  et on prend donc  $x^0 \in F$ . Nous supposons d'abord pour simplifier que  $dg_{x^*} = 0$ , ce qui est équivalent à  $\nabla g_i(x^*) = 0$  pour toutes les composantes  $g_i$  de  $g$ . Rappelons que le développement à l'ordre 2 d'une fonction  $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$  autour d'un point  $z$  a la forme générale

$$\varphi(z+h) = \varphi(z) + \langle \nabla \varphi(z), h \rangle + \frac{1}{2} \langle d^2 \varphi_z h, h \rangle + \|h\|^2 \varepsilon(h),$$

où

$$d^2 \varphi_z = \left( \frac{\partial^2 \varphi}{\partial x_i \partial x_j} (z) \right)_{i,j=1,\dots,n}$$

est la matrice des dérivées secondes ou *hessienne* au point  $z$ , et  $\varepsilon : \mathbb{R}^n \rightarrow \mathbb{R}$  est telle que  $\lim_{h \rightarrow 0} \varepsilon(h) = 0$ . En particulier, si l'on fixe  $z, h \in \mathbb{R}^n$ , la fonction  $\varphi_{z,h} : \mathbb{R} \rightarrow \mathbb{R}$  définie par  $\varphi_{z,h}(t) := \varphi(z+ht)$  vérifie

$$\varphi'_{z,h}(t) = \langle \nabla \varphi(z+ht), h \rangle \text{ et } \varphi''_{z,h}(t) = \langle d^2 \varphi_{z+ht} h, h \rangle.$$

Le développement de Taylor-Lagrange à l'ordre 2 de cette fonction entre  $t = 0$  et  $t = 1$ , s'écrit donc

$$\varphi(z+h) = \varphi(z) + \langle \nabla \varphi(z), h \rangle + \frac{1}{2} \langle d^2 \varphi_{z+hs} h, h \rangle,$$

avec  $s \in [0, 1]$ . En appliquant ceci aux fonctions  $g_k$ , avec  $z = x^*$  et  $h = x - x^*$  pour  $x \in F$ , on obtient

$$g_k(x) - x_k^* = g_k(x) - g_k(x^*) = \frac{1}{2} \sum_{i,j=1,\dots,n} \frac{\partial g_k}{\partial x_i \partial x_j} (x^* + s_k(x - x^*)) (x_i - x_i^*) (x_j - x_j^*),$$

avec  $s_k \in [0, 1]$ , en notant  $x_i$  et  $x_i^*$  les  $i$ -èmes coordonnées de  $x$  et  $x^*$ . En notant

$$M_2 := \max_{y \in F} \max_{k=1,\dots,n} \sum_{i,j=1,\dots,n} \left| \frac{\partial^2 g_k}{\partial x_i \partial x_j} (y) \right|,$$

on obtient ainsi

$$|g_k(x) - x_k^*| \leq \frac{M_2}{2} \left( \max_{i=1,\dots,n} |x_i - x_i^*| \right)^2,$$

pour tout  $k = 1, \dots, n$  ce qui est équivalent à

$$\|g(x) - x^*\|_\infty \leq \frac{M_2}{2} \|x - x^*\|_\infty^2$$

En raisonnant alors comme dans le cas 4 pour les fonctions réelles, on obtient l'estimation de convergence quadratique

$$\|x^k - x^*\|_\infty \leq \frac{2}{M_2} b^{2^k},$$

avec  $b := \frac{M_2 \tilde{r}}{2}$  et  $\tilde{r} := \max_{y \in F} \|y - x^*\|_\infty$ . Dans le cas où  $\rho(dg_{x^*}) = 0$  mais  $dg_{x^*} \neq 0$ , il faut travailler un peu plus pour aboutir à une estimation de ce type. On définit les fonctions itérées de  $g$ , en posant  $g^{[1]} = g$  et  $g^{[k+1]} = g \circ g^{[k]}$ , et l'on remarque que l'on a  $g^{[k]}(x^*) = x^*$  et par la règle de composition des différentielles

$$dg_{x^*}^{[k]} = (dg_{x^*})^k.$$

Comme toutes les valeurs propres de  $dg_{x^*}$  sont nulles, on sait que  $dg_{x^*}$  est semblable à une matrice triangulaire supérieure  $T$  qui n'a que des 0 sur sa diagonale. Il est facile de montrer qu'une telle matrice  $n \times n$  vérifie  $T^n = 0$ . Comme  $dg_{x^*} = P^{-1}TP$ , on a donc

$$dg_{x^*}^{[n]} = (dg_{x^*})^n = (P^{-1}TP)^n = P^{-1}T^nP = 0.$$

Le point fixe  $x^*$  est donc super-attractif pour la méthode du point fixe appliqué à la fonction  $g^{[n]}$ , ce qui signifie que l'on a une estimation du type

$$\|x^{ln} - x^*\|_\infty \leq \frac{2}{M_2} b^{2^l},$$

avec

$$M_2 := \max_{y \in F} \max_{k=1, \dots, n} \sum_{i,j=1, \dots, n} \left| \frac{\partial^2 g_k^{[n]}}{\partial x_i \partial x_j}(y) \right|,$$

$b := \frac{M_2 \tilde{r}}{2}$  et  $\tilde{r} := \max_{y \in F} \|y - x^*\|_\infty$ . Pour  $k$  multiple de  $n$  on a donc l'estimation de convergence

$$\|x^k - x^*\|_\infty \leq \frac{2}{M_2} b^{2^{k/n}},$$

En utilisant le fait que les fonctions  $g^{[l]}$  sont lipschitziennes sur  $F$ , il est facile d'en déduire une estimation du même type pour toutes les valeurs de  $k$ .

La méthode de Newton introduite dans la section §2.3 pour résoudre  $f(x) = 0$  peut aussi être généralisée au cas des fonctions vectorielles : elle repose à nouveau sur l'idée de remplacer  $f$  au voisinage de  $x$  par la 'fonction tangente'

$$\tilde{f}(y) = f(x) + df_x(y - x),$$

qui s'annule au point  $y = x - df_x^{-1}(f(x))$ . On pose donc

$$x^{n+1} = x^n - df_{x^n}^{-1}(f(x^n)).$$

C'est la méthode de Newton-Raphson qui peut être vue comme une méthode de point fixe pour la fonction

$$g(x) := x - df_x^{-1}(f(x)).$$

Cette méthode nécessite que  $df_{x^n}$  soit inversible pour tous les  $x^n$  apparaissant dans la suite. A chaque étape, le calcul de  $df_{x^n}^{-1}(f(x^n))$  revient à résoudre un système linéaire  $n \times n$ .

**Théorème 2.4.1** *Soit  $f$  une fonction de classe  $C^2$ . Si  $x^*$  est solution de l'équation  $f(x) = 0$  et est tel que  $df_{x^*}$  est inversible, alors c'est un point fixe super-attractif de  $g$ . La méthode de Newton-Raphson converge donc quadratiquement vers  $x^*$  si  $x^0$  en est suffisamment proche.*

**Preuve :** Grâce à la continuité de  $x \mapsto df_x$  et donc de  $x \mapsto \det(df_x)$ , il existe  $r > 0$  tel que

$$\|x - x^*\| \leq r \Rightarrow \det(df_x) \neq 0,$$

c'est à dire  $df_x$  est inversible sur la boule  $F = B(x^*, r)$ . En utilisant les formule de Cramer, on voit que  $x \mapsto df_x^{-1}$  est aussi continue sur  $F$  et on note

$$K = \max_{x \in F} \|df_x^{-1}\|_\infty$$

Pour tout  $x \in F$ , on peut écrire

$$g(x) - g(x^*) = x - x^* - df_x^{-1}(f(x)) = x - x^* - df_x^{-1}(f(x) - f(x^*))$$

On rappelle le développement limité de Taylor-Lagrange pour une fonction  $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$ , déjà utilisé dans le cas 4 de l'étude de la méthode du point fixe :

$$\varphi(z + h) = \varphi(z) + \langle \nabla \varphi(z), h \rangle + \frac{1}{2} \langle d^2 \varphi_{z+hs} h, h \rangle,$$

avec  $s \in [0, 1]$ . En appliquant ceci à chaque composante  $f_k$  de  $f$  avec  $z = x$  et  $h = x^* - x$ , on obtient

$$f(x) - f(x^*) = df_x(x^* - x) + \delta.$$

où la  $k$ -ème coordonnée du reste  $\delta$  est

$$\delta_k = \frac{1}{2} \sum_{i,j=1,\dots,n} \frac{\partial^2 f_k}{\partial x_i \partial x_j}(x + s_k(x^* - x))(x_i - x_i^*)(x_j - x_j^*),$$

avec  $s_k \in [0, 1]$ . On obtient ainsi

$$g(x) - g(x^*) = df_x^{-1} \delta,$$

et donc

$$\|g(x) - g(x^*)\|_\infty \leq K \|\delta\|_\infty \leq \frac{KM_2}{2} \|x - x^*\|_\infty^2,$$

où

$$M_2 := \max_{y \in F} \max_{k=1,\dots,n} \sum_{i,j=1,\dots,n} \left| \frac{\partial^2 f_k}{\partial x_i \partial x_j}(y) \right|,$$

Cette estimation permet d'affirmer que  $x^*$  est super-attractif, en suivant le raisonnement du cas 4 de l'étude de la méthode du point fixe pour les fonctions réelles. La suite  $x^n$  converge donc quadratiquement vers  $x^*$  si  $x^0$  en est suffisamment proche.  $\square$

### 3 Approximation des fonctions

Une fonction  $f$  arbitraire définie sur un intervalle  $I$  et à valeur dans  $\mathbb{R}$  peut être représentée par son graphe, ou de manière équivalente par la donnée de l'ensemble de ses valeurs  $f(t)$  pour  $t \in I$ . Ces valeurs sont en nombre infini et il n'est donc pas possible en pratique de les mettre en mémoire sur un ordinateur. On peut alors chercher à remplacer  $f$  par une fonction  $g$  plus simple qui est proche de  $f$  et dépend d'un nombre fini  $n$  de paramètres que l'on peut ainsi mettre en mémoire. Un exemple consiste à choisir  $g$  dans l'ensemble des polynômes de degré  $n - 1$  : on peut alors caractériser  $g$  par ses  $n$  coefficients. Plus généralement, on peut chercher à approcher  $f$  par une fonction  $g$  appartenant à un espace de fonction  $E_n$  de dimension  $n$ . On peut enfin chercher à approcher  $f$  à partir de la donnée de ses valeurs en  $n$  points  $x_1, x_2, \dots, x_n$ . Intuitivement, on approche de mieux en mieux  $f$  lorsque la quantité  $n$  d'information augmente.

La *théorie de l'approximation* étudie de façon rigoureuse le compromis entre la *complexité* donnée par le nombre de paramètres  $n$  et la *précision* que l'on peut obtenir entre  $f$  et  $g$ . Elle s'intéresse aussi à la manière dont on construit en pratique l'approximation  $g$  à partir de  $f$ . Dans ce chapitre, on s'intéresse principalement à l'approximation des fonction par des polynômes, après avoir fait quelques rappels sur les séries trigonométriques. On termine par l'approximation polynomiale par morceaux qui est la plus souvent utilisée en pratique. On considère ici uniquement l'approximation de fonctions d'une seule variable réelle, la généralisation à l'approximation des fonctions à plusieurs variables dépassant le cadre de ce cours.

Afin de mesurer la distance entre deux fonctions définies sur un intervalle  $I$ , on introduit la norme dite  $L^\infty$  ou *norme sup* sur  $I$  : si  $f$  est définie sur  $I$  et à valeur dans  $\mathbb{K} = \mathbb{R}$  ou  $\mathbb{C}$ , on pose

$$\|f\| = \|f\|_\infty := \sup_{x \in I} |f(x)|.$$

Cette norme est aussi appelée norme de la convergence uniforme : en effet une suite de fonctions  $f_n$  converge uniformément vers  $f$  sur  $I$  si et seulement si  $\lim_{n \rightarrow +\infty} \|f - f_n\| = 0$ . On rappelle que l'espace  $\mathcal{C}(I)$  des fonctions continues sur  $I$  est complet pour cette norme. L'erreur de meilleure approximation d'une fonction  $f$  dans un espace de fonctions  $E_n$  est définie par la quantité

$$\inf_{g \in E_n} \|f - g\|,$$

et décrit la qualité de l'approximation de  $f$  par les éléments de  $E_n$ . On s'intéresse en particulier à la décroissance de cette erreur lorsque  $n$  augmente.

#### 3.1 Approximation par les séries trigonométriques

Commençons par quelques rappels sur les séries trigonométriques, qui sont aussi appelée séries de Fourier. Il s'agit de séries de fonctions de la forme

$$\sum_{k \geq 0} a_k \cos(kx) + \sum_{k > 0} b_k \sin(kx),$$

que l'on peut aussi mettre sous la forme

$$\sum_{k \in \mathbb{Z}} c_k e^{ikx},$$

en posant  $c_0 = a_0$  et  $c_{\pm k} = \frac{1}{2}(a_k \pm ib_k)$  pour  $k > 0$ . Lorsque ces séries convergent, leur limites sont des fonctions de période  $2\pi$ , puisque chacun de leur termes possèdent cette propriété.

On appelle *polynôme trigonométrique* de degré  $n$  une fonction du type

$$g(x) = \sum_{|k| \leq n} c_k e^{ikx} = \sum_{0 \leq k \leq n} a_k \cos(kx) + \sum_{0 < k \leq n} b_k \sin(kx).$$

L'ensemble  $\mathcal{T}_n$  des polynomes trigonométriques de degré  $n$  est donc l'espace vectoriel engendré par les fonctions  $e_k : x \mapsto e^{ikx}$  pour  $|k| \leq n$ . On peut montrer que ces fonctions sont indépendantes : on part de la remarque que

$$\int_{-\pi}^{\pi} e_k(x) \overline{e_l(x)} dx = \int_{-\pi}^{\pi} e^{i(k-l)x} dx = 2\pi \text{ si } k = l \text{ et } 0 \text{ si } k \neq l.$$

Par conséquent, si  $\sum_{|k| \leq n} c_k e_k = 0$ , on a

$$0 = \int_{-\pi}^{\pi} \left( \sum_{|k| \leq n} c_k e_k(x) \right) \overline{e_l(x)} dx = c_l.$$

L'espace  $\mathcal{T}_n$  est donc de dimension  $2n + 1$ .

Le problème fondamental de la représentation d'une fonction arbitraire  $2\pi$ -périodique sous la forme d'une série de Fourier est un problème d'approximation : chercher à écrire  $f$  sous la forme d'une série uniformément convergente

$$f(x) = \sum_{n \in \mathbb{Z}} c_n e_n(x),$$

signifie que l'on cherche à approcher  $f$  par la suite de polynômes trigonométriques  $\sum_{|k| \leq n} c_k e_k \in \mathcal{T}_n$ . Si une telle convergence est vérifiée, alors en multipliant l'identité ci-dessus par  $\overline{e_l(x)}$  et en intégrant sur  $[-\pi, \pi]$ , on trouve que le coefficient  $c_l$  dépend de  $f$  suivant la forme

$$c_l = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-ilx} dx.$$

Les  $c_l$  sont souvent notés  $c_l(f)$  et appelés "coefficients de Fourier" de  $f$ . Le polynôme trigonométrique

$$S_n f(x) = \sum_{|k| \leq n} c_k(f) e_k(x),$$

est appelé "somme partielle de Fourier" de  $f$ . Nous rappelons une version simple du *théorème de Dirichlet* qui donne une condition suffisante pour la convergence simple de  $S_n f$  vers  $f$ .

**Théorème 3.1.1** (de Dirichlet). *Soit  $f$  une fonction  $2\pi$  périodique et continue, telle qu'en un point  $x$  il existe une dérivée à gauche et à droite. Alors*

$$\lim_{n \rightarrow +\infty} S_n f(x) = f(x).$$

Le théorème de Dirichlet n'est pas satisfaisant du point de vue numérique car il ne donne aucune estimation sur la façon dont l'erreur  $\|f - S_n f\|$  décroît en fonction de  $n$  (ici  $\|\cdot\|$  désigne la norme sup sur  $\mathbb{R}$ , qui coïncide avec celle sur  $[-\pi, \pi]$  puisque l'on considère des fonctions  $2\pi$ -périodiques). Il est possible d'obtenir de telles estimations si on fait des hypothèses supplémentaires portant sur la *régularité* de  $f$ . En effet, si  $f$  est une fonction  $2\pi$  périodique de classe  $C^1$  sur  $\mathbb{R}$ , une intégration par partie permet d'obtenir

$$c_k(f) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-ikx} dx = \frac{1}{2\pi ik} \int_{-\pi}^{\pi} f'(x) e^{-ikx} dx = \frac{1}{2\pi ik} c_k(f'),$$

et par conséquent

$$|c_k(f)| \leq \frac{1}{2\pi|k|} \int_{-\pi}^{\pi} |f'(x)| dx.$$

Notons que cette estimation reste valide pour des fonctions dont la dérivée n'est pas nécessairement continue mais de valeur absolue intégrable sur  $[-\pi, \pi]$  (au sens où  $f$  est la primitive d'une fonction de valeur absolue intégrable). En itérant l'intégration par partie si  $f$  est suffisamment régulière, on trouve

$$c_k(f) = \frac{1}{2\pi(ik)^{m+1}} c_k(f^{(m+1)}),$$

et par conséquent

$$|c_k(f)| \leq \frac{1}{2\pi|k|^{m+1}} \int_{-\pi}^{\pi} |f^{(m+1)}(x)| dx.$$

Si  $m > 1$ , on en déduit l'estimation

$$\begin{aligned}
\|f - S_n f\| &= \left\| \sum_{|k|>n} c_k(f) e_k \right\| \\
&\leq \sum_{|k|>n} \|c_k(f) e_k\| \\
&= \sum_{|k|>n} |c_k(f)| \\
&\leq \frac{\int_{-\pi}^{\pi} |f^{(m+1)}(x)| dx}{m\pi} \sum_{k>n} |k|^{-(m+1)} \\
&\leq \frac{\int_{-\pi}^{\pi} |f^{(m+1)}(x)| dx}{m\pi} n^{-m},
\end{aligned}$$

où l'on a utilisé l'estimation  $\sum_{k>n} |k|^{-(m+1)} \leq \int_n^{+\infty} t^{-(m+1)} dt \leq \frac{n^{-m}}{m}$ . On a donc prouvé le résultat suivant.

**Théorème 3.1.2** *Soit  $m \geq 2$  et  $f$  une fonction  $2\pi$  périodique de classe  $C^m$  telle que  $f^{(m+1)}$  est intégrable sur  $[-\pi, \pi]$ . On a alors l'estimation*

$$\|f - S_n f\| \leq C n^{-m},$$

avec  $C := \frac{\int_{-\pi}^{\pi} |f^{(m+1)}(x)| dx}{m\pi}$ .

Lorsque  $f$  est seulement supposée continue sur  $\mathbb{R}$ , le théorème de Dirichlet ne permet pas d'affirmer que  $S_n f$  converge uniformément vers  $f$ . On sait en fait depuis le XIXème siècle que l'on peut trouver des fonctions  $f$  continues sur  $\mathbb{R}$  et  $2\pi$ -périodiques telles qu'en certains points  $x \in \mathbb{R}$  la série de Fourier  $S_n f(x)$  diverge quand  $n \rightarrow +\infty$ .

Il est cependant possible d'approcher les fonctions continues par des polynômes trigonométriques qui diffèrent de  $S_n f$ . Cela provient du résultat suivant dû à Jackson et dont la démonstration est difficile.

**Théorème 3.1.3** (de Jackson). *Il existe une constante  $C_1 > 0$ , telle que si  $f$  est une fonction  $2\pi$  périodique et continue, on a pour tout  $n > 0$ ,*

$$\inf_{g \in \mathcal{T}_n} \|f - g\| \leq C_1 \omega(f, \frac{1}{n}),$$

où  $\omega(f, t) := \max_{|x-y| \leq t} |f(x) - f(y)|$  pour tout  $t > 0$ .

Une esquisse de la preuve de ce théorème est donnée à la fin de cette section. La quantité  $\omega(f, t)$  est appelée *module de continuité* de  $f$ . Si  $f$  est une fonction  $2\pi$ -périodique continue sur  $\mathbb{R}$  on vérifie aisément qu'elle est aussi uniformément continue, et on a par conséquent

$$\lim_{t \rightarrow 0} \omega(f, t) = 0.$$

Le théorème 3.1.3 entraîne donc le corollaire suivant.

**Corollaire 3.1.1** *Si  $f$  est  $2\pi$ -périodique et continue sur  $\mathbb{R}$ , alors*

$$\lim_{n \rightarrow +\infty} \inf_{g \in \mathcal{T}_n} \|f - g\| = 0.$$

*Autrement dit, il existe une suite  $(f_n)_{n \geq 0}$  de polynômes trigonométriques  $f_n \in \mathcal{T}_n$  qui converge uniformément vers  $f$ .*

Si la fonction  $f$  est de surcroît  $M$ -Lipschitzienne sur  $\mathbb{R}$ , on a par définition

$$\omega(f, t) \leq Mt.$$

En particulier, si  $f$  est de classe  $\mathcal{C}^1$ , le théorème des accroissements finis montre que qu'elle est  $M$ -Lipschitzienne avec  $M = \|f'\|$ . Le théorème 3.1.3 entraîne donc le corollaire suivant qui précise la vitesse de convergence de la meilleure approximation par les polynômes trigonométriques.

**Corollaire 3.1.2** Si  $f$  est  $2\pi$ -périodique et  $M$ -Lipschitzienne sur  $\mathbb{R}$ , alors

$$\inf_{g \in \mathcal{T}_n} \|f - g\| \leq C_1 M n^{-1}.$$

En particulier si  $f$  est de classe  $\mathcal{C}^1$  on a

$$\inf_{g \in \mathcal{T}_n} \|f - g\| \leq C_1 \|f'\| n^{-1}.$$

Il est possible de généraliser ces résultats sous la forme du théorème suivant que l'on admet ici.

**Théorème 3.1.4** Si  $f$  est  $2\pi$ -périodique, de classe  $\mathcal{C}^{m-1}$  et telle que  $f^{(m-1)}$  est  $M$ -Lipschitzienne sur  $\mathbb{R}$ , alors

$$\inf_{g \in \mathcal{T}_n} \|f - g\| \leq C_m M n^{-m},$$

où la constante  $C_m$  est indépendante de  $n$  et  $f$ . En particulier si  $f$  est de classe  $\mathcal{C}^m$  on a

$$\inf_{g \in \mathcal{T}_n} \|f - g\| \leq C_m \|f^{(m)}\| n^{-m}.$$

Par comparaison au théorème 3.1.2, on note que la condition “ $f \in \mathcal{C}^{m-1}$  et  $f^{(m-1)}$  est  $M$ -Lipschitzienne sur  $\mathbb{R}$ ” est plus faible que “ $f \in \mathcal{C}^m$  et  $f^{(m+1)}$  intégrable”, l'idée principale restant que la vitesse de convergence du procédé d'approximation est liée à la régularité de la fonction  $f$ .

Nous terminons cette section par quelques indications qui permettront aux étudiants les plus audacieux d'établir la preuve du Théorème 3.1.3. Pour tout  $n > 0$  on note  $p$  la partie entière de  $n/4$  et on définit sur  $\mathbb{R}$  la fonction

$$J_n(x) := \lambda_n \left( \frac{\sin((2p+1)x/2)}{\sin(x/2)} \right)^4,$$

prolongée par continuité par  $\lambda_n(2p+1)^4$  en  $x = 2k\pi$  pour  $k \in \mathbb{Z}$ , et où  $\lambda_n > 0$  est choisi tel que  $\int_{-\pi}^{\pi} J_n(x) dx = 1$ . On établit facilement l'identité

$$\sum_{k=-p}^p e^{ikx} = \frac{\sin((2p+1)x/2)}{\sin(x/2)}$$

qui entraîne que  $J_n$  est de la forme

$$J_n(x) = \sum_{k=-4p}^{4p} c_k e^{ikx},$$

avec  $c_k = c_{-k}$  et par conséquent  $J_n \in \mathcal{T}_n$ . On définit alors la fonction

$$f_n(x) := \int_{-\pi}^{\pi} J_n(y) f(x-y) dy.$$

Par changement de variable et en utilisant la périodicité de  $f$  et  $J_n$ , on obtient

$$f_n(x) = \int_{-\pi}^{\pi} J_n(x-y) f(y) dy = \sum_{k=-4p}^{4p} c_k \left( \int_{-\pi}^{\pi} f(y) e^{-iky} dy \right) e^{ikx},$$

Par conséquent  $f_n \in \mathcal{T}_n$ . On étudie l'erreur entre  $f_n$  et  $f$  en écrivant pour tout  $x \in [-\pi, \pi]$

$$|f(x) - f_n(x)| = \left| \int_{-\pi}^{\pi} J_n(y) (f(x) - f(x+y)) dy \right| \leq \int_{-\pi}^{\pi} J_n(y) |f(x) - f(x+y)| dy,$$

où on a utilisé le fait que  $\int_{-\pi}^{\pi} J_n(y) dy = 1$ . On établit facilement

$$|f(x) - f(x+y)| \leq (1 + n|y|) \omega\left(f, \frac{1}{n}\right),$$

et ceci entraîne

$$\|f - f_n\| \leq \omega\left(f, \frac{1}{n}\right) \left( \int_{-\pi}^{\pi} (1 + n|y|) J_n(y) dy \right).$$

On conclut la preuve en établissant que l'intégrale figurant à droite de cette inégalité est bornée par une constante  $C_1$  indépendante de  $n > 0$ .

### 3.2 Approximation polynomiale

On s'intéresse à présent à l'approximation des fonctions par des polynômes. On note

$$\mathcal{P}_n := \text{Vect}\{x \mapsto x^k ; k = 0, 1, \dots, n\},$$

l'espace des polynômes de degré  $n$ . Approcher une fonction  $f$  par un polynôme de degré  $n$  est un procédé très classique en analyse que l'on rencontre par exemple lorsque l'on effectue fait un développement limité de  $f$  en un point  $x_0$ ,

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \dots + \frac{f^{(n)}(x_0)}{n!} (x - x_0)^n := p_n(x), \text{ avec } p_n \in \mathcal{P}_n.$$

L'inconvénient du développement de Taylor est que si on souhaite faire tendre son degré  $n$  vers  $+\infty$ , la convergence de la suite  $p_n$  vers la fonction  $f$  sur un intervalle  $[a, b]$  exige que celle-ci soit développable en série entière dans un voisinage du point  $x_0$  contenant  $[a, b]$ . C'est une hypothèse très forte puisqu'elle entraîne en particulier que  $f$  est de classe  $\mathcal{C}^\infty$  sur cet intervalle. Il est donc légitime de rechercher d'autres procédés d'approximation par des polynômes qui convergent lorsque la fonction  $f$  est moins régulière, par exemple simplement continue.

Un résultat fondamental dû à Weierstrass affirme que toute fonction continue peut être approchée par une suite de polynômes.

**Théorème 3.2.1** (de Weierstrass). *Si  $f$  est continue sur  $I = [a, b]$ , alors*

$$\lim_{n \rightarrow +\infty} \inf_{g \in \mathcal{P}_n} \|f - g\| = 0.$$

*Autrement dit, il existe une suite  $(f_n)_{n \geq 0}$  de polynômes  $f_n \in \mathcal{P}_n$  qui converge uniformément vers  $f$ .*

**Preuve :** On remarque qu'on peut toujours se ramener au cas  $I = [-1, 1]$  en utilisant le changement de variable affine  $\phi(x) = a + \frac{1}{2}(b - a)(x + 1)$  qui envoie  $[-1, 1]$  sur  $[a, b]$ . En effet si  $f$  est continue sur  $[a, b]$  alors la fonction  $f \circ \phi$  est continue sur  $[-1, 1]$ . Si on peut approcher  $f \circ \phi$  uniformément sur  $[-1, 1]$  par une suite de polynômes  $g_n \in \mathcal{P}_n$ , alors les fonctions  $f_n = g_n \circ \phi^{-1}$  sont aussi dans  $\mathcal{P}_n$  puisque  $\phi^{-1}$  est affine et elles approchent uniformément  $f$  sur  $[a, b]$  :

$$\|f - f_n\| = \max_{y \in [a, b]} |f(y) - f_n(y)| = \max_{x \in [-1, 1]} |f(\phi(x)) - f_n(\phi(x))| = \|f \circ \phi - g_n\|.$$

On suppose donc à présent que  $f$  est une fonction continue sur  $I = [-1, 1]$ . On effectue un nouveau changement de variable en posant pour tout  $t \in \mathbb{R}$ ,

$$F(t) = f(\cos(t)).$$

La fonction  $F$  est continue et  $2\pi$ -périodique. D'après le Corollaire 3.1.1, il existe une suite de polynômes trigonométrique  $F_n \in \mathcal{T}_n$  qui converge uniformément vers  $F$ . On remarque qu'il est toujours possible de supposer que  $F_n$  est de la forme

$$F_n(t) = \sum_{k=0}^n c_k \cos(kt).$$

En effet, si ce n'est pas le cas, on remarque que puisque  $F(t) = F(-t)$ , la suite des fonctions  $t \mapsto F_n(-t)$  converge aussi uniformément vers  $F$ , ainsi que la suite  $t \mapsto \frac{1}{2}(F_n(t) + F_n(-t))$  qui a la forme souhaitée. On remarque que les fonctions  $\cos(kt)$  peuvent s'exprimer comme des polynômes de degré  $k$  en la variable

$\cos(t)$  : il existe une famille de polynômes à coefficients réels  $T_k \in \mathcal{P}_k$ , appelés *polynômes de Tchebychev*, telle que pour tout  $k \geq 0$  et  $t \in \mathbb{R}$ ,

$$\cos(kt) = T_k(\cos(t)).$$

Ceci est évident pour les valeurs  $k = 0, 1, 2$  pour lesquelles on a  $T_0(x) = 1$ ,  $T_1(x) = x$  et  $T_2(x) = 2x^2 - 1$ . On peut ensuite le montrer par récurrence en remarquant que

$$\cos((n+1)t) + \cos((n-1)t) = 2 \cos(t) \cos(nt),$$

ce qui conduit à la relation

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x),$$

qui montre que  $T_n \in \mathcal{P}_n$ . On peut par conséquent écrire  $F_n$  sous la forme

$$F_n(t) = \sum_{k=0}^n b_k \cos(t)^k,$$

et on définit alors  $f_n \in \mathcal{P}_n$  par

$$f_n(x) = \sum_{k=0}^n b_k x^k.$$

On conclut en écrivant

$$\|f - f_n\| = \max_{x \in [-1,1]} |f(x) - f_n(x)| = \max_{t \in \mathbb{R}} |f(\cos(t)) - f_n(\cos(t))| = \|F - F_n\|,$$

ce qui montre que  $f_n$  converge uniformément vers  $f$ . □

En examinant la preuve du théorème de Weierstrass, on constate que si  $f$  est  $M$ -Lipschitzienne sur  $[-1, 1]$ , on a

$$|F(t) - F(u)| = |f(\cos(t)) - f(\cos(u))| \leq M |\cos(t) - \cos(u)| \leq M |t - u|,$$

où on a utilisé le fait que  $t \mapsto \cos t$  est 1-Lipschitzienne puisque sa dérivée est inférieure à 1 en valeur absolue. Par conséquent, en utilisant le Corollaire 3.1.2, on obtient

$$\|f - f_n\| = \|F - F_n\| \leq C_1 M n^{-1}.$$

On a ainsi prouvé le résultat suivant.

**Proposition 3.2.1** *Si  $f$  est  $M$ -Lipschitzienne sur  $[-1, 1]$ , alors*

$$\inf_{g \in \mathcal{P}_n} \|f - g\| \leq C_1 M n^{-1}.$$

*En particulier si  $f$  est de classe  $C^1$  on a*

$$\inf_{g \in \mathcal{P}_n} \|f - g\| \leq C_1 \|f'\| n^{-1}.$$

Par changement de variable affine, on obtient les mêmes estimations sur l'intervalle  $[a, b]$ , avec la constante  $C_1$  multipliée par  $b - a$ . Ce résultat se généralise sous la forme suivante que l'on admet.

**Théorème 3.2.2** *Si  $f$  est de classe  $C^{m-1}$  sur  $[a, b]$  et telle que  $f^{(m-1)}$  est  $M$ -Lipschitzienne, alors*

$$\inf_{g \in \mathcal{P}_n} \|f - g\| \leq C_m M n^{-m},$$

*où la constante  $C_m$  dépend de  $m$  et de  $b - a$  et est indépendante de  $n$  et  $f$ . En particulier si  $f$  est de classe  $C^m$  on a*

$$\inf_{g \in \mathcal{P}_n} \|f - g\| \leq C_m \|f^{(m)}\| n^{-m}.$$

Nous terminons cette section en indiquant une autre preuve du théorème de Weierstrass utilisant la famille des *polynômes de Bernstein*. On se ramène dans ce cas par changement de variable affine à une fonction  $f$  définie sur  $I = [0, 1]$ . Pour une telle fonction, on définit le  $n$ -ème polynôme de Bernstein par

$$B_n f(x) := \sum_{k=0}^n f\left(\frac{k}{n}\right) \binom{n}{k} x^k (1-x)^{n-k},$$

où  $\binom{n}{k} := \frac{n!}{k!(n-k)!}$ .

**Théorème 3.2.3** *Si  $f$  est continue sur  $[0, 1]$ , on a  $\lim_{n \rightarrow +\infty} \|f - B_n f\| = 0$ .*

**Preuve :** On note  $e_0(x) = 1$ ,  $e_1(x) = x$  et  $e_2(x) = x^2$  et on examine les polynômes de Bernstein associés à ces trois fonctions. Pour  $e_0$ , on a

$$B_n e_0(x) = \sum_{k=0}^n \binom{n}{k} x^k (1-x)^{n-k} = (x + 1 - x)^n = 1.$$

Pour  $e_1$ , on obtient

$$\begin{aligned} B_n e_1(x) &= \sum_{k=0}^n \frac{k}{n} \binom{n}{k} x^k (1-x)^{n-k} \\ &= \sum_{k=1}^n \binom{n-1}{k-1} x^k (1-x)^{n-k} \\ &= x \left( \sum_{k=1}^n \binom{n-1}{k-1} x^{k-1} (1-x)^{n-1-(k-1)} \right) \\ &= x B_{n-1} e_0(x) = x. \end{aligned}$$

Pour  $e_2$ , on obtient par des considérations similaires

$$\begin{aligned} B_n e_2(x) &= \sum_{k=0}^n \left(\frac{k}{n}\right)^2 \binom{n}{k} x^k (1-x)^{n-k} \\ &= \frac{n-1}{n} \left( \sum_{k=0}^n \frac{k(k-1)}{n(n-1)} \binom{n}{k} x^k (1-x)^{n-k} + \sum_{k=0}^n \frac{k}{n(n-1)} \binom{n}{k} x^k (1-x)^{n-k} \right) \\ &= \frac{n-1}{n} \left( \sum_{k=2}^n \binom{n-2}{k-2} x^k (1-x)^{n-k} + \frac{1}{n-1} B_n e_1(x) \right) \\ &= \frac{n-1}{n} x^2 + \frac{1}{n} x. \end{aligned}$$

Pour  $x \in [0, 1]$ , on peut écrire

$$\begin{aligned} |f(x) - B_n f(x)| &= \left| f(x) - \sum_{k=0}^n f\left(\frac{k}{n}\right) \binom{n}{k} x^k (1-x)^{n-k} \right| \\ &= \left| \sum_{k=0}^n \left( f(x) - f\left(\frac{k}{n}\right) \right) \binom{n}{k} x^k (1-x)^{n-k} \right| \end{aligned}$$

Pour  $\delta > 0$  fixé, on peut estimer la somme ci-dessus en distinguant l'ensemble  $E$  des  $k \in \{0, \dots, n\}$  tels que  $|\frac{k}{n} - x| \leq \delta$  et son complémentaire  $F$ . En notant

$$\Sigma_E := \left| \sum_{k \in E} \left( f(x) - f\left(\frac{k}{n}\right) \right) \binom{n}{k} x^k (1-x)^{n-k} \right|,$$

et  $\Sigma_F$  la somme similaire pour  $k \in F$ , on a donc  $|f(x) - B_n f(x)| \leq \Sigma_E + \Sigma_F$ . On estime le premier terme en écrivant

$$\begin{aligned} \Sigma_E &\leq (\max_{k \in E} |f(x) - f(\frac{k}{n})|) \sum_{k \in E} \binom{n}{k} x^k (1-x)^{n-k} \\ &\leq \max_{|x-y| \leq \delta} |f(x) - f(y)| = \omega(f, \delta). \end{aligned}$$

Pour le second terme, on peut écrire

$$\begin{aligned} \Sigma_F &\leq 2\|f\| \sum_{k \in F} \binom{n}{k} x^k (1-x)^{n-k} \\ &\leq \frac{2\|f\|}{\delta^2} \sum_{k \in F} \left( x - \frac{k}{n} \right)^2 \binom{n}{k} x^k (1-x)^{n-k} \\ &= \frac{2\|f\|}{\delta^2} (x^2 B_n e_0(x) - 2x B_n e_1(x) + B_n e_2(x)) \\ &= \frac{2\|f\|}{\delta^2} \frac{1}{n} (x - x^2) \leq \frac{2\|f\|}{n\delta^2}. \end{aligned}$$

Comme ceci est valable pour tout  $x \in [0, 1]$ , on a ainsi obtenu l'estimation

$$\|f - B_n f\| \leq \omega(f, \delta) + \frac{2\|f\|}{n\delta^2}.$$

Pour tout  $\varepsilon > 0$ , on peut choisir  $\delta > 0$  tel que  $\omega(f, \delta) \leq \varepsilon/2$ , puis  $n_0$  tel que  $\frac{2\|f\|}{n\delta^2} \leq \varepsilon/2$  pour  $n \geq n_0$ , ce qui entraîne  $\|f - B_n f\| \leq \varepsilon$ . On a ainsi montré que  $B_n f$  converge uniformément vers  $f$ .  $\square$

### 3.3 Interpolation polynomiale

Nous étudions à présent un procédé permettant d'obtenir de façon simple une approximation d'une fonction  $f$  par une fonction polynomiale. On se place ici sur un intervalle  $I$  et on se donne  $n + 1$  points distincts sur cet intervalle :

$$x_0 < x_1 < \cdots < x_{n-1} < x_n.$$

Etant donné un ensemble de  $n + 1$  réels  $\{y_0, \dots, y_n\}$ , on se pose la question de l'existence et de l'unicité d'un polynôme de degré  $n$  dont le graphe passe par tous les points  $(x_i, y_i)$ . Ceci est évident dans le cas  $n = 1$  : il existe une unique droite passant par deux points. Le résultat suivant montre que ceci est aussi vrai pour  $n > 1$ .

**Théorème 3.3.1** *Pour tout ensemble de réels  $\{y_0, \dots, y_n\}$ , il existe un unique polynôme  $p_n \in \mathcal{P}_n$  tel que  $p_n(x_i) = y_i$  pour tout  $i = 0, \dots, n$ .*

**Preuve :** Ceci revient à montrer que l'application  $L : \mathcal{P}_n \rightarrow \mathbb{R}^{n+1}$  qui à  $p \in \mathcal{P}_n$  associe le vecteur de coordonnées  $(p(x_0), \dots, p(x_n))$  est bijective. Cette application est linéaire, et  $\dim(\mathcal{P}_n) = \dim(\mathbb{R}^{n+1})$ . Il suffit donc de démontrer qu'elle est injective, c'est à dire que son noyau est réduit au polynôme nul. Or  $L(p) = 0$  signifie que  $p$  s'annule aux  $n + 1$  points distincts  $x_0, \dots, x_n$  ce qui n'est possible que si  $p = 0$  puisque c'est un polynôme de degré  $n$ .  $\square$

**Remarque 3.3.1** *Une autre façon de prouver l'existence et l'unicité de  $p_n$  est de l'exprimer sous la forme  $p_n(x) = \sum_{j=0}^n a_j x^j$ . Les équations  $p_n(x_i) = y_i$  pour  $i = 0, \dots, n$ , sont alors équivalentes au système  $(n + 1) \times (n + 1)$*

$$Va = y$$

où  $a$  et  $y$  sont les vecteurs de coordonnées  $(a_0, \dots, a_n)$  et  $(y_0, \dots, y_n)$  et où  $V = (x_i^j)_{i,j=0,\dots,n}$  est la matrice de Vandermonde associée aux points  $x_0, \dots, x_n$ . Comme ces points sont distincts, on sait que  $V$  est inversible et il existe donc une unique solution.

Le polynôme  $p_n$  est appelé polynôme d'interpolation de Lagrange (ou interpolant de Lagrange) des valeurs  $y_1, \dots, y_n$  aux points  $x_1, \dots, x_n$ . On peut préciser sa forme en introduisant les polynômes  $(\ell_i)_{i=0,\dots,n}$ , définis par

$$\ell_i(x) = \frac{\prod_{j \in \{1, \dots, n\} - \{i\}} (x - x_j)}{\prod_{j \in \{1, \dots, n\} - \{i\}} (x_i - x_j)} = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}.$$

On note que  $\ell_i$  est l'unique polynôme de degré  $n$  tel que  $\ell_i(x_j) = 0$  si  $i \neq j$  et  $\ell_i(x_i) = 1$ . On peut alors écrire

$$p_n(x) = \sum_{i=0}^n y_i \ell_i(x).$$

Puisque  $p_n = 0$  si et seulement si tous les  $y_i$  sont nuls, la famille  $\{\ell_0, \dots, \ell_n\}$  constitue une base de  $\mathcal{P}_n$ . Les fonctions  $\ell_i$  sont parfois appelées fonctions de base de Lagrange de degré  $n$  pour les points  $\{x_0, \dots, x_n\}$ . Il pourra être utile pour l'intuition de tracer les graphes de ces fonctions dans les cas simples  $n = 1$  et  $n = 2$ .

**Définition 3.3.1** *Si  $f$  est une fonction définie sur  $I$ , on appelle polynôme d'interpolation de Lagrange de  $f$  aux points  $x_0, \dots, x_n$ , l'unique  $p_n \in \mathcal{P}_n$  tel que  $p_n(x_i) = f(x_i)$ , pour  $i = 0, \dots, n$ .*

D'après les remarques précédentes, ce polynôme peut s'écrire sous la forme

$$p_n(x) := \sum_{i=0}^n f(x_i) \ell_i(x),$$

dite *forme de Lagrange*. On remarque aussi que  $f \in \mathcal{P}_n$  si et seulement si  $p_n = f$ . Il est possible d'exprimer  $p_n$  sous une forme différente dite *forme de Newton*. On définit par récurrence les *différences divisées* de  $f$  en posant pour tout les points  $x_0 < \dots < x_n$ ,

$$f[x_i] := f(x_i),$$

puis en définissant les différences d'ordre 1 par

$$f[x_i, x_{i+1}] = \frac{f[x_{i+1}] - f[x_i]}{x_{i+1} - x_i},$$

les différences d'ordre 2 par

$$f[x_i, x_{i+1}, x_{i+2}] = \frac{f[x_{i+1}, x_{i+2}] - f[x_i, x_{i+1}]}{x_{i+2} - x_i},$$

jusqu'à la différence d'ordre  $n$  par

$$f[x_0, \dots, x_n] = \frac{f[x_1, \dots, x_n] - f[x_0, \dots, x_{n-1}]}{x_n - x_0}.$$

Le calcul des différences divisées se fait ainsi de proche en proche par ordre croissant.

**Proposition 3.3.1** *Le polynôme d'interpolation de Lagrange de  $f$  aux points  $x_0, \dots, x_n$  peut aussi s'écrire*

$$p_n(x) = \sum_{i=0}^n f[x_0, \dots, x_i] \prod_{j=0}^{i-1} (x - x_j).$$

**Preuve :** On procède par récurrence sur  $n$ . Il est évident que  $f[x_0] = f(x_0)$  est le polynôme (constant) d'interpolation de Lagrange de  $f$  au point  $x_0$ , et de même que  $f[x_0] + f[x_0, x_1](x - x_0)$  est le polynôme affine d'interpolation de Lagrange aux points  $x_0$  et  $x_1$ . On suppose la proposition vérifiée à l'ordre  $n - 1$ . On remarque que  $p_n - p_{n-1}$  est un polynôme de degré  $n$  qui s'annule aux points  $x_0, \dots, x_{n-1}$  et par conséquent il existe  $\tau \in \mathbb{R}$  tel que

$$p_n(x) = p_{n-1}(x) + \tau \prod_{j=0}^{n-1} (x - x_j) = \sum_{i=0}^{n-1} f[x_0, \dots, x_i] \prod_{j=0}^{i-1} (x - x_j) + \tau \prod_{j=0}^{n-1} (x - x_j).$$

Il reste à montrer que  $\tau = f[x_0, \dots, x_n]$ . Pour cela on remarque que  $\tau$  est le coefficient de  $x^n$  dans le polynôme d'interpolation de Lagrange  $p_n$ , et on montre par récurrence que ce coefficient est égal à  $f[x_0, \dots, x_n]$ . C'est évident pour  $n = 0$  et  $n = 1$ . En supposant cela vrai à l'ordre  $n - 1$ , on pose

$$q_n(x) = \frac{(x - x_0)r_{n-1}(x) - (x - x_n)p_{n-1}(x)}{x_n - x_0},$$

où  $r_{n-1}$  est le polynôme d'interpolation de  $f$  pour les points  $x_1, \dots, x_n$ . On voit ainsi que  $q_n \in \mathcal{P}_n$  et on vérifie aisément que

$$q_n(x_i) = f(x_i), \quad i = 0, \dots, n.$$

Par conséquent  $q_n = p_n$ . D'après l'hypothèse de récurrence, le coefficient de  $x^{n-1}$  dans  $r_{n-1}$  et  $p_{n-1}$  est respectivement  $f[x_1, \dots, x_n]$  et  $f[x_0, \dots, x_{n-1}]$ . Par conséquent, le coefficient de  $x^n$  dans  $p_n$  est donné par

$$\frac{f[x_1, \dots, x_n] - f[x_0, \dots, x_{n-1}]}{x_n - x_0} = f[x_0, \dots, x_n],$$

ce qui conclut la preuve. □

Il existe un autre procédé d'interpolation dû à Hermite et qui fait intervenir les valeurs de  $f$  ainsi que de ses dérivées. On se donne ici deux points  $a < b$  et on part du résultat suivant.

**Théorème 3.3.2** *Pour tout ensemble de réels  $\{\alpha_0, \dots, \alpha_n, \beta_0, \dots, \beta_n\}$ , il existe un unique polynôme  $q \in \mathcal{P}_{2n+1}$  tel que*

$$q^{(i)}(a) = \alpha_i \quad \text{et} \quad q^{(i)}(b) = \beta_i, \quad i = 0, \dots, n.$$

où l'on a utilisé la convention  $q^{(0)} = q$ .

**Preuve :** Comme dans la preuve du Théorème 3.3.1, il suffit de montrer que l'application linéaire  $L : \mathcal{P}_{2n+1} \rightarrow \mathbb{R}^{2n+2}$  qui à  $p \in \mathcal{P}_{2n+1}$  associe le vecteur de coordonnées  $(p(a), \dots, p^{(n)}(a), p(b), \dots, p^{(n)}(b))$  est injective. Or si ce vecteur s'annule, cela signifie que  $p$  est de la forme

$$p(x) = (x - a)^n (x - b)^n r(x),$$

où  $r$  est un polynôme, ce qui n'est possible que si  $r = 0$  puisque  $p$  est au plus de degré  $2n + 1$ . Par conséquent  $p = 0$ .  $\square$

**Définition 3.3.2** Soit  $f$  une fonction de classe  $\mathcal{C}^n$  sur un intervalle ouvert  $I$  et soit  $a < b$  deux points de cet intervalle. On définit le polynôme d'interpolation de Hermite d'ordre  $n$  de  $f$  aux points  $a$  et  $b$  comme l'unique  $q_{2n+1} \in \mathcal{P}_{2n+1}$  tel que

$$q_{2n+1}^{(i)}(a) = f^{(i)}(a) \text{ et } q_{2n+1}^{(i)}(b) = f^{(i)}(b), \quad i = 0, \dots, n.$$

Il est possible d'exprimer l'interpolation de Hermite à l'aide de fonctions de base sous la forme

$$q_{2n+1}(x) = \sum_{i=0}^n (f^{(i)}(a) \ell_{a,i}(x) + f^{(i)}(b) \ell_{b,i}(x)).$$

A titre d'exercice, on pourra chercher l'expression des fonctions  $\ell_{a,i}$  et  $\ell_{b,i}$  lorsque  $n = 1$  ce qui correspond à une interpolation par des polynômes de degré 3.

### 3.4 Estimation de l'erreur d'interpolation

Afin d'étudier l'erreur entre  $f$  et son polynôme d'interpolation de Lagrange  $p_n$  aux points  $x_0, \dots, x_n$ , on donne tout d'abord un résultat qui généralise le théorème de Rolle.

**Lemme 3.4.1** Soit  $f$  une fonction de classe  $\mathcal{C}^n$  sur un intervalle  $I$  et qui s'annule en  $n$  points distincts  $a_0 < \dots < a_n$  contenus dans cet intervalle. Alors il existe un point  $z \in ]a_0, a_n[$  tel que  $f^{(n)}(z) = 0$ .

**Preuve :** On procède par récurrence. Pour  $n = 1$  c'est le théorème de Rolle. On suppose la propriété vraie à l'ordre  $n - 1$ . Si  $f$  s'annule aux points  $a_0, \dots, a_n$ , alors  $f'$  s'annule en  $n$  points  $b_0, \dots, b_{n-1}$  avec  $a_i < b_i < a_{i+1}$ . Par l'hypothèse de récurrence,  $f^{(n)} = (f')^{(n-1)}$  s'annule en un point  $z \in ]b_0, b_{n-1}[ \subset ]a_0, a_n[$ .  $\square$

Afin de décrire l'erreur d'interpolation on introduit la fonction

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

**Théorème 3.4.1** Soit  $f$  une fonction de classe  $\mathcal{C}^{n+1}$  sur  $I$  et  $p_n$  son polynôme d'interpolation de Lagrange en  $n$  points distincts  $x_0 < \dots < x_n$  contenus dans  $I$ , alors pour tout  $x \in I$  il existe  $y \in I$  qui dépend de  $x$ , tel que

$$f(x) - p_n(x) = f^{(n+1)}(y) \Pi_n(x).$$

Le point  $y$  est contenu dans  $] \min\{x, x_0\}, \max\{x, x_n\}[$ , c'est à dire dans  $]x, x_n[$  si  $x < x_0$ , dans  $]x_0, x[$  si  $x > x_n$ , dans  $]x_0, x_n[$  si  $x \in [x_0, x_n]$ .

**Preuve :** Dans le cas où  $x$  est égal à l'un des  $x_i$ , il n'y a rien à prouver puisque les deux membres de l'égalité sont nuls. On suppose que  $x$  est différent de tous les  $x_i$ , ce qui entraîne  $\Pi_n(x) \neq 0$ . Par conséquent, il existe un nombre  $\mu \in \mathbb{R}$  (qui dépend de  $x$ ) tel que

$$f(x) - p_n(x) = \mu \Pi_n(x).$$

La fonction  $g(t) = f(t) - p_n(t) - \mu \Pi_n(t)$  s'annule aux  $n+2$  points distincts  $x_0, \dots, x_n$  et  $x$ . Par conséquent, d'après le Lemme 3.4.1 il existe un point  $y \in ] \min\{x, x_0\}, \max\{x, x_n\}[$  tel que

$$0 = g^{(n+1)}(y) = f^{(n+1)}(y) - p_n^{(n+1)}(y) - \mu \Pi_n^{(n+1)}(y) = f^{(n+1)}(y) - \mu,$$

et par conséquent  $\mu = f^{(n+1)}(y)$ . □

A partir de ce résultat on peut estimer l'erreur d'interpolation sur un intervalle  $[a, b]$  qui contient  $[x_0, x_n]$ . Une première conséquence du Théorème 3.4.1 est l'estimation en norme sup sur  $[a, b]$ .

$$\|f - p_n\| \leq \|f^{(n+1)}\| \|\Pi_n\|$$

On peut estimer la norme sup de  $\Pi_n$  en écrivant

$$\begin{aligned} \|\Pi_n\| &= \frac{1}{(n+1)!} \max_{x \in [a, b]} \prod_{i=0}^n |x - x_i| \\ &\leq \frac{1}{(n+1)!} (b-a)^{n+1}. \end{aligned}$$

Ceci entraîne l'estimation

$$\|f - p_n\| \leq \frac{1}{(n+1)!} \|f^{(n+1)}\| (b-a)^{n+1}.$$

Cette estimation peut être améliorée pour des choix particuliers des points d'interpolation. Par exemple, dans le cas de points  $a = x_0 < \dots < x_n$  *équidistants* c'est à dire

$$x_i = a + \frac{i}{n}(b-a),$$

il est facile d'établir que  $\prod_{i=0}^n |x - x_i| \leq \frac{n!(b-a)^{n+1}}{n^{n+1}}$ , ce qui conduit à l'estimation

$$\|f - p_n\| \leq \frac{1}{n^{n+2}} \|f^{(n+1)}\| (b-a)^{n+1},$$

asymptotiquement meilleure que la précédente quand  $n \rightarrow +\infty$ .

Ces estimations nous permettent d'établir un premier résultat sur la convergence de l'interpolant de Lagrange  $p_n$  vers  $f$  lorsque  $n \rightarrow +\infty$ , dans le cas où  $f$  est très régulière au sens où elle admet un développement en série entière convergent sur l'intervalle  $[a, b]$ .

**Théorème 3.4.2** *Soit  $f$  une fonction qui admet un développement en série entière au point  $\frac{a+b}{2}$  de rayon de convergence  $R > \frac{3}{2}(b-a)$ . Alors la suite  $p_n$  converge uniformément vers  $f$  sur  $[a, b]$ .*

**Preuve :** D'après l'hypothèse sur  $f$ , pour tout  $0 < r < R$ , la série

$$\sum_{k \geq 0} \frac{1}{k!} |f^{(k)}(\frac{a+b}{2})| r^k,$$

est convergente. En notant  $C(r)$  sa somme, on a en particulier,

$$|f^{(k)}(\frac{a+b}{2})| \leq C(r) k! r^{-k}.$$

Comme on a supposé  $R > \frac{3}{2}(b-a)$ , on peut choisir  $r$  dans l'intervalle  $[\frac{b-a}{2}, R[$ . Pour tout  $x \in [a, b]$ , on peut dériver terme à terme la série entière

$$f(x) = \sum_{k \geq 0} \frac{1}{k!} f^{(k)}(\frac{a+b}{2}) (x - \frac{a+b}{2})^k.$$

En posant  $u := x - \frac{a+b}{2}$ , on obtient après  $n$  dérivations

$$f^{(n)}(x) = \sum_{k \geq 0} \frac{1}{k!} f^{(k)}(\frac{a+b}{2}) \frac{d^n}{du^n} (u^k)$$

Lorsque  $0 \leq u \leq \frac{a+b}{2}$ , on a  $\frac{d^n}{du^n}(u^k) \geq 0$  et on peut donc écrire

$$\begin{aligned} |f^{(n)}(x)| &\leq \sum_{k \geq 0} \frac{1}{k!} |f^{(k)}(\frac{a+b}{2})| \frac{d^n}{du^n}(u^k) \\ &\leq C(r) \sum_{k \geq 0} r^{-k} \frac{d^n}{du^n}(u^k) \\ &= C(r) \frac{d^n}{du^n} \left( \sum_{k \geq 0} (\frac{u}{r})^k \right) \\ &= C(r) \frac{d^n}{du^n} \left( \frac{r}{r-u} \right) \\ &= \frac{C(r)n!r}{(r-u)^{n+1}} \\ &\leq C(r)n!r \left| r - \frac{b-a}{2} \right|^{-(n+1)} \end{aligned}$$

Lorsque  $-\frac{a+b}{2} \leq u \leq 0$ , on fait le même calcul en posant  $v = -u$  et en dérivant par rapport à  $v$ , et on aboutit aussi à l'estimation

$$|f^{(n)}(x)| \leq C(r)n!r \left| r - \frac{b-a}{2} \right|^{-(n+1)}.$$

On a par conséquent

$$\|f^{(n+1)}\| \leq C(r)(n+1)!r \left| r - \frac{b-a}{2} \right|^{-(n+1)},$$

ce qui combiné à l'estimation  $\|f - p_n\| \leq \frac{1}{(n+1)!} \|f^{(n+1)}\| (b-a)^{n+1}$ , conduit à

$$\|f - p_n\| \leq rC(r)\rho^{n+1} \quad \text{avec } \rho := \frac{b-a}{r - \frac{b-a}{2}}$$

Lorsque  $R > \frac{3}{2}(b-a)$ , il est possible de choisir  $r < R$  tel que  $0 < \rho < 1$ , ce qui entraîne la convergence uniforme.  $\square$

**Remarque 3.4.1** Dans le cas où les points  $x_i$  sont équidistants avec  $x_0 = a$  et  $x_n = b$ , on peut utiliser l'estimation  $\|f - p_n\| \leq \frac{1}{n^{n+2}} \|f^{(n+1)}\| (b-a)^{n+1}$  afin d'obtenir le résultat du théorème ci-dessus sous la condition plus faible  $R > (\frac{1}{2} + \frac{1}{e})(b-a)$  (indication : utiliser la formule de Stirling qui donne un équivalent de  $n!$  quand  $n \rightarrow +\infty$ ).

La preuve du Théorème 3.4.2 nous indique que la convergence de  $p_n$  vers  $f$  est très rapide, puisque  $\|f - p_n\| \leq C\rho^{n+1}$  avec  $0 < \rho < 1$ , mais ceci est au prix d'hypothèses très fortes sur  $f$  qui est supposée développable en série entière sur un intervalle  $]\frac{a+b}{2} - R, \frac{a+b}{2} + R[$  contenant  $[a, b]$  avec  $R$  suffisamment grand, et en particulier  $\mathcal{C}^\infty$  sur cet intervalle. Lorsque de telles hypothèses ne sont pas satisfaites, la convergence de  $p_n$  vers  $f$  n'est plus garantie, et elle demande un examen approfondi faisant intervenir le choix des points d'interpolation  $x_i$  dans l'intervalle  $[a, b]$ .

Le choix le plus naturel consiste à prendre des points équidistants, mais dans ce cas il est possible de mettre en évidence des problèmes de convergence même pour des fonctions très régulières : il existe des fonctions  $f$  développables en série entières sur un intervalle contenant  $[a, b]$  mais telles que  $p_n$  ne converge pas vers  $f$ . C'est le *phénomène de Runge* que l'on peut illustrer numériquement sur  $[a, b] = [-1, 1]$  en considérant la fonction  $f(x) = (x^2 + \alpha)^{-1}$  avec  $\alpha > 0$  : lorsque  $\alpha$  est suffisamment petit, on constate la divergence de la suite  $p_n$  lorsque  $n \rightarrow +\infty$  qui se traduit en particulier par des oscillations au voisinage des extrémités de l'intervalle. Le choix des points d'interpolation joue aussi un rôle important dans l'étude de la stabilité numérique du procédé d'interpolation qui est l'objet de la section suivante. Cette étude conduit à proposer d'autres choix que celui des points équidistants, et qui donnent de meilleurs résultats de convergence.

### 3.5 Stabilité

Etant donné un choix de points  $a \leq x_0 < \dots < x_n \leq b$ , désignons par  $\mathcal{I}_n$  l'opérateur qui à une fonction  $f$  continue sur  $[a, b]$  associe son polynôme d'interpolation  $p_n$  de degré  $n$  aux points  $x_0, \dots, x_n$ . On dit que  $\mathcal{I}_n$  est l'opérateur d'interpolation aux points  $x_0, \dots, x_n$ . On a donc

$$\mathcal{I}_n : \mathcal{C}([a, b]) \mapsto \mathcal{P}_n, \quad \mathcal{I}_n f(x) = p_n(x) = \sum_{i=0}^n f(x_i) \ell_i(x),$$

où les  $\ell_i$  sont les fonctions de bases de Lagrange aux points  $x_0, \dots, x_n$ . Il est immédiat de vérifier que  $\mathcal{P}_n$  est une application linéaire c'est à dire un élément de  $\mathcal{L}(\mathcal{C}([a, b]), \mathcal{P}_n)$ . On appelle *constante de Lebesgue* du procédé d'interpolation de Lagrange aux points  $x_0, \dots, x_n$  la norme de l'opérateur  $\mathcal{I}_n$  subordonnée à la norme sup sur  $[a, b]$ , c'est à dire

$$\Lambda_n = \sup_{f \in \mathcal{C}([a, b]), \|f\| \leq 1} \|\mathcal{I}_n f\| = \sup_{f \in \mathcal{C}([a, b]), f \neq 0} \frac{\|\mathcal{I}_n f\|}{\|f\|},$$

où  $\|f\| := \sup_{x \in [a, b]} |f(x)|$ . La constante de Lebesgue joue un rôle centrale dans l'étude de la stabilité du procédé d'interpolation puisque pour toute paire de fonctions  $f$  et  $g$  on a

$$\|\mathcal{I}_n f - \mathcal{I}_n g\| \leq \Lambda_n \|f - g\|$$

Ceci signifie que si on fait une erreur de norme  $\varepsilon > 0$  sur la fonction  $f$ , il en résulte une erreur de norme au plus  $\Lambda_n \varepsilon$  sur son polynôme d'interpolation.

**Remarque 3.5.1** *Il est facile de vérifier que constante de Lebesgue peut-être aussi être définie comme la norme de l'application linéaire qui à un vecteur  $y = (y_0, \dots, y_n)$  associe le polynôme d'interpolation aux points  $(x_i, y_i)$ , c'est à dire*

$$\mathcal{J}_n : \mathbb{R}^{n+1} \mapsto \mathcal{P}_n, \quad \mathcal{J}_n y = \sum_{i=0}^n y_i \ell_i(x).$$

Plus précisément, on a

$$\Lambda_n = \sup_{y \in \mathbb{R}^{n+1}, \|y\|_\infty \leq 1} \|\mathcal{J}_n y\| = \sup_{y \in \mathbb{R}^{n+1}, y \neq 0} \frac{\|\mathcal{J}_n y\|}{\|y\|_\infty}.$$

Ceci signifie que si on commet une erreur de  $\varepsilon$  sur les données  $y_i$ , il en résulte une erreur de norme au plus  $\Lambda_n \varepsilon$  sur le polynôme d'interpolation.

La constante de Lebesgue joue aussi un rôle important dans l'étude de l'erreur entre  $f$  et son polynôme d'interpolation, comme le montre le résultat suivant.

**Théorème 3.5.1** *Pour tout  $f \in \mathcal{C}([a, b])$ , on a*

$$\|f - \mathcal{I}_n f\| \leq (1 + \Lambda_n) \inf_{g \in \mathcal{P}_n} \|f - g\|$$

**Preuve :** Pour tout  $g \in \mathcal{P}_n$ , on peut écrire

$$\|f - \mathcal{I}_n f\| \leq \|f - g\| + \|\mathcal{I}_n f - g\| = \|f - g\| + \|\mathcal{I}_n f - \mathcal{I}_n g\| \leq (1 + \Lambda_n) \|f - g\|,$$

où on a utilisé le fait que  $\mathcal{I}_n g = g$ . Comme  $g$  est arbitraire on obtient le résultat annoncé.  $\square$

En combinant ce résultat avec ceux qui décrivent l'erreur de meilleure approximation par des polynômes, on obtient des estimations sur la l'erreur d'interpolation. Par exemple, en utilisant le Théorème 3.2.2, on obtient le résultat suivant.

**Corollaire 3.5.1** *Si  $f$  est de classe  $\mathcal{C}^{m-1}$  sur  $[a, b]$  et telle que  $f^{(m-1)}$  est  $M$ -Lipschitzienne, alors*

$$\|f - \mathcal{I}_n f\| \leq C_m M (1 + \Lambda_n) n^{-m},$$

où la constante  $C_m$  dépend de  $m$  et de  $b - a$  et est indépendante de  $n$  et  $f$ . En particulier si  $f$  est de classe  $\mathcal{C}^m$  on a

$$\|f - \mathcal{I}_n f\| \leq C_m \|f^{(m)}\| (1 + \Lambda_n) n^{-m}.$$

Pour préciser ces estimations, il est important de comprendre si la constante de Lebesgue augmente lorsque  $n \rightarrow +\infty$ , et si sa croissance peut compenser le facteur de décroissance  $n^{-m}$ . Donnons d'abord un moyen de calcul de  $\Lambda_n$ .

**Proposition 3.5.1** *On a*

$$\Lambda_n = \max_{x \in [a, b]} \sum_{i=0}^n |\ell_i(x)| = \left\| \sum_{i=0}^n |\ell_i| \right\|$$

où les  $\ell_i$  sont les fonctions de bases de Lagrange.

**Preuve :** Pour tout  $x \in [a, b]$ , on a

$$|\mathcal{I}_n f(x)| = \left| \sum_{i=0}^n f(x_i) \ell_i(x) \right| \leq \left( \max_{i=0, \dots, n} |f(x_i)| \right) \sum_{i=0}^n |\ell_i(x)| \leq \|f\| \left\| \sum_{i=0}^n |\ell_i| \right\|,$$

ce qui entraîne

$$\|\mathcal{I}_n f\| \leq \|f\| \left\| \sum_{i=0}^n |\ell_i| \right\|,$$

et par conséquent  $\Lambda_n \leq \left\| \sum_{i=0}^n |\ell_i| \right\|$ . Pour démontrer l'inégalité inverse, on considère le point  $x^*$  tel que

$$\sum_{i=0}^n |\ell_i(x^*)| = \max_{x \in [a, b]} \sum_{i=0}^n |\ell_i(x)| = \left\| \sum_{i=0}^n |\ell_i| \right\|,$$

et on pose  $y_i = 1$  si  $\ell_i(x^*) > 0$  et  $-1$  sinon. Il est facile de construire une fonction  $f$  telle que  $f(x_i) = y_i$  et  $\|f\| = 1$  (on prend par exemple  $f$  continue et affine sur chaque intervalle  $[x_i, x_{i+1}]$  avec les valeurs prescrites aux points  $x_i$ ). Pour cette fonction, on a

$$\|\mathcal{I}_n f\| \geq |\mathcal{I}_n f(x^*)| = \left| \sum_{i=0}^n y_i \ell_i(x^*) \right| = \sum_{i=0}^n |\ell_i(x^*)| = \left\| \sum_{i=0}^n |\ell_i| \right\|,$$

et par conséquent  $\Lambda_n \geq \left\| \sum_{i=0}^n |\ell_i| \right\|$  □

Considérons à présent le cas particulier où les points  $x_i$  sont équidistants avec  $x_i = a + \frac{i}{n}(b-a)$ . Le résultat suivant nous montre que la constante de Lebesgue croît exponentiellement lorsque  $n$  augmente.

**Proposition 3.5.2** *Pour les points équidistants on a  $\Lambda_n \geq \frac{2^n}{4n^2}$ .*

**Preuve :** Tout  $x \in [a, b]$  peut s'écrire  $x = a + \frac{s}{n}(b-a)$  avec  $s \in [0, n]$  et on a

$$\ell_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j} = \prod_{j \neq i} \frac{s - j}{i - j}.$$

Au point  $x^* = a + \frac{1}{2n}(b-a)$  qui correspond à  $s^* = \frac{1}{2}$  on a

$$\begin{aligned} |\ell_i(x^*)| &= \frac{\prod_{j \neq i} |\frac{1}{2} - j|}{i!(n-i)!} \\ &\geq \frac{\prod_{j \in \{2, \dots, n\} - \{i\}} (j-1)}{4i!(n-i)!} \\ &\geq \frac{n!}{4n^2 i!(n-i)!} = \frac{1}{4n^2} \binom{n}{i}. \end{aligned}$$

Et par conséquent

$$\sum_{i=0}^n |\ell_i(x^*)| \geq \frac{2^n}{4n^2}.$$

ce qui entraîne le résultat puisque  $\Lambda_n \geq \sum_{i=0}^n |\ell_i(x^*)|$ . □

**Remarque 3.5.2** *On peut aussi majorer  $\Lambda_n$  en remarquant que pour tout  $s \in [k, k+1]$  on a*

$$\prod_{j \neq i} \frac{|s - j|}{|i - j|} \leq \frac{(k+1)(n-k)!}{i!(n-i)!} \leq n \binom{n}{i}$$

ce qui conduit à  $\Lambda_n \leq n2^n$ .

On voit ainsi que le choix de points équidistants conduit à des problèmes de stabilité numérique lorsque  $n$  est grand. D'autre part, la croissance exponentielle de  $\Lambda_n$  est plus forte que toute décroissance en  $n^{-m}$ , et par conséquent les estimations telles que celles du Corollaire 3.5.1 ne se traduisent par aucune propriété de convergence. Un meilleur choix est celui des *points de Tchebychev*. Lorsque l'on travaille sur l'intervalle  $[-1, 1]$  ces points sont donnés par

$$u_i = \cos\left(\frac{(2i+1)\pi}{2n+2}\right), \quad i = 0, \dots, n.$$

On remarquera que la répartition de ces points est plus dense au voisinage des extrémités de l'intervalle. On note aussi que ce sont exactement les  $n+1$  racines du polynôme de Tchebychev  $T_{n+1} \in \mathcal{P}_{n+1}$  qui a été introduit dans la preuve du théorème de Weierstrass. Dans le cas d'un intervalle  $[a, b]$  quelconque, on définit les points de Tchebychev en transportant les points définis sur  $[-1, 1]$  par l'application affine  $t \mapsto x = \frac{a+b}{2} + \frac{b-a}{2}u$ . On pose donc

$$x_i = \frac{a+b}{2} + \frac{b-a}{2}u_i = \frac{a+b}{2} + \frac{b-a}{2}\cos\left(\frac{(2i+1)\pi}{2n+2}\right), \quad i = 0, \dots, n.$$

Le résultat suivant nous montre que la constante de Lebesgue a une croissance logarithmique lorsque l'on utilise les points de Tchebychev.

**Proposition 3.5.3** *Pour les points de Tchebychev on a  $\Lambda_n \leq C \log(n)$  pour tout  $n > 0$ , où  $C$  est une constante indépendante de  $n$ .*

**Preuve :** Tout  $x \in [a, b]$  peut s'écrire  $x = \frac{a+b}{2} + \frac{b-a}{2}u$  avec  $u \in [-1, 1]$ . On a

$$\ell_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j} = \prod_{j \neq i} \frac{u - u_j}{u_i - u_j}.$$

En remarquant que  $T_{n+1}(u) = c_n \prod_{j=0}^n (u - u_j)$  avec  $c_n \in \mathbb{R}$  on en déduit

$$\ell_i(x) = \frac{T_{n+1}(u)}{(u - u_i)T'_{n+1}(u_i)}.$$

En dérivant la relation  $T_{n+1}(\cos(t)) = \cos((n+1)t)$ , on voit que

$$\sin(t)T'_{n+1}(\cos(t)) = (n+1)\sin((n+1)t).$$

Tout  $u \in [-1, 1]$  peut s'écrire  $u = \cos(t)$  avec  $t \in [0, \pi]$ , et en particulier  $u_i = \cos(t_i)$  avec  $t_i = \frac{(2i+1)\pi}{2n+2}$ . On a donc

$$\ell_i(x) = \frac{\cos((n+1)t)\sin(t_i)}{(n+1)(\cos(t) - \cos(t_i))\sin((n+1)t_i)} = (-1)^i \frac{\cos((n+1)t)\sin(t_i)}{(n+1)(\cos(t) - \cos(t_i))},$$

où on a utilisé le fait que  $\sin((n+1)t_i) = \sin((i + \frac{1}{2})\pi) = (-1)^i$ . On remarque à présent que

$$\cos(t) - \cos(t_i) = -2\sin\left(\frac{t+t_i}{2}\right)\sin\left(\frac{t-t_i}{2}\right).$$

On remarque que puisque  $\frac{|t-t_i|}{2} \leq \frac{\pi}{2}$  on a

$$\left|\sin\left(\frac{t-t_i}{2}\right)\right| \geq \frac{2}{\pi} \frac{|t-t_i|}{2}.$$

D'autre part, on peut écrire

$$\begin{aligned} \left|\sin\left(\frac{t+t_i}{2}\right)\right| &\geq \min\left\{\sin\left(\frac{t_i}{2}\right), \sin\left(\frac{t_i+\pi}{2}\right)\right\} \\ &= \min\left\{\sin\left(\frac{t_i}{2}\right), \cos\left(\frac{t_i}{2}\right)\right\} \\ &\geq \sin\left(\frac{t_i}{2}\right)\cos\left(\frac{t_i}{2}\right) \\ &= \frac{1}{2}\sin(t_i). \end{aligned}$$

En combinant ces remarques avec l'expression obtenue pour  $\ell_i(x)$ , on obtient

$$|\ell_i(x)| \leq \frac{\pi |\cos((n+1)t)|}{(n+1)|t-t_i|}$$

Soit  $t_j$  le point de Tchebychev le plus proche de  $t$ . Pour  $i \neq j-1, j, j+1$ , on a  $|t-t_i| \geq \frac{\pi}{n+1}(|j-i|-1)$  et on peut donc écrire

$$|\ell_i(x)| \leq \frac{\pi}{(n+1)|t-t_i|} \leq \frac{1}{|i-j|-1}.$$

Pour  $i = j-1, j, j+1$ , on a en utilisant le théorème des accroissements finis,

$$|\cos((n+1)t)| = |\cos((n+1)t) - \cos((n+1)t_i)| \leq (n+1)|t-t_i|,$$

et par conséquent

$$|\ell_i(x)| \leq \pi.$$

Ces deux estimations conduisent finalement à

$$\sum_{i=1}^n |\ell_i(x)| \leq 3\pi + \sum_{i \neq j-1, j, j+1} \frac{1}{|i-j|-1} \leq 3\pi + 2 \sum_{k=1}^{n-1} \frac{1}{k} \leq 3\pi + 2 + \log(n-1).$$

On peut trouver une constante  $C$  telle que  $3\pi + 2 + \log(n-1) \leq C \log n$  pour tout  $n > 0$  ce qui conduit au résultat annoncé.  $\square$

Un corollaire immédiat de ce résultat montre que l'erreur pour l'approximation par l'interpolation aux points de Tchebychev décroît presque à la même vitesse que l'erreur de meilleure approximation par les polynômes.

**Corollaire 3.5.2** *Si  $f$  est de classe  $C^{m-1}$  sur  $[a, b]$  et telle que  $f^{(m-1)}$  est  $M$ -Lipschitzienne, et si  $\mathcal{I}_n$  est l'opérateur d'interpolation avec les points de Tchebychev, on a*

$$\|f - \mathcal{I}_n f\| \leq C_m M n^{-m} \log(n),$$

où la constante  $C_m$  dépend de  $m$  et de  $b-a$  et est indépendante de  $n$  et  $f$ . En particulier si  $f$  est de classe  $C^m$  on a

$$\|f - \mathcal{I}_n f\| \leq C_m \|f^{(m)}\| n^{-m} \log(n).$$

### 3.6 Approximation des moindres carrés

Dans le procédé d'interpolation, on a besoin des valeurs de  $f$  en  $n+1$  points pour construire un polynôme de degré  $n$ . Si l'on dispose des valeurs de  $f$  en  $m+1$  points  $x_0 < \dots < x_m$  avec  $m > n$ , on peut chercher à construire un polynôme de degré  $n$  qui approche  $f$  par un autre procédé. Plus précisément, étant donné un vecteur  $y$  de  $m+1$  coordonnées  $y_0, \dots, y_m$ , on définit le polynôme des *moindres carrés* de degré  $n$  aux points  $x_0, \dots, x_m$  comme le polynôme  $q_n \in \mathcal{P}_n$  qui minimise la quantité

$$\sum_{i=0}^m |q(x_i) - y_i|^2,$$

parmi tous les polynômes  $q \in \mathcal{P}_n$ . Dans le cas où les  $y_i$  sont les valeurs d'une fonction  $f$  aux points  $x_i$ , on dit que  $q_n$  est l'approximation des moindres carrés de degré  $n$  de  $f$  aux points  $x_0, \dots, x_m$ , qui minimise donc la quantité

$$\sum_{i=0}^m |q(x_i) - f(x_i)|^2.$$

Si l'on écrit  $q_n(x) = \sum_{k=0}^n a_k x^k$ , alors on voit que la recherche de  $q_n$  est équivalente à celle du vecteur  $a$  de coordonnées  $a_0, \dots, a_n$  qui minimise la norme euclidienne

$$\|Va - y\|$$

où  $V$  est une matrice  $(m+1) \times (n+1)$  dont les coefficients sont donnés par  $v_{i,j} = x_i^j$ . Il s'agit donc de la méthode des moindres carrés déjà abordée dans la section §1.5. On sait qu'il existe toujours une solution et que celle-ci est unique lorsque  $V$  est injective. C'est le cas ici puisque  $Va = 0$  équivaut à l'annulation du polynôme  $\sum_{k=0}^n a_k x^k$  aux points  $x_0, \dots, x_m$ , ce qui n'est possible que si ce polynôme est nul, i.e.  $a = 0$ , puisque  $m \geq n$ . Les équations normales qui caractérisent  $a$  sont données par le système  $(n+1) \times (n+1)$

$$V^t V a = V^t y,$$

avec

$$V^t V = \left( \sum_{k=0}^m x_k^{i+j} \right)_{i,j=0,\dots,n} \text{ et } V^t y = \left( \sum_{k=0}^m x_k^j y_k \right)_{j=0,\dots,n}.$$

Dans le cas  $n = 0$ , on trouve ainsi que la solution constante du problème des moindres carrés  $q_0(x) = a_0$  est donnée par la moyenne des valeurs  $y_k$  :

$$a_0 = \frac{1}{m+1} \sum_{k=0}^m y_k.$$

Dans le cas  $n = 1$ , la solution affine  $q_1(x) = a_0 + a_1 x$  est appelée en statistiques *droite de régression* pour les points  $\{(x_i, y_i), i = 1, \dots, n\}$ , et ses coefficients se calculent simplement à partir des valeurs  $x_k$  et  $y_k$  en résolvant un système  $2 \times 2$ .

Un autre type d'approximation des moindres carrés pour une fonction  $f$  continue sur un intervalle  $[a, b]$  est obtenu en cherchant à minimiser la quantité

$$\int_a^b |f(x) - q(x)|^2 dx,$$

parmi tous les polynômes  $q \in \mathcal{P}_n$ . Ce procédé est intuitivement lié au précédent en remarquant que si on choisit des points  $a = x_0 < \dots < x_m = b$  équidistants, la quantité

$$\frac{1}{m} \sum_{i=0}^m |f(x_i) - q(x_i)|^2,$$

qui est minimisée par le polynôme des moindres carrés aux points  $x_0, \dots, x_m$  est alors une somme de Riemann qui approche l'intégrale ci-dessus lorsque le nombre de points  $m$  augmente. Il est facile de voir que l'on définit une norme sur  $\mathcal{C}([a, b])$  en posant

$$\|g\|_2 := \left( \int_a^b |g(x)|^2 dx \right)^{1/2}.$$

Cette norme est appelée norme  $L^2$  sur l'intervalle  $[a, b]$ . On remarque qu'elle est dérivée du produit scalaire

$$\langle f, g \rangle := \int_a^b f(x)g(x)dx,$$

au sens où  $\|g\|_2 := \sqrt{\langle g, g \rangle}$ . On recherche donc le polynôme  $q_n \in \mathcal{P}_n$  solution de

$$\|f - q_n\|_2 = \min_{q \in \mathcal{P}_n} \|f - q\|_2.$$

Afin de prouver l'existence et l'unicité du polynôme  $q_n$ , on introduit la suite des *polynômes de Legendre* qui est définie en appliquant le procédé d'orthogonalisation de Gram-Schmidt aux fonctions  $e_k := x \mapsto x^k$ .

**Définition 3.6.1** La suite des polynômes de Legendre orthonormés sur  $[a, b]$  est définie par récurrence en posant  $L_0 = \frac{e_0}{\|e_0\|_2}$  et

$$L_n = \frac{e_n - \sum_{k=0}^{n-1} \langle e_n, L_k \rangle L_k}{\|e_n - \sum_{k=0}^{n-1} \langle e_n, L_k \rangle L_k\|_2},$$

c'est à dire  $L_0(x) = (b-a)^{-1/2}$  et  $L_n(x) = \frac{x^n - \sum_{k=0}^{n-1} \left( \int_a^b x^n L_k(x) \right) L_k(x)}{\left( \int_a^b x^{2n} dx \right)^{1/2}}$ .

On déduit immédiatement de cette définition que les polynômes de Legendre forment un ensemble orthonormé au sens où

$$\langle L_i, L_j \rangle = 0 \text{ si } i \neq j \text{ et } \langle L_i, L_i \rangle = \|L_i\|_2^2 = 1.$$

La famille  $\{L_0, \dots, L_n\}$  est une base orthonormée de  $\mathcal{P}_n$ , et il est aussi facile de vérifier que  $L_n$  est exactement de degré  $n$ .

**Proposition 3.6.1** *Le polynôme  $q_n$  qui minimise  $\|f - q\|_2$  parmi tous les  $q \in \mathcal{P}_n$  est donné par*

$$q_n := \sum_{k=0}^n \langle f, L_k \rangle L_k.$$

*C'est la projection orthogonale de  $f$  sur  $\mathcal{P}_n$  qui est caractérisée par la propriété  $\langle f - q_n, q \rangle = 0$  pour tout  $q \in \mathcal{P}_n$ .*

**Preuve :** Si  $q_n$  est donné par la formule ci-dessus, on remarque que pour tout  $i = 0, \dots, n$ , on a

$$\langle f - q_n, L_i \rangle = \langle f, L_i \rangle - \sum_{k=0}^n \langle f, L_k \rangle \langle L_k, L_i \rangle = \langle f, L_i \rangle - \langle f, L_i \rangle = 0.$$

Puisque tout  $q \in \mathcal{P}_n$  peut s'écrire comme une combinaison linéaire des  $L_i$  on a bien la propriété

$$\langle f - q_n, q \rangle = 0.$$

Pour tout  $q \in \mathcal{P}_n$  on a d'autre part

$$\|f - q\|_2^2 = \|f - q_n + q_n - q\|_2^2 = \|f - q_n\|_2^2 + \|q_n - q\|_2^2 \geq \|f - q_n\|_2^2,$$

puisque  $\langle f - q_n, q_n - q \rangle = 0$ . □

**Remarque 3.6.1** *Il est possible de donner un sens à la projection orthogonale de  $f$  sur  $\mathcal{P}_n$  lorsque  $f$  n'est pas une fonction continue : il suffit en effet que  $f$  soit intégrable sur  $[a, b]$  pour que les produits scalaires  $\langle f, L_k \rangle$  soient bien définis.*

La norme  $L^2$  sur  $[a, b]$  peut-être majorée par la norme sup sur  $[a, b]$  suivant

$$\|g\|_2 \leq (b - a)^{1/2} \|g\|,$$

pour toute fonction  $g \in \mathcal{C}([a, b])$ . On obtient ainsi pour l'erreur de projection orthogonale

$$\|f - q_n\|_{L^2} \leq (b - a)^{1/2} \|f - q\|,$$

pour tout  $q \in \mathcal{P}_n$  c'est à dire

$$\|f - q_n\|_{L^2} \leq (b - a)^{1/2} \inf_{q \in \mathcal{P}_n} \|f - q\|.$$

En combinant cette estimation avec les résultats de meilleure approximation polynomiale obtenue dans la section §3.2, on obtient immédiatement des résultats de convergence de  $q_n$  vers  $f$ .

**Proposition 3.6.2** *Pour toute fonction  $f \in \mathcal{C}([a, b])$ , l'erreur de projection orthogonale sur  $\mathcal{P}_n$  vérifie*

$$\lim_{n \rightarrow +\infty} \|f - q_n\|_2 = 0.$$

*Si  $f$  est de classe  $\mathcal{C}^m$  sur  $[a, b]$ , on a*

$$\|f - q_n\|_2 \leq C_m \|f^{(m)}\| n^{-m}.$$

*où la constante  $C_m$  dépend de  $m$  et de  $b - a$  et est indépendante de  $n$  et  $f$ .*

La convergence de  $q_n$  vers  $f$  signifie que l'on peut écrire

$$f = \sum_{k \geq 0} \langle f, L_k \rangle L_k,$$

au sens où la série converge vers  $f$  en norme  $L^2$ . En ce sens, la famille  $\{L_k\}_{k \geq 0}$  constitue une *base orthonormée* pour décrire les fonctions continues sur  $[a, b]$ . En combinant le fait que  $\|f - q_n\|_2$  tend vers 0 avec l'égalité de Pythagore

$$\|f\|_2^2 = \|q_n\|_2^2 + \|f - q_n\|_2^2 = \sum_{k=0}^n |\langle f, L_n \rangle|^2 + \|f - q_n\|_2^2,$$

on obtient l'égalité dite de Parseval

$$\|f\|_2^2 = \sum_{k=0}^{+\infty} |\langle f, L_n \rangle|^2,$$

qui est classique pour les bases orthonormées en dimension finie. Un autre exemple de base orthonormée de fonctions  $\{f_k\}_{k \in \mathbb{Z}}$  est associé aux séries de Fourier dont la convergence peut s'écrire

$$f = \sum_{k \in \mathbb{Z}} \langle f, f_k \rangle f_k,$$

où  $\langle \cdot, \cdot \rangle$  désigne le produit scalaire  $L^2$  sur  $[-\pi, \pi]$  et  $f_k(x) := (2\pi)^{-1/2} e^{ikx}$ . Le concept général de base orthonormée en dimension infinie peut-être rendu plus rigoureux dans le cadre des espaces de Hilbert qui n'est pas abordé dans ce cours. Il est intéressant de remarquer qu'une base orthonormée telle que  $L_n$  est une suite uniformément bornée en norme  $L^2$  puisque  $\|L_n\|_2 = 1$  mais que pour tout  $n \neq m$  on a  $\|L_n - L_m\|_2 = \sqrt{2}$  ce qui entraîne qu'on ne peut pas en extraire de sous-suite convergente. Ceci traduit le fait qu'en dimension infinie un ensemble fermé et borné n'est pas nécessairement compact.

**Remarque 3.6.2** Les polynômes de Legendre sont plus usuellement définis sur l'intervalle  $[a, b] = [-1, 1]$  et renormalisés de manière à ce que  $L_n(1) = 1$  pour tout  $n$  (il s'agit donc d'une suite de polynômes orthogonaux mais non-orthonormés). On peut facilement établir quelques propriétés importantes de cette famille, en particulier la formule de Rodrigues

$$L_n(x) = \frac{(-1)^n}{2^n n!} \frac{d^n}{dx^n} \left( (1-x^2)^n \right),$$

et la formule de récurrence

$$L_{n+1}(x) = \frac{2n+1}{n+1} x L_n(x) - \frac{n}{n+1} L_{n-1}(x),$$

initialisée par  $L_0(x) = 1$  et  $L_1(x) = x$ .

**Remarque 3.6.3** En utilisant le changement de variable  $x = \cos(t)$ , on voit que les polynômes de Tchebychev abordés dans la preuve du théorème de Weierstrass vérifient pour tout  $m \neq n$

$$\begin{aligned} \int_{-1}^1 T_n(x) T_m(x) (1-x^2)^{-1/2} dx &= - \int_{-\pi}^{\pi} T_n(\cos(t)) T_m(\cos(t)) dt \\ &= - \int_{-\pi}^{\pi} \cos(nt) \cos(mt) dt = 0. \end{aligned}$$

Il s'agit donc d'une suite de polynômes orthogonaux au sens du produit scalaire

$$\langle f, g \rangle := \int_{-1}^1 f(x) g(x) (1-x^2)^{-1/2} dx.$$

Plus généralement, la théorie des polynômes orthogonaux établit l'existence de bases orthonormées de polynômes pour un produit scalaire de type

$$\langle f, g \rangle := \int_I f(x) g(x) w(x) dx.$$

où  $I$  est un intervalle borné ou non, et  $w(x)$  une fonction positive telle que  $\int_I |x|^n w(x) dx < \infty$  pour tout  $n \geq 0$ . Citons en particulier les polynômes de Hermite ( $I = \mathbb{R}$  et  $w(x) = e^{-x^2}$ ) et de Laguerre ( $I = [0, +\infty[$  et  $w(x) = e^{-x}$ ).

### 3.7 Interpolation polynomiale par morceaux

Nous avons observé que l'interpolation polynomiale sur un intervalle  $[a, b]$  fait apparaître des problèmes de stabilité lorsque l'on fait tendre le degré  $n$  vers  $+\infty$ , en particulier si l'on choisi des points d'interpolation équidistants. Un procédé alternatif permettant d'éviter ces difficultés, et très utilisé en pratique, consiste à découper l'intervalle en morceaux en appliquant sur chacun d'entre eux un procédé d'interpolation de degré fixé. On fait ensuite tendre la taille de ces morceaux vers 0.

Plus précisément, pour  $n \geq 0$  on se donne une subdivision

$$a = a_0 < a_1 < \cdots < a_{n-1} < a_n = b,$$

et on définit sa finesse par

$$h = \max_{i=0, \dots, n-1} (a_{i+1} - a_i).$$

En se fixant un entier  $m > 0$ , on se donne sur chacun des intervalles  $[a_i, a_{i+1}]$  une autre subdivision

$$a_i = x_{i,0} < x_{i,1} < \cdots < x_{i,m-1} < x_{i,m} = a_{i+1}.$$

Un choix classique consiste à prendre des points équidistants sur chaque intervalle  $[a_i, a_{i+1}]$  c'est à dire  $x_{i,j} = a_i + \frac{j}{m}(a_{i+1} - a_i)$ . Etant donné une fonction  $f$ , on peut alors définir son interpolation polynomiale de Lagrange de degré  $m$  par morceaux sur la subdivision  $a_0, \dots, a_n$ , comme l'unique fonction  $f_n$  dont la restriction à chaque intervalle  $[a_i, a_{i+1}]$  est un polynôme de degré  $m$  et qui vérifie

$$f_n(x_{i,j}) = f(x_{i,j}), \quad i = 0, \dots, n-1, \quad j = 0, \dots, m.$$

On remarque que dans le cas  $m = 1$  qui correspond à l'interpolation affine par morceaux, il s'agit tout simplement de l'approximation du graphe de  $f$  par une "ligne brisée" aux points  $(a_i, f(a_i))$ , et on remarque que  $f_n$  est continue. Plus généralement on note que l'interpolation polynomiale par morceaux se raccorde de façon continue aux points  $a_i$  car on a

$$x_{i,m} = x_{i+1,0} = a_{i+1}.$$

En utilisant sur l'intervalle  $[a_i, a_{i+1}]$  l'estimation de l'erreur d'interpolation établie dans la section §3.4, on trouve que si  $f$  est de classe  $\mathcal{C}^{m+1}$  on a pour tout  $x \in [a_i, a_{i+1}]$ ,

$$|f(x) - f_n(x)| \leq \frac{1}{(m+1)!} \max_{t \in [a_i, a_{i+1}]} |f^{(m+1)}(t)| (a_{i+1} - a_i)^{m+1},$$

ce qui entraîne pour la norme sup sur  $[a, b]$

$$\|f - f_n\| \leq C_m \|f^{(m+1)}\| h^{m+1},$$

avec  $C_m = \frac{1}{(m+1)!}$ . Lorsque l'on choisit une subdivision uniforme de  $[a, b]$  c'est à dire  $a_i = a + \frac{i}{n}(b-a)$  on a  $h = (b-a)/n$  et par conséquent l'estimation d'erreur prend la forme

$$\|f - f_n\| \leq C_m \|f^{(m+1)}\| n^{-(m+1)},$$

avec  $C_m = \frac{(b-a)^{m+1}}{(m+1)!}$ . Notons que le nombre total de valeurs de  $f$  nécessaires pour définir  $f_n$  est égal à  $nm+1$  qui est le cardinal de l'ensemble  $\Gamma_n$  de tous les points  $x_{i,j}$  (en ne comptant pas deux fois les points  $x_{i,m}$  et  $x_{i+1,0}$  qui coïncident). On peut décomposer  $f_n$  suivant

$$f_n(x) = \sum_{\gamma \in \Gamma_n} f(\gamma) \ell_\gamma(x),$$

où  $\ell_\gamma(x)$  est l'unique fonction polynomiale de degré  $m$  par morceaux sur les intervalles  $[a_i, a_{i+1}]$  qui vérifie  $\ell_\gamma(\gamma) = 1$  et  $\ell_\gamma(\mu) = 0$  pour  $\mu \in \Gamma_n - \{\gamma\}$ . On peut vérifier que les fonctions  $\ell_\gamma$  constituent une base de l'espace vectoriel des fonctions polynomiale de degré  $m$  par morceaux sur les intervalles  $[a_i, a_{i+1}]$  et

continues sur  $[a, b]$ . Dans le cas  $m = 1$ , l'ensemble  $\Gamma_n$  coïncide avec  $\{a_0, \dots, a_n\}$  et le graphe de la fonction de base  $\ell_{a_i}$  a la forme d'un "chapeau" à support dans  $[a_{i-1}, a_{i+1}]$ .

L'interpolation polynomiale de Lagrange par morceaux est globalement continue sur  $[a, b]$  mais les dérivées de  $f_n$  sont en général discontinues aux points de raccords  $a_i$  entre les polynômes, ce qui signifie que l'approximation n'est pas de classe  $\mathcal{C}^1$ . Il est possible d'obtenir des approximations polynomiales par morceaux plus régulières en utilisant sur chaque intervalle  $[a_i, a_{i+1}]$  l'interpolation de Hermite que nous avons introduit dans la section §3.3. Plus précisément on définit l'interpolation de Hermite par morceaux de degré  $2m + 1$  comme l'unique fonction  $f_n$  dont la restriction à chaque intervalle  $[a_i, a_{i+1}]$  est un polynôme de degré  $2m + 1$  et qui vérifie

$$f_n^{(k)}(a_i) = f^{(k)}(a_i), \quad i = 0, \dots, n, \quad k = 0, \dots, m$$

Il est immédiat de vérifier que la fonction  $f_n$  ainsi définie est de classe  $\mathcal{C}^m$  sur  $[a, b]$ . En analysant l'erreur du procédé d'interpolation de Hermite, on peut prouver qu'elle vérifie une estimation d'erreur sur  $[a, b]$  du type

$$\|f - f_n\| \leq C_m \|f^{(2m+2)}\| h^{2m+2},$$

où la constante  $C_m$  ne dépend que de  $m$ .

Nous terminons en évoquant un procédé d'approximation très utilisé pour la modélisation géométrique des courbes : les *fonctions splines*. Etant donné une subdivision  $a = a_0 < \dots < a_n = b$ , on dit qu'une fonction  $g$  est une spline d'ordre  $m$  sur  $[a, b]$  pour cette subdivision, si sa restriction à chaque intervalle  $[a_i, a_{i+1}]$  est un polynôme de degré  $m$  et si  $g$  est globalement de classe  $\mathcal{C}^{m-1}$  sur  $[a, b]$ . Un résultat important, et facile à démontrer, est que l'on peut décrire toutes les fonctions de ce type comme des combinaisons linéaires des fonctions élémentaires

$$x \mapsto (x - a_i)_+^m = \left( \max\{0, (x - a_i)\} \right)^m, \quad i = 1, \dots, n - 1,$$

ainsi que des fonctions  $x \mapsto x^k$  pour  $k = 0, \dots, m$ . L'ensemble de ces fonctions constitue une base de l'espace des splines d'ordre  $m$  sur  $[a, b]$  pour la subdivision  $a_1, \dots, a_n$ , qui est donc de dimension  $n + m$ . En pratique, on décrit souvent les fonctions splines en utilisant une autre base constituée de fonctions dont des supports sont mieux localisés autour des points  $a_i$  : pour  $i = 1, \dots, n + m$ , il existe une fonction spline  $B_i$  dite *B-spline* dont le support est contenu dans l'intervalle  $[a_{i-m-1}, a_i]$ , en posant  $a_{i-m-1} = a$  si  $i \leq m$  et  $a_i = b$  si  $i \geq n$ . L'ensemble de ces fonctions constitue une base de l'espace des splines d'ordre  $m$  sur  $[a, b]$  pour la subdivision  $a_1, \dots, a_n$ . Dans le cas  $m = 1$  on retrouve les fonctions de base pour l'interpolation affine par morceaux.

Un résultat important et difficile à prouver est l'existence et l'unicité d'une spline d'interpolation dans le cas où  $m$  est impair.

**Théorème 3.7.1** *Si  $m$  est impair, pour tout ensemble de réels  $\{y_0, \dots, y_n\}$  avec  $n \geq m$  et  $\{\alpha_1, \dots, \alpha_{m-1}\}$ , il existe une unique fonction spline  $f_n$  d'ordre  $m$  sur  $[a, b]$  pour la subdivision  $a_0, \dots, a_n$  telle que*

$$f_n(a_i) = y_i, \quad i = 0, \dots, n \quad \text{et} \quad f_n^{(k)}(a) = \alpha_k, \quad k = 1, \dots, m - 1.$$

*En particulier si  $f$  est une fonction continue, il existe une unique spline d'interpolation  $f_n$  d'ordre  $m$  définie par*

$$f_n(a_i) = f(a_i), \quad i = 0, \dots, n \quad \text{et} \quad f_n^{(k)}(a) = \alpha_k, \quad k = 1, \dots, m - 1.$$

Une variante de ce résultat affirme l'existence et l'unicité d'une spline d'interpolation périodique.

**Théorème 3.7.2** *Si  $m$  est impair, pour tout ensemble de réels  $\{y_0, \dots, y_{n-1}\}$  avec  $n \geq m$ , il existe une unique fonction spline  $f_n$  d'ordre  $m$  sur  $[a, b]$  pour la subdivision  $a_0, \dots, a_n$  telle que*

$$f_n(a_i) = y_i, \quad i = 1, \dots, n \quad \text{et} \quad f_n^{(k)}(a) = f_n^{(k)}(b), \quad k = 0, \dots, m - 1.$$

*En particulier si  $f$  est une fonction continue telle que  $f(a) = f(b)$ , il existe une unique spline d'interpolation  $f_n$  d'ordre  $m$  définie par*

$$f_n(a_i) = f(a_i), \quad i = 0, \dots, n \quad \text{et} \quad f_n^{(k)}(a) = f_n^{(k)}(b), \quad k = 1, \dots, m - 1.$$

## 4 Calcul approché des intégrales

Le calcul exact de l'intégrale d'une fonction  $f$  sur un intervalle  $[a, b]$  est possible lorsqu'on dispose d'une primitive de cette fonction. Cela n'est pas toujours le cas, même pour des fonctions très simples telles que  $x \mapsto e^{-x^2}$  dont la primitive ne s'exprime pas sous une forme explicite permettant de la calculer. On peut alors chercher à calculer une approximation de  $\int_a^b f(x)dx$  au moyen d'une formule numérique faisant intervenir les valeurs de  $f$  en certains points de l'intervalle  $[a, b]$ . De telles formules sont appelées *quadratures*. Notons qu'une somme de Riemann

$$\Sigma(f) = \sum_{i=0}^{k-1} (a_{i+1} - a_i) f(x_i),$$

avec  $a = a_0 < \dots < a_k = b$  et  $x_i \in [a_i, a_{i+1}]$  est un exemple d'une telle formule. Ce chapitre présente les quadratures les plus employées en pratique, ainsi qu'une analyse de la précision avec laquelle elle permettent d'approcher une intégrale.

### 4.1 Méthodes de quadrature simples et composées

Dans la somme de Riemann ci-dessus, la quantité  $(a_{i+1} - a_i)f(x_i)$  peut être vue comme une approximation de l'intégrale  $\int_{a_i}^{a_{i+1}} f(x)dx$ . Pour étudier cette approximation en simplifiant les notations, on se place sur un intervalle  $[a, b]$  et on étudie l'approximation de  $\int_a^b f(x)dx$  par la formule de quadrature

$$(b - a)f(c),$$

avec  $c \in [a, b]$ . Les trois choix de  $c$  les plus communément employés sont

1. La quadrature du rectangle à gauche :  $c = a$ .
2. La quadrature du rectangle à droite :  $c = b$ .
3. La quadrature du point milieu :  $c = \frac{a+b}{2}$ .

Si  $f$  est de classe  $\mathcal{C}^1$  sur  $[a, b]$ , on peut estimer l'erreur pour la formule du rectangle à gauche en écrivant

$$\int_a^b f(x)dx = \int_a^b \left( f(a) + \int_a^x f'(t)dt \right) dx = (b - a)f(a) + \int_a^b \int_a^x f'(t)dt dx,$$

d'où

$$\left| \int_a^b f(x)dx - (b - a)f(a) \right| \leq \int_a^b \int_a^x |f'(t)| dt dx \leq \|f'\| \int_a^b \int_a^x dt dx = \frac{(b - a)^2}{2} \|f'\|,$$

avec  $\|f'\|$  la norme sup de  $f'$  sur  $[a, b]$ . Un calcul similaire donne la même estimation pour la formule du rectangle à droite. Si on revient maintenant à la somme de Riemann

$$\Sigma(f) = \sum_{i=0}^{k-1} (a_{i+1} - a_i) f(x_i),$$

en prenant  $x_i = a_i$  ou  $x_i = a_{i+1}$ , on obtient la formule dite des rectangles à gauche ou à droite, dont on peut estimer l'erreur en la décomposant sur chaque intervalle suivant

$$\left| \int_a^b f(x)dx - \Sigma(f) \right| \leq \sum_{i=0}^{k-1} \left| \int_{a_i}^{a_{i+1}} f(t)dt - (a_{i+1} - a_i) f(x_i) \right| \leq \frac{1}{2} \|f'\| \sum_{i=0}^{k-1} (a_{i+1} - a_i)^2.$$

En notant  $h = \max |a_{i+1} - a_i|$  la finesse de la subdivision, on obtient ainsi l'estimation

$$\left| \int_a^b f(x)dx - \Sigma(f) \right| \leq \frac{1}{2} \|f'\| \left( \sum_{i=0}^{k-1} (a_{i+1} - a_i) \right) h = \frac{b - a}{2} \|f'\| h.$$

Dans le cas de la formule du point milieu, on peut améliorer l'estimation d'erreur en remarquant que si  $f$  est de classe  $\mathcal{C}^2$  sur  $[a, b]$ , on a d'après la formule de Taylor avec reste intégral

$$\begin{aligned}\int_a^b f(x)dx &= \int_a^b \left( f\left(\frac{a+b}{2}\right) + \left(x - \frac{a+b}{2}\right)f'\left(\frac{a+b}{2}\right) + \int_{\frac{a+b}{2}}^x (x-t)f''(t)dt \right) dx \\ &= (b-a)f\left(\frac{a+b}{2}\right) + \int_a^b \int_{\frac{a+b}{2}}^x (x-t)f''(t)dt dx,\end{aligned}$$

d'où

$$\left| \int_a^b f(t)dt - (b-a)f\left(\frac{a+b}{2}\right) \right| \leq \|f''\| \int_a^b \int_{\frac{a+b}{2}}^x (x-t)dt dx = \frac{(b-a)^3}{24} \|f''\|,$$

Si l'on revient à la somme de Riemann avec  $x_i = \frac{a_i + a_{i+1}}{2}$ , on obtient ainsi

$$\left| \int_a^b f(x)dx - \Sigma(f) \right| \leq \frac{1}{24} \|f''\| \sum_{i=0}^{k-1} (a_{i+1} - a_i)^3 \leq \frac{b-a}{24} \|f''\| h^2,$$

qui est une meilleure estimation que celle de la méthode des rectangles lorsque  $h \rightarrow 0$ . Le principe qui généralise l'analyse ci-dessus est le suivant :

1. On part d'une quadrature dite "simple" qui donne une approximation de  $\int_a^b f(t)dt$  utilisant l'évaluation de  $f$  en un petit nombre de points sur  $[a, b]$ .
2. On en déduit une quadrature dite "composée" sur  $[a, b]$  en sommant les intégrales approchées par la quadrature simple sur chaque intervalle  $[a_i, a_{i+1}]$  d'une subdivision  $a = a_0 < \dots < a_k = b$ .

Une quadrature simple fréquemment utilisée consiste à approcher  $\int_a^b f(x)dx$  par l'intégrale de son polynôme d'interpolation affine  $p_1(t)$  aux points  $a$  et  $b$ , c'est à dire par

$$\int_a^b p_1(x)dx = \int_a^b \left( f(a) + (t-a) \frac{f(b) - f(a)}{b-a} \right) dt = (b-a) \frac{f(a) + f(b)}{2}.$$

La quadrature composée associée est appelée *formule des trapèze* et est donnée par la somme

$$T(f) = \sum_{i=0}^{k-1} (a_{i+1} - a_i) \frac{f(a_i) + f(a_{i+1})}{2}$$

En utilisant la formule d'erreur établie dans la section §3.4, on obtient pour la quadrature simple

$$\left| \int_a^b (f(x) - p_1(x))dx \right| \leq \|f'\| \int_a^b |\Pi_1(x)|dx = \|f'\| \frac{1}{2} \int_a^b (x-a)(b-x)dx = \frac{(b-a)^3}{12} \|f''\|,$$

et on en déduit pour la formule des trapèze l'estimation

$$\left| \int_a^b f(x)dx - T(f) \right| \leq \frac{(b-a)}{12} \|f''\| h^2,$$

qui est du même ordre que celle de la somme de Riemann avec règle du point milieu. On peut généraliser cette construction en remplaçant dans la quadrature simple la fonction affine  $p_1$  par le polynôme d'interpolation  $p_n \in \mathcal{P}_n$  pour la subdivision équidistante  $x_j = a + \frac{j}{n}(b-a)$  pour  $j = 0, \dots, n$ . La règle de quadrature simple peut s'écrire

$$\int_a^b p_n(x)dx = \sum_{j=0}^n \left( \int_a^b \ell_j(x)dx \right) f(x_j).$$

où les  $\ell_j$  sont les fonctions de bases de Lagrange aux points  $x_j$ . C'est la *quadrature de Newton-Cotes* de degré  $n$ . En utilisant le changement de variable  $x = \phi(y) = \frac{a+b}{2} + \frac{b-a}{2}y$ , on peut se ramener sur l'intervalle  $[-1, 1]$  en posant  $\tilde{p}_n = p_n \circ \phi$  ce qui donne

$$\int_a^b p_n(x)dx = \frac{b-a}{2} \int_{-1}^1 \tilde{p}_n(y)dy.$$

On remarque que  $\tilde{p}_n \in \mathcal{P}_n$  est le polynôme d'interpolation de  $\tilde{f} = f \circ \phi$  aux points équidistants  $y_i = -1 + \frac{2i}{n}$  puisque

$$\tilde{p}_n(y_j) = p_n(\phi(y_j)) = p_n(x_j) = f(x_j) = \tilde{f}(y_j).$$

Par conséquent, on peut écrire

$$\int_a^b p_n(x) dx = \frac{b-a}{2} \sum_{j=0}^n \omega_j f(x_j), \quad \omega_j = \int_{-1}^1 \ell_j(y) dy,$$

où  $\ell_j$  désigne ici la fonction de base de Lagrange sur associée au point  $y_j$  de la subdivision uniforme de  $[-1, 1]$ . Les poids  $\omega_j$  sont donc indépendants de  $a$  et  $b$ . Pour  $n = 2$  on trouve les poids  $(\frac{1}{3}, \frac{4}{3}, \frac{1}{3})$  qui donnent la *formule de Simpson*

$$\int_a^b p_2(x) dx = (b-a) \left( \frac{1}{6} f(a) + \frac{2}{3} f\left(\frac{a+b}{2}\right) + \frac{1}{6} f(b) \right),$$

dont la version composée est

$$S(f) = \sum_{i=0}^{k-1} (a_{i+1} - a_i) \left( \frac{1}{6} f(a_i) + \frac{2}{3} f\left(\frac{a_i + a_{i+1}}{2}\right) + \frac{1}{6} f(a_{i+1}) \right).$$

Pour  $n = 4$ , on trouve les poids  $(\frac{7}{45}, \frac{32}{45}, \frac{4}{15}, \frac{32}{45}, \frac{7}{45})$  qui donnent la *formule de Boole-Villarceau*. On note que si on interpole la fonction constante  $f = 1$ , on a toujours  $p_n = f$  et la quadrature est alors exacte, ce qui entraîne que l'on a nécessairement

$$\sum_{j=0}^n \omega_j = 2.$$

## 4.2 Etude de convergence

On s'intéresse à des formules de quadratures de la forme

$$\frac{b-a}{2} \sum_{j=0}^n \omega_j f(x_j),$$

pour l'approximation de  $\int_a^b f(x) dx$ , où les  $x_j$  sont de la forme  $x_j = \phi(y_j) = \frac{a+b}{2} + \frac{b-a}{2} y_j$  avec  $\{y_0, \dots, y_n\}$  fixés dans l'intervalle  $[-1, 1]$ , et où les poids  $\omega_j$  vérifient la relation  $\sum_{j=0}^n \omega_j = 2$ . Toutes les méthodes de la section précédente sont de ce type.

**Définition 4.2.1** La quadrature est dite d'ordre  $m$  si et seulement si elle est exacte pour les polynômes de degré inférieur ou égal à  $m$  : pour tout  $p \in \mathcal{P}_m$  on a

$$\int_a^b p(x) dx = \frac{b-a}{2} \sum_{j=0}^n \omega_j p(x_j),$$

**Remarque 4.2.1** Afin de vérifier qu'une quadrature sur  $[a, b]$  est d'ordre  $m$ , il suffit de le vérifier pour sur l'intervalle  $[-1, 1]$ . En effet si la quadrature  $\sum_{j=0}^n \omega_j f(x_j)$  sur  $[-1, 1]$  est d'ordre  $m$ , on a alors pour tout  $p \in \mathcal{P}_m$ , en posant  $\tilde{p} = p \circ \phi \in \mathcal{P}_m$ ,

$$\frac{b-a}{2} \sum_{j=0}^n \omega_j p(x_j) = \frac{b-a}{2} \sum_{j=0}^n \omega_j \tilde{p}(y_j) = \frac{b-a}{2} \int_{-1}^1 \tilde{p}(y) dy = \int_a^b p(x) dx.$$

D'autre part, par linéarité, il est suffisant de vérifier que la quadrature est exacte pour les fonctions  $x \mapsto x^k$  pour  $k = 0, \dots, m$ . Une quadrature est donc d'ordre  $m$  si on a

$$\sum_{j=0}^n \omega_j y_j^k = \int_{-1}^1 y^k dy = \frac{1 + (-1)^k}{k+1}, \quad k = 0, \dots, m.$$

En utilisant les remarques ci-dessus, on vérifie que les quadratures des rectangles à gauche et à droite sont d'ordre  $m = 0$ , celles du point milieu et du trapèze sont d'ordre  $m = 1$ , et il est aisé de vérifier qu'elles ne sont pas d'ordre supérieur. Les quadratures de Newton-Cotes de degré  $n$  sont clairement d'ordre  $n$ . En utilisant la symmétrie des points d'interpolation sur  $[-1, 1]$ , on voit que si  $n$  est pair la quadrature est aussi exacte pour  $x \mapsto x^{n+1}$  et elle est donc d'ordre  $n + 1$ . Ainsi la règle de Simpson est d'ordre 3, celle de Boole-Villarceau d'ordre 5, etc.

Afin d'étudier l'erreur d'une méthode quadrature donnée, on note

$$E(f) = \int_a^b f(x)dx - \frac{b-a}{2} \sum_{j=0}^n \omega_j f(x_j),$$

et on introduit pour tout  $t \in \mathbb{R}$  et  $m \in \mathbb{N}$  la fonction

$$g_{t,m}(x) := (x-t)_+^m = \left(\max\{x-t, 0\}\right)^m$$

**Théorème 4.2.1** *Si la méthode de quadrature est d'ordre  $m$ , on a pour toute fonction  $f \in \mathcal{C}^{m+1}([a, b])$*

$$E(f) = \frac{1}{m!} \int_a^b K_m(t) f^{(m+1)}(t) dt,$$

où  $K_m(t)$  est l'erreur pour fonction  $g_{t,m}$ , i.e.

$$K_m(t) := E(g_{t,m}) = \int_a^b g_{t,m}(x) dx - \frac{b-a}{2} \sum_{j=0}^n \omega_j g_{t,m}(x_j).$$

La fonction  $K_m$  est appelée "noyau de Peano" de la quadrature.

**Preuve :** En utilisant la formule de Taylor avec reste intégral au point  $a$  on a

$$f(x) = p(x) + \frac{1}{m!} \int_a^x (x-t)^m f^{(m+1)}(t) dt = p(x) + \frac{1}{m!} \int_a^b g_{t,m}(x) f^{(m+1)}(t) dt$$

avec  $p(x) = \sum_{k=0}^m \frac{1}{k!} f^{(k)}(a)(x-a)^k$ . Ceci entraîne d'une part que

$$\int_a^b f(x) dx = \int_a^b p(x) dx + \frac{1}{m!} \int_a^b \left( \int_a^b g_{t,m}(x) dx \right) f^{(m+1)}(t) dt,$$

et d'autre part que

$$\frac{b-a}{2} \sum_{j=0}^n \omega_j f(x_j) = \frac{b-a}{2} \sum_{j=0}^n \omega_j p(x_j) + \frac{1}{m!} \int_a^b \left( \frac{b-a}{2} \sum_{j=0}^n \omega_j g_{t,m}(x_j) \right) f^{(m+1)}(t) dt.$$

En soustrayant ces deux identités et en utilisant le fait que la quadrature est exacte pour  $p$ , on obtient le résultat annoncé.  $\square$

Une conséquence immédiate de ce résultat est que pour toute fonction  $f \in \mathcal{C}^{m+1}([a, b])$ , on a

$$|E(f)| \leq \frac{\int_a^b |K_m(t)| dt}{m!} \|f^{(m+1)}\|,$$

où  $\|f^{(m+1)}\|$  est la norme sup de  $f^{(m+1)}$  sur  $[a, b]$ . Afin d'évaluer l'intégrale  $\int_a^b |K_m(t)| dt$ , on utilise le changement de variable  $y = \phi(x)$  pour relier  $K_m$  au noyau de Peano  $k_m$  pour la quadrature sur  $[-1, 1]$ ,

$$k_m(t) := \int_{-1}^1 g_{t,m}(x) dx - \sum_{j=0}^n \omega_j g_{t,m}(x_j).$$

Pour cela on pose  $\tilde{g}_{m,t}(y) = g_{m,t} \circ \phi$  et on obtient d'abord

$$K_m(t) = \frac{b-a}{2} \left( \int_{-1}^1 \tilde{g}_{t,m}(y) dy - \sum_{j=0}^n \omega_j \tilde{g}_{t,m}(y_j) \right).$$

On remarque ensuite que

$$\tilde{g}_{\phi(t),m}(y) = g_{\phi(t),m}(\phi(y)) = (\phi(y) - \phi(t))_+^m = \left( \frac{b-a}{2}(y-t) \right)_+^m = \left( \frac{b-a}{2} \right)^m g_{t,m}(y),$$

et par conséquent

$$K_m(\phi(t)) = \left( \frac{b-a}{2} \right)^{m+1} k_m(t),$$

c'est à dire  $K_m \circ \phi = \left( \frac{b-a}{2} \right)^{m+1} k_m$ . Ceci entraîne immédiatement

$$\int_a^b |K_m(t)| dt = \left( \frac{b-a}{2} \right)^{m+2} \int_{-1}^1 |k_m(t)| dt.$$

Nous avons donc établi le résultat suivant.

**Corollaire 4.2.1** *Si la méthode de quadrature est d'ordre  $m$ , on a pour toute fonction  $f \in \mathcal{C}^{m+1}([a, b])$*

$$|E(f)| \leq C(b-a)^{m+2} \|f^{(m+1)}\|,$$

avec

$$C := \frac{1}{m! 2^{m+2}} \int_{-1}^1 |k_m(t)| dt.$$

Une conséquence immédiate porte sur la méthode de quadrature composée obtenue à partir de la quadrature étudiée c'est à dire

$$Q(f) = \sum_{i=0}^{k-1} \frac{a_{i+1} - a_i}{2} \sum_{j=0}^n \omega_j f(x_{i,j}),$$

avec  $x_{i,j} = \frac{a_i + a_{i+1}}{2} + \frac{a_{i+1} - a_i}{2} y_j$ , dont on peut évaluer la précision en sommant les estimations d'erreur obtenue sur chaque intervalle  $[a_i, a_{i+1}]$  comme on l'a déjà fait dans la section précédente pour les méthodes des rectangles et des trapezes.

**Corollaire 4.2.2** *Si la méthode de quadrature simple est d'ordre  $m$ , on a pour tout  $f \in \mathcal{C}^{m+1}([a, b])$*

$$\left| \int_a^b f(x) dx - Q(f) \right| \leq C(b-a) \|f^{(m+1)}\| h^{m+1},$$

avec  $C := \frac{1}{m! 2^{m+2}} \int_{-1}^1 |k_m(t)| dt$  et  $h = \max |a_{i+1} - a_i|$ .

Le calcul de la constante  $C$  peut être facilité lorsque le noyau de Peano  $k_m(t)$  est de signe constant sur  $[-1, 1]$ . On a dans ce cas

$$C = \frac{1}{m! 2^{m+2}} \left| \int_{-1}^1 k_m(t) dt \right|,$$

et d'autre part, pour la fonction  $e_{m+1}(x) := x^{m+1}$ , la formule d'erreur du Théorème 4.2.1 pour la quadrature sur l'intervalle  $[-1, 1]$  donne exactement

$$E(e_{m+1}) = (m+1) \int_{-1}^1 k_m(t) dt.$$

On peut donc écrire

$$C = \frac{1}{(m+1)!2^{m+2}} |E(e_{m+1})| = \frac{1}{(m+1)!2^{m+2}} \left| \frac{1 + (-1)^{m+1}}{m+2} - \sum_{j=0}^n \omega_j y_j^{m+1} \right|$$

Dans le cas de la quadrature du trapeze, un calcul simple montre que  $k_1(t) = \frac{1}{2}(t^2 - 1)$  si  $|t| \leq 1$  et  $k_1(t) = 0$  si  $|t| > 1$ . Le noyau de Peano est de signe constant ce qui conduit à

$$C := \frac{1}{16} \left| \frac{2}{3} - 2 \right| = \frac{1}{12}.$$

On retrouve ainsi pour la quadrature composée l'estimation

$$\left| \int_a^b f(x) dx - T(f) \right| \leq \frac{b-a}{12} \|f''\| h^2$$

Plus généralement, il est possible de prouver que le noyau de Peano  $k_m(t)$  est de signe constant sur  $[-1, 1]$  pour toutes les méthodes de Newton-Cotes. Dans le cas de la formule de Simpson, on obtient ainsi

$$C = \frac{1}{768} \left| \frac{2}{5} - \frac{2}{3} \right| = \frac{1}{2880}.$$

### 4.3 Les méthodes de Gauss

On a vu que les méthodes de quadrature de Newton-Cotes de degré  $n$  utilisent  $n+1$  points et sont d'ordre  $n$  ou  $n+1$  suivant la parité de  $n$ . La recherche de la quadrature ayant l'ordre le plus élevé possible pour un nombre de points prescrit conduit à la *méthode de Gauss-Legendre*.

**Théorème 4.3.1** *Il existe une unique formule de quadrature sur  $[-1, 1]$  de la forme*

$$\sum_{j=0}^n \omega_j f(x_j),$$

qui soit d'ordre  $2n+1$ , c'est à dire exacte si  $f \in \mathcal{P}_{2n+1}$ . Les poids  $\omega_i$  sont positifs.

**Preuve :** On montre d'abord que si une telle quadrature existe elle est unique. Soit  $p_{n+1}$  le polynôme de degré  $n+1$  défini par

$$p_{n+1}(x) = \prod_{j=0}^n (x - x_j).$$

Pour tout  $q \in \mathcal{P}_n$ , le produit  $p_{n+1}q$  est de degré  $2n+1$  et par conséquent

$$\int_{-1}^1 p_{n+1}(x)q(x)dx = \sum_{j=0}^n \omega_j p_{n+1}(x_j)q(x_j) = 0.$$

Ceci montre que  $p_{n+1}$  est orthogonal à  $\mathcal{P}_n$  au sens du produit scalaire  $\langle u, v \rangle = \int_{-1}^1 u(x)v(x)dx$ . Par conséquent on a

$$p_{n+1} = \alpha_{n+1} L_{n+1},$$

où  $L_{n+1}$  est le polynôme de Legendre de degré  $n+1$  introduit dans la section §3.6 et  $\alpha_{n+1}$  est un nombre tel que le coefficient directeur de  $p_{n+1}$  est égal à 1. Les points  $x_i$  sont donc uniquement déterminés : ce sont les racines de  $L_{n+1}$ . En introduisant les fonctions de bases de Lagrange

$$\ell_i(x) := \prod_{j \neq i} \frac{x - x_j}{x_i - x_j},$$

qui sont dans  $\mathcal{P}_n$  et donc a-fortiori dans  $\mathcal{P}_{2n+1}$ , on obtient

$$\int_{-1}^1 \ell_i(x) dx = \sum_{j=0}^n \omega_j \ell_i(x_j) = \omega_i,$$

ce qui montre que les poids  $\omega_j$  sont aussi uniquement déterminés. L'unicité est donc établie. Montrons à présent que pour ce choix des points  $x_j$  et des poids  $\omega_j$ , la quadrature est en effet exacte pour les polynômes de degré inférieur ou égal à  $2n + 1$ . Si  $p$  est un tel polynôme on peut effectuer sa division euclidienne par  $p_{n+1}$  et l'écrire

$$p = qp_{n+1} + r,$$

avec  $q, r \in \mathcal{P}_n$ . Comme  $q$  et  $p_{n+1}$  sont orthogonaux, on a

$$\int_{-1}^1 p(x) dx = \int_{-1}^1 r(x) dx.$$

Puisque  $r \in \mathcal{P}_n$  on peut le décomposer suivant

$$r(x) = \sum_{j=0}^n r(x_j) \ell_j(x),$$

et par conséquent on a

$$\int_{-1}^1 p(x) dx = \sum_{j=0}^n \omega_j r(x_j) = \sum_{j=0}^n \omega_j p(x_j),$$

ce qui montre que la quadrature est exacte pour  $p$ . Finalement, on remarque que  $\ell_i^2 \in \mathcal{P}_{2n}$  et par conséquent

$$\int_{-1}^1 \ell_i(x)^2 dx = \sum_{j=0}^n \omega_j \ell_i(x_j)^2 = \omega_i,$$

ce qui montre la positivité des poids  $\omega_i$ . □

Voici la forme explicite de la méthode de Gauss-Legendre pour des petites valeurs de  $n$  :

1. Pour  $n = 0$ , on trouve le point  $\{0\}$  et le poids  $\{2\}$  : c'est la méthode du point milieu.
2. Pour  $n = 1$ , on trouve les points  $\{-\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\}$  et les poids  $\{1, 1\}$ . La méthode est d'ordre 3.
3. Pour  $n = 2$ , on trouve les points  $\{-\sqrt{\frac{3}{5}}, 0, \sqrt{\frac{3}{5}}\}$  et les poids  $\{\frac{5}{9}, \frac{8}{9}, \frac{5}{9}\}$ . La méthode est d'ordre 5.

Il est possible de montrer que le noyau de Peano de la méthode de Gauss-Legendre est positif, ce qui permet comme pour les méthodes de Newton-Cotes de calculer facilement la constante  $C$  intervenant dans l'estimation d'erreur pour la quadrature simple et composée.

**Remarque 4.3.1** *De façon plus générale, les méthodes de quadrature de Gauss sont des quadratures visant à approcher l'intégrale*

$$\int_I f(x)w(x)dx,$$

où  $w$  est une fonction positive donnée, à l'aide de  $n + 1$  évaluations de  $f$  et qui sont exactes pour les polynômes de degré  $2n+1$ . La méthode de Gauss-Legendre correspond au cas  $w = 1$ . Les méthodes de Gauss plus générales utilisent les polynômes orthogonaux pour le produit scalaire  $\langle u, v \rangle := \int_I u(x)v(x)w(x)dx$ , qui ont été évoqués à la fin de la section §3.6.

## 5 Approximation des équations différentielles

### 5.1 Equations différentielles linéaires

Une équation différentielle est une équation du type

$$y'(t) = f(y(t), t),$$

dont la solution  $y : y \mapsto y(t)$  est une fonction de  $\mathbb{R}$  dans  $\mathbb{R}^n$  et où  $f : (x, t) \mapsto f(x, t)$  est une fonction continue de  $\mathbb{R}^n \times \mathbb{R}$  dans  $\mathbb{R}^n$ . Il peut arriver dans certains cas que  $f$  ne soit définie que sur  $U \times \mathbb{R}$  où  $U$  est un ouvert de  $\mathbb{R}^n$  appelé *espace des configurations*. Par analogie avec les modèles d'évolution temporelle qui constituent l'essentiel des exemples d'équations différentielles,  $t$  est souvent appelée "variable de temps". Si  $I \subset \mathbb{R}$  est un intervalle on dit que  $y$  est solution de l'équation ci-dessus sur  $I$  lorsqu'elle est dérivable et vérifie l'équation pour tout  $t$  dans l'intérieur  $\overset{\circ}{I}$  de  $I$ , et est continue sur  $I$ . Etant donné  $t_0 \in I$  et  $y_0 \in \mathbb{R}^n$ , le *problème de Cauchy* consiste à rechercher la solution de l'équation qui prend la valeur  $y_0$  au temps  $t_0$ , c'est à dire

$$y'(t) = f(y(t), t), \quad t \in \overset{\circ}{I}, \quad y(t_0) = y_0.$$

Par changement de variable on ramène souvent au cas où  $I = [0, T]$  et  $t_0 = 0$ , c'est à dire

$$y'(t) = f(y(t), t), \quad t \in ]0, T[ \quad y(t_0) = y_0.$$

La théorie permettant d'établir l'existence et l'unicité des solution des équation différentielles fait l'objet du cours de second semestre "Méthodes numériques pour les équations différentielles", et n'est pas abordée ici. L'objectif principal de ce bref chapitre est d'introduire des méthodes numériques permettant le calcul approché des solutions et d'analyser ces méthodes pour des cas très simples d'équations différentielles.

Une équation différentielle est dite *autonome* lorsque  $f$  ne dépend que de la variable  $x$ , et qu'elle peut donc s'écrire

$$y'(t) = f(y(t)).$$

L'équation est dite *linéaire* si et seulement si pour  $t$  fixé, l'application  $x \mapsto f(x, t)$  est affine, c'est à dire

$$f(x, t) = A(t)x + b(t),$$

où  $A : \mathbb{R} \mapsto \mathcal{M}_n(\mathbb{R})$  et  $b : \mathbb{R} \mapsto \mathbb{R}^n$  sont des fonctions continues. L'équation est dite *linéaire homogène* si et seulement si l'application  $x \mapsto f(x, t)$  est linéaire, c'est à dire de la forme ci-dessus avec  $b = 0$ . Les équations différentielles linéaires homogènes autonomes sont donc de la forme

$$y'(t) = Ay(t),$$

où  $A$  est une matrice  $n \times n$  donnée. Dans le cas  $n = 1$ , il est bien connu que le problème de Cauchy

$$y'(t) = ay(t), \quad y(0) = y_0,$$

admet une unique solution sur  $\mathbb{R}$  donnée par  $y(t) = ae^t$ . Ce résultat se généralise pour  $n > 1$  à l'aide de l'exponentielle matricielle introduite dans la section §1.4.

**Théorème 5.1.1** *Pour tout  $A \in \mathcal{M}_n(\mathbb{R})$  et  $y_0 \in \mathbb{R}^n$ , le problème de Cauchy*

$$y'(t) = Ay(t), \quad y(0) = y_0,$$

*admet une unique solution sur  $\mathbb{R}$  donnée par*

$$y(t) = \exp(tA)y_0.$$

*Plus généralement, si  $b \in \mathbb{R}^n$ , le problème de Cauchy*

$$y'(t) = Ay(t) + b, \quad y(0) = y_0,$$

*admet une unique solution sur  $\mathbb{R}$  donnée par*

$$y(t) = \exp(tA)y_0 + \int_0^t \exp((t-s)A)b ds.$$

**Preuve :** Pour l'équation linéaire homogène, on remarque que la fonction

$$t \mapsto e_A(t) := \exp(tA) = \sum_{k \geq 0} \frac{1}{k!} t^k A^k,$$

peut se dériver terme à terme et que l'on a

$$e'_A(t) = Ae_A(t).$$

En posant

$$z(t) = \exp(-tA)y(t),$$

on voit ainsi que si  $y$  est solution de l'équation si et seulement si

$$z'(t) = -A \exp(-tA)y(t) + \exp(-tA)y'(t) = \exp(-tA)(y'(t) - Ay(t)) = 0,$$

c'est à dire  $z(t)$  est une fonction constante  $z(t) = c$ . Les solutions de l'équation sont donc de la forme

$$y(t) = \exp(tA)c,$$

et la condition  $y(0) = y_0$  est vérifiée si et seulement  $c = y_0$ . Pour l'équation linéaire non-homogène, on utilise la méthode de la *variation de la constante* : en posant à nouveau

$$z(t) = \exp(-tA)y(t),$$

on voit que si  $y$  est solution de l'équation si et seulement si

$$z'(t) = \exp(-tA)b,$$

c'est à dire

$$z(t) = c + \int_0^t \exp(-sA)b \, ds.$$

Les solutions de l'équation sont donc de la forme

$$y(t) = \exp(tA)c + \int_0^t \exp((t-s)A)b \, ds.$$

et la condition  $y(0) = y_0$  est vérifiée si et seulement  $c = y_0$ . □

**Remarque 5.1.1** Certains modèles d'équations différentielles font intervenir des dérivées d'ordre supérieures de la fonction  $y$ , et ont la forme générale

$$y^{(m)}(t) = f(y(t), y'(t), \dots, y^{(m-1)}(t), t),$$

où  $f$  est une fonction continue allant de  $\mathbb{R}^n \times \dots \times \mathbb{R}^n \times \mathbb{R}$  dans  $\mathbb{R}^n$ . Le problème de Cauchy est alors assujéti de conditions du type

$$y(0) = y_0, \quad y'(0) = y_1, \dots, \quad y^{(m-1)}(0) = y_{m-1},$$

et on parle alors d'équation différentielles d'ordre  $m$ . De telles équations peuvent en fait se ramener à des équations d'ordre 1 de la forme

$$Y'(t) = F(Y(t), t),$$

en introduisant l'inconnue  $Y(t) := (y(t), y'(t), \dots, y^{(m-1)}(t)) \in \mathbb{R}^{mn}$ . Par exemple, pour  $A \in \mathcal{M}_n(\mathbb{R})$ , le problème de Cauchy pour une équation linéaire homogène du deuxième ordre du type

$$y''(t) = Ay(t), \quad y(0) = y_0, \quad y'(0) = y_1,$$

peut se reformuler sous la forme

$$Y'(t) = BY(t) \quad Y(0) = (y_0, y_1),$$

avec  $Y(t) = (y(t), y'(t)) \in \mathbb{R}^{2n}$  et  $B = \begin{pmatrix} 0 & A \\ I & 0 \end{pmatrix} \in \mathcal{M}_{2n}(\mathbb{R})$ .

## 5.2 La méthode des différences finies

Dans la pratique, on rencontre de nombreuses équations différentielles pour lesquelles l'expression de la solution n'est pas connue explicitement. Il est alors pertinent de rechercher des méthodes numériques permettant d'approcher la solution du problème de Cauchy

$$y'(t) = f(y(t), t), \quad t \in ]0, T[, \quad y(0) = y_0.$$

La méthode des différences finies est fondée sur l'idée que si  $y$  est une fonction dérivable, alors le quotient

$$\Delta_h y(t) := \frac{y(t+h) - y(t)}{h},$$

approche la valeur  $y'(t)$  lorsque  $h > 0$  est petit. Ce quotient est appelé *différence finie* de pas  $h$  de  $y$  au point  $t$ , et l'application linéaire  $\Delta_h : g \mapsto \Delta_h g$  est appelée opérateur de différence finie de pas  $h$ . On précise parfois qu'il s'agit de la *différence finie avant*, par distinction avec la *différence finie arrière* qui est donnée pour  $h > 0$  par

$$\Delta_{-h} y(t) := \frac{y(t) - y(t-h)}{h},$$

et qui est aussi une approximation de  $y'(t)$  lorsque  $h > 0$  est petit. En se fixant un  $h > 0$  appelé *pas de temps* et en posant  $t_k = kh$ , pour  $k = 0, 1, 2, \dots$ , on voit ainsi que l'on a

$$\Delta_h y(t_k) = \frac{y(t_{k+1}) - y(t_k)}{h} \approx y'(t_k) = f(y(t_k), t_k),$$

ou encore

$$y(t_{k+1}) \approx y(t_k) + hf(y(t_k), t_k).$$

Si cette formule était exacte, on pourrait calculer de proche en proche les valeurs  $y(t_k)$  puisque l'on connaît  $y(t_0) = y_0$ . Comme elle n'est qu'une approximation, on introduit une quantité  $y^k$  qui a vocation à approcher  $y(t_k)$  et que l'on calcule de proche en proche en posant  $y^0 := y_0$  et

$$y^{k+1} := y^k + hf(y^k, t_k).$$

C'est le *schéma d'Euler explicite* à un pas. Dans le cas de l'équation  $y' = Ay$ , ce schéma prend la forme simple

$$y^{k+1} = (I + hA)y^k.$$

Ce schéma permet ainsi de calculer des valeurs approchées de  $y$  sur la grille de points  $(t_k)_{k \geq 0}$ . On peut en déduire une approximation  $y_h$  de la fonction  $y$  en tout  $t \geq 0$  en interpolant ces valeurs, par exemple par une fonction affine par morceaux : on pose

$$y^h(t) = y_k + \frac{y_{k+1} - y_k}{h}(t - t_k), \quad \text{si } t \in [t_k, t_{k+1}].$$

Notons que si  $n > 1$  il s'agit de valeurs vectorielles  $y^k = (y_1^k, \dots, y_n^k)$  ce qui revient à dire que  $y^h$  est une fonction à valeur vectorielle donc chaque composante  $y_i^h$  est obtenue par interpolation des valeurs  $(y_i^k)$  aux points  $t_k$ .

Il existe d'autres schémas aux différences finies que le schéma d'Euler implicite. En particulier, en partant de la remarque que  $\frac{y(t_{k+1}) - y(t_k)}{h}$  est la différence arrière  $\Delta_{-h} y(t_{k+1})$  et peut donc aussi être vue comme une approximation de  $y'(t_{k+1})$ , c'est à dire

$$y(t_{k+1}) - hf(y(t_{k+1}), t_n) \approx y(t_k),$$

on obtient un schéma différent en posant  $y^0 := y_0$  et

$$y^{k+1} - hf(y^{k+1}, t_{k+1}) := y^k.$$

C'est le *schéma d'Euler implicite*, qui est appelé ainsi car la valeur  $y^{k+1}$  n'est pas donnée explicitement en fonction de  $y^k$  mais comme la solution d'une équation qu'il faut résoudre soit de manière exacte, soit

par une méthode d'approximation comme celles étudiées dans le chapitre 2. Dans le cas de l'équation  $y' = Ay$ , ce schéma prend la forme simple

$$(I - Ah)y^{k+1} = y^n,$$

et on voit que la solution est définie lorsque  $I - Ah$  est inversible ce qui est toujours vrai pour  $h$  suffisamment petit. On a alors

$$y^{n+1} = (I - hA)^{-1}y^k.$$

Un autre exemple de schéma part de la remarque que  $\frac{y(t_{k+1}) - y(t_k)}{h}$  est aussi une approximation de  $y'(t_{k+\frac{1}{2}})$ , en posant  $t_{k+\frac{1}{2}} = (k + \frac{1}{2})h = \frac{t_k + t_{k+1}}{2}$ . On a ainsi

$$y(t_{k+1}) \approx y(t_k) + hf(y(t_{k+\frac{1}{2}}), t_{k+\frac{1}{2}}) \approx y(t_k) + hf\left(y(t_k) + \frac{h}{2}f(y(t_k), t_k), t_{k+\frac{1}{2}}\right),$$

ce qui mène à un schéma explicite dit du *point milieu*

$$y^{k+1} := y^k + hf\left(y_k + \frac{h}{2}f(y_k, t_k), t_{k+\frac{1}{2}}\right).$$

Dans le cas de l'équation  $y' = Ay$ , ce schéma prend la forme simple

$$y^{k+1} = (I + Ah + \frac{1}{2}h^2A^2)y^k.$$

Nous avons présentés les trois schémas les plus simples, mais il existe de nombreux autres schémas aux différences finies.

### 5.3 Etude de convergence

L'étude de convergence d'un schémas aux différences finies consiste à étudier la proximité entre  $y^k$  et  $y(t_k)$  lorsque le pas de temps  $h$  tends vers 0. Comme on s'est placé sur un intervalle  $[0, T]$ , on ne considère que les valeurs de  $k$  telles que  $0 \leq t_k \leq T$ , c'est à dire  $0 \leq k \leq \frac{T}{h}$ .

**Définition 5.3.1** *Le schéma est dit convergent si on a*

$$\lim_{h \rightarrow 0} \max_{0 \leq k \leq \frac{T}{h}} \|y(t_k) - y^k\| = 0,$$

où  $\|\cdot\|$  désigne une norme sur  $\mathbb{R}^n$ .

Notons que le choix de la norme n'importe pas dans cette définition puisque toutes les normes sur  $\mathbb{R}^n$  sont équivalentes. On étudie pour commencer la convergence du schéma d'Euler explicite pour un problème de Cauchy

$$y'(t) = f(y(t), t), \quad t \in ]0, T[, \quad y(t_0) = y_0,$$

dont on suppose que la solution  $y(t)$  existe sur  $[0, T]$ . On introduit l'erreur du schéma

$$e^k = y(t_k) - y^k.$$

Remarquons que l'on a  $e^0 = 0$ .

Dans un premier temps, on s'intéresse à la précision de l'approximation de la dérivée  $g'(t)$  d'une fonction  $g : [0, T] \rightarrow \mathbb{R}$  par la différence finie  $\Delta_h g(t)$ . Si  $g \in \mathcal{C}^2$  la formule de Taylor-Lagrange permet d'écrire

$$\Delta_h g(t) = \frac{g(t) + hg'(t) + \frac{h^2}{2}g''(s) - g(t)}{h} = g'(t) + \frac{h}{2}g''(s),$$

avec  $s \in [t, t+h]$ . Par conséquent si  $[t, t+h] \subset [0, T]$  on a l'estimation

$$|g'(t) - \Delta_h g(t)| \leq \frac{h}{2}|g''(s)| \leq \frac{M_2}{2}h,$$

avec  $M_2 = \|g''\| = \max_{s \in [0, T]} |g''(s)|$ , qui montre que la précision de l'approximation de  $g'(t)$  par  $\Delta_h g(t)$  est d'ordre  $h$ .

Si on applique cette remarque à chacune des  $n$  composantes  $y_i$  de la solution  $y$  en supposant que celle-ci est de classe  $\mathcal{C}^2$  (c'est à dire que toutes les composantes le sont), on obtient

$$\|y'(t) - \Delta_h y(t)\|_\infty \leq \frac{M_2}{2} h,$$

avec  $M_2 := \max_{i=1, \dots, n} \|y_i''\|$ . Notons que la régularité  $\mathcal{C}^2$  de la solution  $y(t)$  est assurée dans le cas des équations linéaires autonomes  $y' = Ax + b$ , au vu du Théorème 5.1.1.

On introduit à présent une quantité qui mesure de combien la solution exacte  $y(t_k)$  ne vérifie pas l'équation définissant le schéma explicite, en posant

$$c_k := \frac{y(t_{k+1}) - y(t_k)}{h} - f(y(t_k), t_k).$$

Cette quantité est appelée *erreur de consistance du schéma*. D'après ce qu'on vient de voir, on peut écrire

$$\|c_k\|_\infty = \|\Delta_h y(t_k) - y'(t_k)\|_\infty \leq \frac{M_2}{2} h.$$

Le schéma est dit consistant à l'ordre 1 car  $h$  apparait à la puissance 1 à droite de cette inégalité. En faisant la soustraction entre la définition de l'erreur de consistance et l'identité

$$0 = \frac{y^{k+1} - y^k}{h} - f(y^k, t_k),$$

qui définit le schéma, on obtient une relation entre  $e^k$  et  $e^{k+1}$  :

$$e^{k+1} = e^k + h(f(y(t_k), t_k) - f(y^k, t_k)) + hc_k.$$

Afin d'aller plus loin, on fait une hypothèse sur  $f$  donnée par la définition suivante

**Définition 5.3.2** *La fonction  $f$  est Lipschitzienne en  $x$  uniformément en  $t$  si et seulement si il existe une constante  $L$  telle que pour tout  $t \in [0, T]$  et  $x, \tilde{x} \in \mathbb{R}^n$  on a*

$$\|f(x, t) - f(\tilde{x}, t)\| \leq L\|x - \tilde{x}\|,$$

où  $\|\cdot\|$  désigne une norme sur  $\mathbb{R}^n$ .

Notons que le choix de la norme n'importe pas dans cette définition, à une modification près de la constante  $L$  puisque toutes les normes sur  $\mathbb{R}^n$  sont équivalentes. Notons aussi que cette hypothèse est toujours vérifiée dans le cas des équations différentielles linéaires autonomes  $y' = Ax + b$  avec la constante  $L = \|A\|$ . Sous l'hypothèse que  $f$  est Lipschitzienne en  $x$  uniformément en  $t$ , et en utilisant la norme  $\|\cdot\|_\infty$ , la relation entre  $e^k$  et  $e^{k+1}$  entraîne

$$\|e^{k+1}\|_\infty \leq (1 + hL)\|e^k\|_\infty + h\|c_k\|_\infty$$

et par conséquent d'après l'estimation établie pour  $\|c_k\|_\infty$

$$\|e^{k+1}\|_\infty \leq (1 + hL)\|e^k\|_\infty + \frac{M_2}{2} h^2.$$

Par itération (ou par récurrence) et en utilisant  $e^0 = 0$ , on obtient ainsi

$$\begin{aligned} \|e^k\|_\infty &\leq (1 + hL)\|e^{k-1}\|_\infty + hL \frac{M_2}{2} h^2 \\ &\leq (1 + hL)^2 \|e^{k-2}\|_\infty + (1 + (1 + hL)) \frac{M_2}{2} h^2 \\ &\leq \dots \leq (1 + (1 + hL) + \dots + (1 + hL)^{k-1}) \frac{M_2}{2} h^2 \\ &\leq \frac{(1+hL)^k - 1}{L} \frac{M_2}{2} h. \end{aligned}$$

En utilisant le fait que  $hk \leq T$ , et l'inégalité  $1 + \alpha \leq e^\alpha$  pour  $\alpha > 0$ , on obtient ainsi pour tout  $k$  tel que  $t_k \in [0, T]$  l'estimation

$$\|e_k\| \leq Ch, \quad C := \frac{(e^{LT} - 1)M_2}{2L}.$$

On a ainsi démontré le résultat suivant.

**Théorème 5.3.1** *Si la solution  $y$  de l'équation différentielle existe et est de classe  $\mathcal{C}^2$  sur  $[0, T]$  et si  $f$  est Lipschitzienne en  $x$  uniformément en  $t$ , alors le schéma explicite est convergent et vérifie de plus l'estimation d'erreur*

$$\max_{0 \leq k \leq \frac{T}{h}} \|y(t_k) - y^k\|_\infty \leq Ch,$$

où la constante  $C$  dépend de  $T$ , de la constante de Lipschitz  $L$  et des normes sup des dérivées secondes des composantes de  $y$  sur  $[0, T]$ . On dit que ce schéma est d'ordre 1 (ou converge à l'ordre 1).

**Remarque 5.3.1** *L'hypothèse que  $f$  est Lipschitzienne en  $x$  uniformément en  $t$  joue un rôle central dans la théorie qui établit l'existence et l'unicité de la solution du problème de Cauchy, et que nous ne présentons pas ici.*

Il est possible de mettre en oeuvre une analyse similaire pour le schéma implicite, en partant de la remarque que la différence finie arrière  $\Delta_{-h}y(t)$  est aussi une approximation de  $y'(t)$  dont l'erreur en norme  $\ell^\infty$  est contrôlée par  $\frac{M_2}{2}h$ . L'erreur de consistance, qui est ici définie par

$$c_k := \frac{y(t_{k+1}) - y(t_k)}{h} - f(y(t_{k+1}), t_{k+1}),$$

est à nouveau majorée par

$$\|c_k\|_\infty \leq \frac{M_2}{2}h.$$

Un étude plus approfondie permet de prouver que si  $f$  est Lipschitzienne en  $x$  uniformément en  $t$ , alors le schéma implicite est bien défini pour tout  $k$  lorsque  $h > 0$  est suffisamment petit, et que l'on aboutit à une estimation d'erreur du premier ordre similaire à celle du théorème 5.3.1.

L'analyse de convergence pour le schéma explicite du point milieu fait appel à l'étude de la *différence finie centrée*

$$\delta_h g(t) := \frac{g(t + \frac{h}{2}) - g(t - \frac{h}{2})}{h},$$

qui est aussi une approximation de  $g'(t)$ . Si on utilise la formule de Taylor-Lagrange au point  $t$  à l'ordre 3, en supposant que  $g$  est de classe  $\mathcal{C}^3$ , on constate que les termes d'ordre deux s'annulent par symétrie et que l'on a ainsi

$$\delta_h g(t) = g'(t) + \frac{\frac{1}{6}(\frac{h}{2})^3(g^{(3)}(s) + g^{(3)}(u))}{h},$$

avec  $s \in [t, t + h/2]$  et  $u \in [t - h/2, t]$ . On aboutit ainsi à l'estimation

$$|g'(t) - \delta_h g(t)| \leq \frac{M_3}{24}h^2, \quad M_3 := \|g^{(3)}\|$$

qui montre que la différence finie centrée est une approximation plus précise de la dérivée que les différences finies avant et arrière. On obtient de même pour la fonction vectorielle  $y$  solution de l'équation

$$\|y'(t) - \delta_h y(t)\|_\infty \leq \frac{M_3}{24}h^2,$$

avec  $M_3 := \max_{i=1}^n \|y_i^{(3)}\|$ . L'erreur de consistance est ici définie par

$$c_k := \frac{y(t_{k+1}) - y(t_k)}{h} - f\left(y(t_k) + \frac{h}{2}f(y(t_k), t_k), t_{k+\frac{1}{2}}\right).$$

En faisant l'hypothèse que  $y$  est de classe  $\mathcal{C}^3$  et que  $f$  est lipschitzienne en  $x$  uniformément en  $t$ , on peut la majorer en écrivant

$$\begin{aligned} \|c_k\|_\infty &\leq \left\| \frac{y(t_{k+1}) - y(t_k)}{h} - y'(t_{k+\frac{1}{2}}) \right\|_\infty + \left\| y'(t_{k+\frac{1}{2}}) - f\left(y(t_k) + \frac{h}{2}f(y(t_k), t_k), t_{k+\frac{1}{2}}\right) \right\|_\infty \\ &= \left\| \delta_h y(t_{k+\frac{1}{2}}) - y'(t_{k+\frac{1}{2}}) \right\|_\infty + \left\| f(y(t_{k+\frac{1}{2}}), t_{k+\frac{1}{2}}) - f\left(y(t_k) + \frac{h}{2}f(y(t_k), t_k), t_{k+\frac{1}{2}}\right) \right\|_\infty \\ &\leq \frac{M_3}{24}h^2 + L \left\| y(t_{k+\frac{1}{2}}) - y(t_k) - \frac{h}{2}f(y(t_k), t_k) \right\|_\infty \\ &= \frac{M_3}{24}h^2 + L \frac{h}{2} \left\| \Delta_{\frac{h}{2}} y(t_k) - y'(t_k) \right\|_\infty \\ &\leq \left( \frac{M_3}{24} + L \frac{M_2}{8} \right) h^2. \end{aligned}$$

En faisant la soustraction entre la définition de l'erreur de consistance et l'identité

$$0 = \frac{y^{k+1} - y^k}{h} - f\left(y_k + \frac{h}{2}f(y_k, t_k), t_{k+\frac{1}{2}}\right),$$

qui définit le schéma, on établit la relation entre  $e^k$  et  $e^{k+1}$

$$e^{k+1} = e^k + h\left(f\left(y(t_k) + \frac{h}{2}f(y(t_k), t_k), t_{k+\frac{1}{2}}\right) - f\left(y_k + \frac{h}{2}f(y_k, t_k), t_{k+\frac{1}{2}}\right)\right) + hc_k.$$

En utilisant l'hypothèse que  $f$  est lipschitzienne en  $x$  uniformément en  $t$ , on obtient

$$\begin{aligned} \|e^{k+1}\|_\infty &\leq \|e^k\|_\infty + hL\|y(t_k) + \frac{h}{2}f(y(t_k), t_k) - y_k - \frac{h}{2}f(y_k, t_k)\|_\infty + h\|c_k\|_\infty \\ &\leq (1 + hL)\|e^k\|_\infty + hL\|\frac{h}{2}f(y(t_k), t_k) - \frac{h}{2}f(y_k, t_k)\|_\infty + h\|c_k\|_\infty \\ &\leq (1 + hL + \frac{1}{2}(hL)^2)\|e^k\|_\infty + \left(\frac{M_3}{24} + L\frac{M_2}{8}\right)h^3. \end{aligned}$$

En terminant les calcul comme pour le schéma d'Euler explicite, on obtient finalement le résultat suivant

**Théorème 5.3.2** *Si la solution  $y$  de l'équation différentielle existe et est de classe  $C^3$  sur  $[0, T]$  et si  $f$  est Lipschitzienne en  $x$  uniformément en  $t$ , le schéma du point milieu est convergent, et vérifie de plus l'estimation d'erreur*

$$\max_{0 \leq k \leq \frac{T}{h}} \|y(t_k) - y^k\|_\infty \leq Ch^2.$$

où la constante  $C$  dépend de  $T$ , de la constante de Lipschitz  $L$  et des normes sup des dérivées secondes et troisièmes des composantes de  $y$  sur  $[0, T]$ . On dit que ce schéma est d'ordre 2 (ou converge à l'ordre 2).

Pour terminer, notons qu'il est possible d'itérer la formule des différences finies lorsque l'on cherche à approcher les dérivées d'ordre supérieur d'une fonction  $g$  à partir de ses valeurs en des points. Par exemple, si on cherche à approcher  $g''(t)$  on peut ainsi écrire

$$g''(t) \approx \Delta_h g'(t) \approx \Delta_h \Delta_h g(t) = \frac{\Delta_h g(t+h) - \Delta_h g(t)}{h} = \frac{g(t+2h) - 2g(t+h) + g(t)}{h^2}.$$

Il s'agit de la différence finie d'ordre deux avant de pas  $h$  au point  $t$  qui est notée

$$\Delta_h^2 g(t) := \frac{g(t+2h) - 2g(t+h) + g(t)}{h^2}.$$

Une autre approximation de  $g''(t)$  est la différence finie d'ordre deux centrée

$$\Delta_h \Delta_{-h} g(t) = \delta_h^2 g(t) = \frac{g(t+h) - 2g(t) + g(t-h)}{h^2}.$$

On pourra vérifier, en utilisant la formule de Taylor-Lagrange que la différence centrée approche  $g''(t)$  à l'ordre  $h^2$  lorsque  $g \in C^4$ , avec l'estimation

$$\|\delta_h^2 g(t) - g''(t)\| \leq \frac{M_4}{12} h^2, \quad M_4 := \|g^{(4)}\|$$

alors que l'approximation par  $\Delta_h^2 g(t)$  est d'ordre  $h$ . Les différences finies itérées se relient naturellement aux différences divisées rencontrées dans la section §3.3 : on vérifie aisément par récurrence que si  $t_n = nh$  on a alors

$$\Delta_h^m g(t_n) = m!g[t_n, t_{n+1}, \dots, t_{n+m}].$$