Approximation numérique des fonctions

2017-2018

Notes du Cours 3M234

Martin Campos Pinto, d'après le cours d'Albert Cohen et Bruno Després

17 janvier 2018

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, où 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 leurs performances. 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.

Ces notes contiennent la totalité des résultats exposés en Amphi mais la réciproque est fausse : les cours d'Amphi mettent l'accent sur les résultats les plus importants en laissant de côté certains aspects plus techniques. Il est de ce fait vivement conseillé d'assister aux Amphis plutôt que de se baser uniquement sur la lecture de ces notes.

Avertissement : comme tout document écrit, ce cours contient très certainement des imprécisions qui ont échappé au regard vigilant de ses auteurs. Toutes les remarques ou questions permettant d'en améliorer la rédaction pourront être envoyées par mail à l'adresse campos@ljll.math.upmc.fr

Quelques références :
- *Analyse numérique : une approche mathématique*, Schatzman, éd. Dunod
- *Analyse numérique et équations différentielles*, Demailly, éd. EDP Sciences
- *Introduction à l'analyse numérique matricielle et à l'optimisation*, Ciarlet, éd. Dunod
- *Analyse numérique des équations différentielles*, Mignot-Crouzeix, éd. Masson

Références en anglais :
- *Numerical analysis : a mathematical introduction*, Schatzman, Oxford University Press
- *Interpolation and Approximation*, Davis, Dover Publications

Table des matières

1 Introduction : approximation et convergence

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 $f(x)$ pour $x \in I$. 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 de paramètres que l'on peut ainsi mettre en mémoire. Un exemple consiste à choisir g dans l'ensemble des polynômes de degré n : on peut alors caractériser g par ses $n+1$ coefficients. Plus généralement, on peut chercher à approcher f par une fonction g appartenant à un espace de fonctions E_N de dimension N. Intuitivement, on approche de mieux en mieux f lorsque la quantité 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.

1.1 Cadre théorique et questions générales

Les méthodes d'approximations que nous étudierons dans ce cours peuvent se décomposer suivant le schéma général suivant :

(i) choix d'un espace fonctionnel de dimension finie N, noté E_N, dans lequel seront définies les approximations de la fonction cible f ;

(ii) choix d'un approximant particulier f_N dans l'espace E_N ;

(iii) identification d'un espace fonctionnel X contenant E_N, et pour lequel l'opérateur d'approximation

$$\mathcal{A}_N : X \to E_N, \quad f \mapsto f_N$$

est bien défini, et stable.

On rappelle qu'un espace fonctionnel est un espace vectoriel normé dont les éléments sont des fonctions, et on se restreindra ici à des méthodes *linéaires*, au sens où $\mathcal{A}_N$ sera toujours un opérateur linéaire (de nombreuses méthodes non-linéaires sont utilisées en pratique, mais leur étude dépasse le cadre de ce cours).

Par *stable*, on entend que $\mathcal{A}_N f$ doit dépendre continuement de f, au sens des topologies de X et E_N. Comme E_N est de dimension finie toutes les normes sont équivalentes et on peut choisir de lui associer la norme de X. Comme $\mathcal{A}_N$ est linéaire, sa stabilité revient à l'existence d'une constante C_N telle que

$$\|\mathcal{A}_N f\|_X \leq C_N \|f\|_X, \qquad \forall f \in X. \tag{1.1}$$

Ici la constante peut a priori varier avec N, mais elle ne doit pas dépendre de f (voir la remarque 1.1.2 sur l'emploi du terme de "constante"). La stabilité est une propriété cruciale que devra vérifier toute méthode raisonnable, car un opérateur instable peut donner des résultats arbitrairement éloignés lorsqu'on l'applique à des fonctions arbitrairement proches (ce qui n'est évidemment pas acceptable en pratique). En effet, s'il n'est pas possible de trouver un nombre C_N tel que (1.1) est vrai, alors on vérifiera que pour toute

fonction $f \in X$ et tout réel $\varepsilon > 0$ arbitrairement petit, il existe une fonction $\tilde{f}_\varepsilon$ telle que

$$\|f - \tilde{f}_\varepsilon\|_X \leq \varepsilon \quad \text{et} \quad \|\mathcal{A}_N f - \mathcal{A}_N \tilde{f}_\varepsilon\|_X \geq \frac{1}{\varepsilon}.$$

Un exemple important qui sera considéré à plusieurs reprises dans ce cours est celui où l'espace d'approximation est constitué des polynômes de degré inférieur ou égal à n,

$$\mathbb{P}_n := \text{Vect}\{x \mapsto x^k \ ; \ k = 0, 1, \cdots, n\} \tag{1.2}$$

dont la dimension est $N = \dim(\mathbb{P}_n) = n + 1$. L'espace X sera le plus souvent celui des fonctions continues sur un intervalle I, noté $\mathcal{C}^0(I)$, qu'on munira de sa norme usuelle, dite L^∞ ou *norme sup* sur l'intervalle I,

$$\|g\|_{L^\infty(I)} := \sup_{x \in I} |g(x)|. \tag{1.3}$$

Cette norme est aussi appelée *norme de la convergence uniforme*, car une suite de fonctions f_n converge uniformément vers f sur I si et seulement si $\lim_{n\to\infty} \|f - f_n\|_{L^\infty(I)} = 0$. On rappelle que l'espace $\mathcal{C}^0(I)$ des fonctions continues sur I est complet pour cette norme.

L'étude des propriétés d'une méthode d'approximation passe évidemment par celle des *erreurs d'approximation*

$$\|f - \mathcal{A}_N f\|_X \tag{1.4}$$

et de leur comportement lorsque N tend vers l'infini. Ces erreurs peuvent être étudiées sur ordinateur, on parle alors d'étude *numérique*. L'un des objectifs de l'analyse mathématique est d'établir des estimations *a priori* (i.e., qui ne requièrent pas de calculer f_N) sur leur amplitude. On dit que la méthode est convergente lorsque ces erreurs tendent vers 0, et lorsque c'est le cas on peut vouloir déterminer sa *vitesse de convergence*. Plus précisément, on pourra utiliser la définition suivante.

Définition 1.1.1 (convergence et ordre de convergence) *On dit que les approximations $\mathcal{A}_N$* **convergent** *sur une partie $Y \subset X$ si pour toute fonction $f \in Y$, on a*

$$\lim_{N\to\infty} \|f - \mathcal{A}_N f\|_X = 0.$$

S'il existe un réel $\gamma > 0$ tel que pour tout $f \in Y$, il existe $C(f) > 0$ tel que

$$\|f - \mathcal{A}_N f\|_X \leq C(f) \frac{1}{N^\gamma}, \qquad N \in \mathbb{N}^*,$$

alors on dit que les approximations **convergent à l'ordre γ.**

Au-delà de l'opérateur $\mathcal{A}_N$, une question importante est celle des propriétés d'approximation de l'espace E_N. On appelle *erreur de meilleure approximation* la quantité

$$\inf_{g \in E_N} \|f - g\|_X \tag{1.5}$$

qui n'est rien d'autre que la distance entre f et l'espace E_N vu comme un sous-espace de X, et il est naturel de chercher à établir des estimations théoriques pour cette quantité. Il est naturel que cette distance dépende de la fonction f. En pratique on verra qu'elle est souvent liée à la *régularité* de f, c'est-à-dire à l'amplitude de ses dérivées. On peut préciser cette assertion en énonçant un résultat important (et non-trivial) qui sera établi dans la section 4.2.2 de ce cours, et qui concerne l'erreur de meilleure approximation polynomiale. Il fait intervenir la norme suivante qui généralise la norme uniforme (1.3) aux fonctions de régularité supérieure,

$$\|g\|_{\mathcal{C}^m(I)} := \max_{k=0,\dots,m} \|g^{(k)}\|_{L^\infty(I)}. \tag{1.6}$$

On pourra vérifier que l'espace des fonctions de classe $\mathcal{C}^m$ sur I est complet pour cette norme.

Théorème 1.1.1 (meilleure approximation polynomiale) *Si f est une fonction de classe $\mathcal{C}^m$ sur un intervalle I, son erreur de meilleure approximation polynomiale vérifie*

$$\inf_{g\in\mathbb{P}_n} \|f-g\|_{L^\infty(I)} \leq C_m \frac{\|f\|_{\mathcal{C}^m(I)}}{n^m} \qquad \forall n \geq 1$$

avec une constante C_m qui dépend de m et de la longueur de l'intervalle I, mais est indépendante de n et de f.

Il est naturel d'évaluer la précision des opérateurs $\mathcal{A}_N$ à l'aune des erreurs de meilleure approximation (1.5). Dans le meilleur des cas les fonctions $\mathcal{A}_N f$ réalisent les erreurs de meilleure approximation,

$$\|f-\mathcal{A}_N f\|_X = \inf_{g\in E_N} \|f-g\|_X, \qquad \forall f\in X, \quad N\in\mathbb{N},$$

et on dit que la méthode d'approximation est *optimale*. Dans le cas plus fréquent (mais toujours favorable) où il existe une constante C indépendante de N (et de f) telle que

$$\|f-\mathcal{A}_N f\|_X \leq C \inf_{g\in E_N} \|f-g\|_X, \qquad \forall f\in X, \quad N\in\mathbb{N}, \tag{1.7}$$

on dira que la méthode d'approximation est *quasi-optimale*. Le résultat suivant, assez simple à vérifier, illustre l'intérêt de ce cadre théorique.

Proposition 1.1.1 *Si les opérateurs $\mathcal{A}_N$ sont* uniformément stables *au sens où*

$$\|\mathcal{A}_N f\|_X \leq C_{\mathcal{A}} \|f\|_X, \qquad \forall f\in X, \quad N\in\mathbb{N} \tag{1.8}$$

pour une constante $C_{\mathcal{A}}$ indépendante de N *(et de f), et s'ils préservent les espaces E_N au sens où*

$$\mathcal{A}_N g = g \qquad \forall g\in E_N, \tag{1.9}$$

alors ils sont quasi-optimaux, en effet on a

$$\|f-\mathcal{A}_N f\|_X \leq (1+C_{\mathcal{A}}) \inf_{g\in E_N} \|f-g\|_X$$

avec la même constante $C_{\mathcal{A}}$ que dans (1.8).

Preuve. Pour deux fonctions quelconques $f \in X$ et $g \in E_N$, on a

$$\|f - \mathcal{A}_N f\|_X \leq \|f - g\|_X + \|\mathcal{A}_N g - \mathcal{A}_N f\|_X \leq \|f - g\|_X + C_{\mathcal{A}} \|g - f\|_X$$

où l'on a utilisé (1.9) dans la première inégalité et (1.8) (associé à la linéarité de $\mathcal{A}_N$) dans la deuxième. Le résultat s'en déduit en prenant une borne inférieure sur $g \in E_N$. □

Remarque 1.1.1 *Si $\mathcal{A}_N$ est stable, sa norme d'opérateur (subordonnée à celle de X) est définie comme la plus petite constante C_N vérifiant l'inégalité* (1.1)*, c'est-à-dire*

$$|||\mathcal{A}_N|||_X := \sup_{\substack{f \in X \\ f \neq 0}} \frac{\|\mathcal{A}_N f\|_X}{\|f\|_X}. \tag{1.10}$$

Les opérateurs $\mathcal{A}_N$ sont donc uniformément stables au sens de (1.8) *si et seulement si la suite $(|||\mathcal{A}_N|||_X)_{N \in \mathbb{N}}$ est bornée.*

Remarque 1.1.2 *Malgré l'utilisation de la même notation, les nombres C apparaissant dans les inégalités* (1.1) *et* (1.7) *sont a priori sans rapport. Il en sera de même dans la suite du cours : la lettre C sera utilisée de façon récurrente dans les inégalités pour désigner des réels (généralement positifs) indépendants de certaines variables apparaissant dans l'inégalité, sans qu'il existe a priori de lien entre eux. On tentera de préciser la dépendance de ces "constantes" par rapport à d'autres variables : ainsi dans l'inégalité* (1.1) *les nombres C_N ne dépendent pas de f, mais ils peuvent varier avec N.*

Pour évaluer les performances pratiques des méthodes d'approximation qui seront vues dans le cours, nous les appliquerons aux trois fonctions suivantes qui ont l'intérêt de présenter divers types de régularité. Ces fonctions seront considérées sur l'intervalle $[-1, 1]$ de façon à faciliter la comparaison des résultats :

(i) la fonction sinus

$$f_{\sin} : x \mapsto \sin(2\pi x) \tag{1.11}$$

qui est extrêmement régulière (elle est de classe $\mathcal{C}^\infty$ et développable en série entière sur $\mathbb{R}$),

(ii) une fonction régulière introduite par le mathématicien Runge pour étudier la stabilité des interpolations polynomiales,

$$f_{\text{Runge}} : x \mapsto \frac{1}{1 + 25x^2} \tag{1.12}$$

qui est également $\mathcal{C}^\infty$ mais moins régulière que le sinus au sens où l'amplitude de ses dérivées augmente rapidement (voir par exemple la discussion section 2.1),

(iii) enfin la fonction valeur absolue

$$f_{\text{abs}} : x \mapsto |x| \tag{1.13}$$

qui est continue mais pas $\mathcal{C}^1$.

1.2 Un exemple important : l'interpolation polynomiale

Pour se familiariser avec quelques questions importantes qui seront vues dans ce cours, on considère le cas de l'interpolation polynomiale sur un intervalle $I = [a, b]$. Cette méthode classique consiste à approcher une fonction f par un polynôme p_n prenant les mêmes valeurs que f en certains points choisis à l'avance dans l'intervalle. Si l'on choisit p_n dans $\mathbb{P}_n$ l'espace des polynômes de degré inférieur ou égal à n, cf. (1.2), la dimension de l'espace d'approximation est $N = n + 1$ et nous avons besoin de $n + 1$ points dans I,

$$a \leq x_0 < x_1 < \cdots < x_n \leq b. \tag{1.14}$$

Le polynôme interpolant est alors défini de façon unique : c'est évident pour $n = 0$ ou 1, et on peut prouver que c'est vrai pour tout $n > 1$: cette étude sera menée dans la section 2.2, où nous établirons que l'opérateur d'interpolation est stable pour la norme L^∞ (proposition 2.2.1). La question de la stabilité uniforme est plus délicate, et dépend fortement de la *position* des noeuds d'interpolation (1.14). Son étude sera menée dans le chapitre 4, où nous établirons des estimations d'erreur a priori en utilisant le cadre théorique présenté plus haut, et notamment le théorème 1.1.1.

Dans le cadre de cette introduction nous nous proposons de mener une discussion informelle sur les performances numériques des interpolations lorsqu'on les applique aux fonctions tests $f_{\sin}$ et f_{abs} définies en (1.11) et (1.13). Nous nous plaçons donc sur l'intervalle $I = [-1, 1]$, et nous commençons par considérer le cas le plus simple, à savoir une répartition régulière des noeuds d'interpolation, $x_i = -1 + \frac{2i}{n}$ pour $i = 0, \ldots, n$ (avec $n \geq 1$ pour simplifier).

Les polynômes interpolants p_n obtenus de cette façon sont représentés sur la figure 1 pour divers degrés, et les premières observations que nous pouvons faire sont les suivantes :

- avec la fonction sinus (colonne de gauche), des oscillations apparaissent à proximité des noeuds extrêmes pour $n = 5$. Ces oscillations détériorent la précision des approximations mais celle-ci s'améliore ensuite nettement pour $n \geq 9$, et pour $n = 15$ les deux courbes sont pratiquement confondues ;
- avec la fonction valeur absolue (colonne de droite), on voit également des oscillations mais leur comportement est différent lorsque n augmente : leur apparition est plus tardive, et leur amplitude ne semble pas se résorber.

Pour confirmer ces observations nous traçons dans la figure 2 l'évolution des erreurs d'approximation (1.4) lorsque n varie entre 1 et 20. Ces erreurs sont ici définies en norme uniforme (1.3), et en pratique elles sont évaluées de façon approchée sur une grille fine de l'intervalle I. On vérifie ainsi que le comportement des interpolations est radicalement différent suivant le type de fonction f qui est approchée : elles convergent pour la fonction sinus, et elles divergent pour la fonction valeur absolue.

Ces résultats sont en accord avec le principe général évoqué plus haut : plus f est régulière, et plus ses approximations auront tendance à être précises. Toutefois, nous pouvons faire au moins deux commentaires sur la mauvaise qualité des interpolations de la fonction valeur absolue.

D'une part, on constate que les erreurs les plus importantes sont situées à proximité des extrémités de l'intervalle I, ce qui semble en contradiction avec l'idée suivant laquelle les difficultés de l'approximation sont liées au manque de régularité de la fonction f :

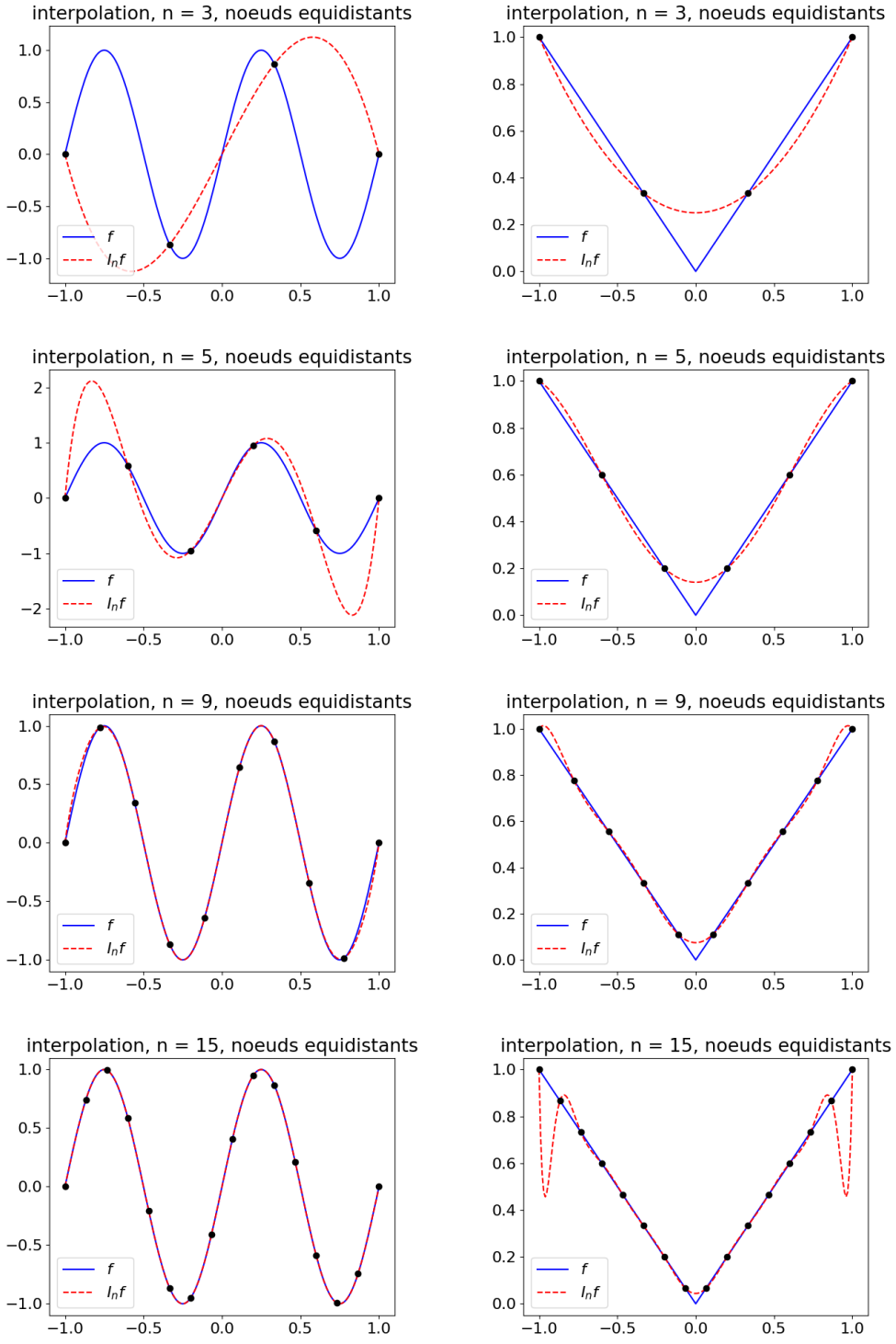

FIGURE 1 – Approximation des fonctions $f = f_{\sin}$ (à gauche) et $f = f_{\text{abs}}$ (à droite) par leurs polynômes interpolateurs p_n de degrés $n = 3, 5, 9$ et 15. L'interpolation utilise ici $n + 1$ noeuds équidistants sur l'intervalle $[-1, 1]$.

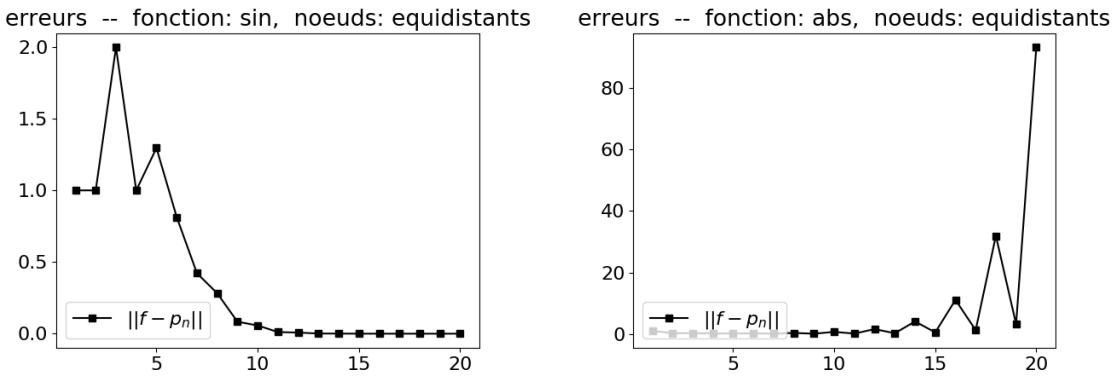

FIGURE 2 – Tracé des erreurs $\|f - p_n\|_{L^\infty(I)}$ en fonction de n, pour les interpolations sur des noeuds équidistants représentées dans la figure 1.

dans la mesure où f est infiniment régulière (et mieux : affine) en dehors du point $x = 0$, on pourrait s'attendre à ce que les erreurs d'approximation soient plutôt concentrées au voisinage de ce point.

D'autre part, il est intéressant de comparer ces résultats avec les estimations disponibles pour les erreurs de meilleure approximation : la fonction valeur absolue n'étant que $\mathcal{C}^0$ le théorème 1.1.1 ne permet pas de conclure à la convergence des approximations polynomiales, mais nous connaissons le célèbre théorème de Weierstrass suivant lequel toute fonction continue sur un intervalle borné est limite uniforme d'une suite de polynômes.

Théorème 1.2.1 (de Weierstrass) *Si f est continue sur I, alors*

$$\inf_{g \in \mathbb{P}_n} \|f - g\|_{L^\infty(I)} \to 0 \quad \textit{lorsque} \quad n \to \infty.$$

Il existe donc une suite $(g_n)_{n\in\mathbb{N}}$ de polynômes $g_n \in \mathbb{P}_n$ qui converge vers la fonction $f(x) = |x|$, et le fait que les approximations représentées sur la figure 1 ne convergent pas signifie essentiellement que le polynôme interpolant p_n représente un (très) mauvais choix parmi tous les polynômes possibles de degré n. Dans les termes de la section 1.1, cela peut s'exprimer en disant que l'approximation polynomiale par interpolation sur des noeuds équidistants est (très) loin d'être optimale. Pour faire mieux, il faut changer l'approximation : on peut par exemple définir $p_n \in \mathbb{P}_n$ par un principe d'approximation non-interpolante (ce qui sera vu plus bas), ou bien plus simplement changer les noeuds d'interpolation.

Il est en effet connu que les noeuds équidistants ont de mauvaises propriétés d'interpolation (on verra au chapitre 2.2 qu'ils donnent également de mauvais résultats pour certaines fonctions $\mathcal{C}^\infty$ comme la fonction de Runge), et qu'on obtient de bien meilleures approximations avec les noeuds de Tchebychev dont les valeurs sur l'intervalle $[-1, 1]$ sont

$$x_i = \cos\left(\frac{\pi(i + \frac{1}{2})}{n+1}\right), \qquad i = 0, 1, \cdots, n.$$

Ces noeuds correspondent à la projection sur l'axe x d'une subdivision régulière du demi-cercle, de sorte qu'ils seront plus concentrés près des extrémités de l'intervalle I qu'en son centre : intuitivement on peut penser que cette répartition permettra de réduire les oscillations observées près des bords.

Les résultats représentés sur la figure 3 montrent que c'est effectivement le cas. Pour la fonction sinusoidale, on n'observe pas de différence majeure : la convergence des approximations est toujours observée. Pour la fonction valeur absolue en revanche, les résultats sont nettement améliorés : les grandes oscillations ont disparu et les approximations semblent converger vers la fonction f lorsque n augmente. On observe d'ailleurs que les erreurs sont maintenant localisées au voisinage du point $x = 0$, ce qui semble plus en accord avec le principe (exprimé dans le théorème 1.1.1) suivant lequel la qualité des approximations devrait être d'abord limitée par la régularité de la fonction f.

A nouveau, on peut confirmer ces observations en traçant dans la figure 4 les courbes des erreurs $\|f - p_n\|_{L^\infty(I)}$ correspondant à l'interpolation sur les noeuds de Tchebychev. On vérifie ainsi que les approximations convergent pour les deux fonctions considérées, et que la convergence est plus lente pour la fonction valeur absolue, moins régulière que la fonction sinus.

Pour mieux apprécier les vitesses de convergence vers 0, on a retracé dans la figure 5 les courbes d'erreurs des figures 2 et 4 en utilisant une échelle logarithmique. La superposition avec des courbes dont la décroissance est prescrite (de la forme $x \mapsto x^{-\gamma}$, avec $\gamma > 0$) permet de préciser ces vitesses de convergence : en échelle logarithmique ces courbes deviennent en effet des droites, dont la pente peut facilement se comparer avec la pente moyenne tracée par les erreurs d'approximation : si la courbe (logarithmique) des erreurs décroît avec une pente inférieure où égale à $-\gamma$ on peut en déduire une convergence d'ordre γ, cf. la définition 1.1.1.

Pour la fonction sinusoidale on vérifie ainsi que la vitesse de convergence s'accélère à mesure que n augmente (que les noeuds soient équidistants ou de Tchebychev), et pour la fonction valeur absolue, on observe que lorsque les approximations convergent (i.e. avec les noeuds de Tchebychev), l'ordre de convergence est proche de 1. On remarquera à ce sujet que cette vitesse n'était pas prédite par le théorème 1.1.1 (ce qui ne met pas le théorème en défaut).

1.3 Résumé et plan du cours

Nous pouvons donner le résumé suivant des notions entrevues dans cette introduction.

- Une **méthode d'approximation** peut être vue comme un **opérateur** (linéaire) qui associe à toute fonction f d'un espace fonctionnel X de **dimension infinie** une fonction f_N choisie dans un espace E_N de **dimension finie** N.
- L'interpolation polynomiale est un exemple important : étant donné $n+1$ **points** dans un intervalle I, elle consiste à approcher $f \in \mathcal{C}^0(I)$ par **un unique polynôme de degré** $\leq n$ qui interpole f en ces points.
- Plus la fonction est **régulière** (au sens où ses dérivées sont d'amplitude réduite), plus ses approximations auront tendance à être **précises**.
- La précision des interpolations dépend également de la **position des noeuds** : avec des noeuds **équidistants** les interpolations semblent très **sous-optimales**

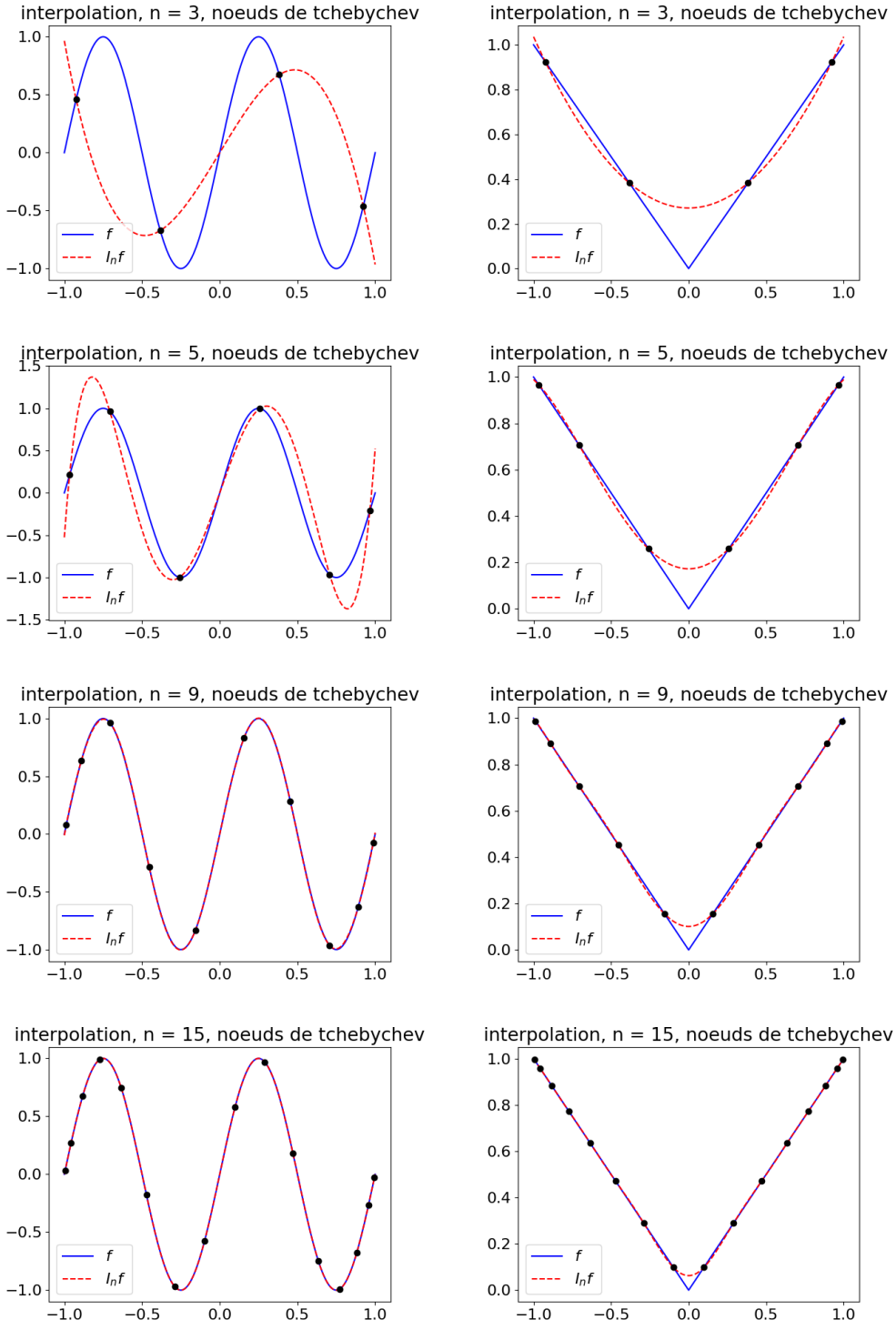

FIGURE 3 – Approximation des fonctions $f(x) = f_{\sin}$ (à gauche) et $f(x) = f_{\text{abs}}$ (à droite) par leurs polynômes interpolateurs p_n de degrés $n = 3, 5, 9$ et 15. L'interpolation est ici définie aux $n + 1$ noeuds de Tchebychev sur l'intervalle $[-1, 1]$.

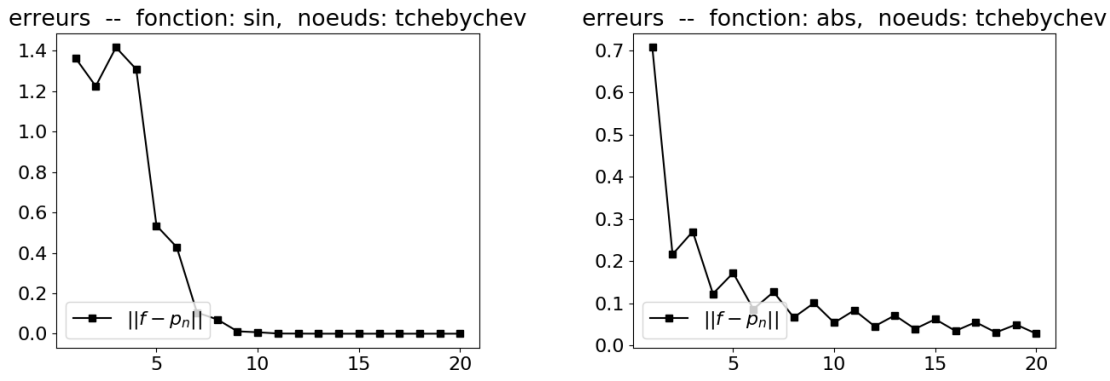

FIGURE 4 – Tracé des erreurs $\|f - p_n\|_{L^\infty}$ en fonction de n, pour les interpolations sur des noeuds de Tchebychev représentées dans la figure 3.

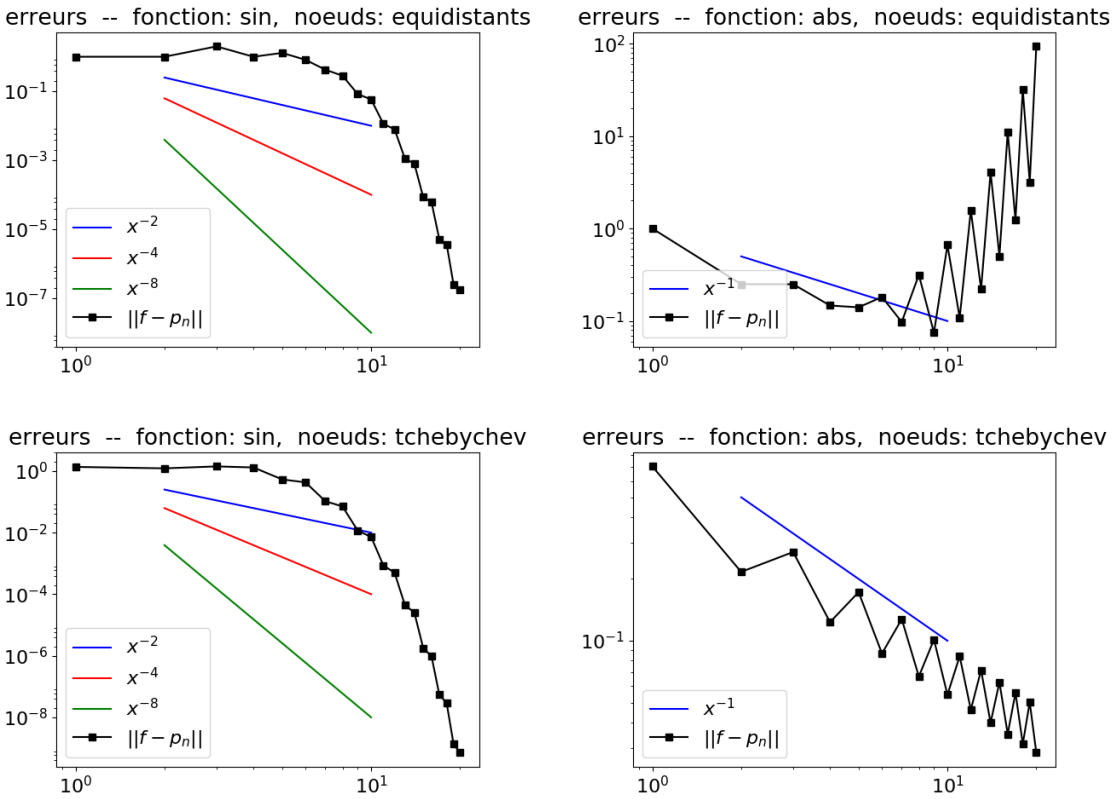

FIGURE 5 – Reproduction des courbes d'erreur représentées dans les figures 2 et 4 en échelle logarithmique, pour une meilleure appréciation des vitesses de convergence. La superposition avec des courbes dont la décroissance est prescrite (de la forme $x \mapsto x^{-\gamma}$ avec $\gamma > 0$) permet ici de préciser ces vitesses de convergence.

au regard de l'estimation énoncée dans le théorème 1.1.1. Avec des **noeuds de Tchebychev**, leur qualité est bien **meilleure**.

- **L'analyse** mathématique va nous permettre de prouver des **estimations a priori** qui établiront la **convergence**, et parfois la **vitesse de convergence**, du procédé d'approximation lorsque N tend vers l'infini.

Organisation du cours : Le plan de ces notes est le suivant. Dans le chapitre 2 nous présenterons en détail la construction de plusieurs méthodes importantes d'approximation de fonctions (interpolations polynomiales, approximations polynomiales non-interpolantes, approximations polynomiales par morceaux, approximations par des séries de Fourier) et nous établirons certaines de leurs propriétés, lorsque leur démonstration peut être faite en quelques lignes. Nous changerons temporairement de sujet au chapitre 3 pour étudier l'approximation de *solution d'équations*, essentiellement par des méthodes de point fixe. Nous reprendrons l'analyse des méthodes d'approximation de fonctions au chapitre 4, qui sera consacré aux preuves les plus élaborées. Nous suivrons alors deux objectifs : le premier sera de démontrer l'estimation du théorème 1.1.1 en nous appuyant sur des résultats concernant l'approximation par séries trigonométriques. Le deuxième objectif sera d'obtenir des estimations a priori pour l'interpolation polynomiale, de façon notamment à mieux comprendre les résultats numériques observés dans cette introduction. Le chapitre 5 concluera le cours par l'étude de diverses méthodes permettant de calculer des intégrales de façon approchée. Ces méthodes, qualifiées parfois de *quadratures numériques*, représentent une application très importante des techniques étudiées dans les chapitres précédents.

2 Quelques méthodes d'approximation de fonctions

On présente ici plusieurs méthodes d'approximation de fonctions. On commencera par décrire des méthodes d'approximation polynomiales, et on en établira quelques propriétés lorsque leur démonstration peut se faire en quelques lignes. On étudiera en particulier l'approximation polynomiale par morceaux qui est très utilisée en pratique. On terminera par quelques résultats sur les séries trigonométriques, qui seront à nouveau utilisées dans le chapitre 4 pour démontrer des estimations a priori importantes pour l'approximation polynomiale. 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.

2.1 Développements polynomiaux de Taylor

Un exemple très classique d'approximation polynomiale est donné par les polynômes de Taylor qui apparaissent dans le développement limité d'une fonction au voisinage d'un point $x_0 \in I$. Rappelons-les, en supposant pour simplifier que f est $\mathcal{C}^\infty$ sur l'intervalle I : pour tout $x \in I$, des intégrations par parties successives donnent

$$\begin{aligned}
f(x) &= f(x_0) + \int_{x_0}^{x} f'(y)\,\mathrm{d}y \\
&= f(x_0) - [(x-y)f'(y)]_{x_0}^{x} + \int_{x_0}^{x} (x-y)f''(y)\,\mathrm{d}y \\
&= f(x_0) + (x-x_0)f'(x_0) + \int_{x_0}^{x} (x-y)f''(y)\,\mathrm{d}y \\
&= f(x_0) + (x-x_0)f'(x_0) - \left[\frac{(x-y)^2}{2}f''(y)\right]_{x_0}^{x} + \int_{x_0}^{x} \frac{(x-y)^2}{2} f^{(3)}(y)\,\mathrm{d}y \\
&= f(x_0) + (x-x_0)f'(x_0) + \frac{(x-x_0)^2}{2}f''(x_0) + \int_{x_0}^{x} \frac{(x-y)^2}{2} f^{(3)}(y)\,\mathrm{d}y \\
&(\cdots) \\
&= f(x_0) + (x-x_0)f'(x_0) + \cdots + \frac{(x-x_0)^n}{n!}f^{(n)}(x_0) + \int_{x_0}^{x} \frac{(x-y)^n}{n!} f^{(n+1)}(y)\,\mathrm{d}y.
\end{aligned} \tag{2.15}$$

Cette égalité est bien sûr la formule de Taylor avec reste intégral.

Définition 2.1.1 *Soit f une fonction de classe $\mathcal{C}^n$ sur un voisinage de x_0. Son* **polynôme de Taylor** *de degré n en x_0 est donné par*

$$\mathcal{T}_n f(x) = \sum_{k=0}^{n} \frac{(x-x_0)^k}{k!} f^{(k)}(x_0).$$

En pratique l'approximation par des polynômes de Taylor de degrés élevés est d'un intérêt limité, et ceci pour deux raisons au moins. D'une part, l'évaluation des dérivées d'ordre élevé n'est pas toujours une tâche immédiate pour une fonction arbitraire. D'autre

part, l'opérateur $\mathcal{T}_n$ n'est bien défini que pour des fonctions de classe $\mathcal{C}^n$, de sorte que la question de sa convergence n'a de sens que pour des fonctions de classe $\mathcal{C}^\infty$. En fait, si l'on souhaite que $\mathcal{T}_n f$ converge vers f (au sens de la norme sup sur I), alors f doit être développable en série entière sur un intervalle contenant I, ce qui est une hypothèse très forte et bien plus restrictive que la simple régularité $\mathcal{C}^\infty$. Cela se conçoit assez facilement si l'on pense au fait que les valeurs d'une telle fonction f doivent être prescrites sur tout l'intervalle I par ses valeurs sur un ouvert arbitrairement petit contenant x_0 : d'un point de vue pratique, on comprend intuitivement que ce type de propriété globale ne sera en général pas vérifié par une fonction quelconque, quand bien même celle-ci serait très régulière localement.

D'un point de vue théorique, il est toutefois intéressant de constater que des estimations a priori peuvent s'obtenir facilement.

Proposition 2.1.1 *Soit $f \in \mathcal{C}^{(n+1)}(I)$ et $x_0 \in I$. On a*

$$\|f - \mathcal{T}_n f\|_{L^\infty(I)} \leq \frac{|I|^{n+1}}{(n+1)!} \|f^{(n+1)}\|_{L^\infty(I)} \tag{2.16}$$

où $|I|$ est la longueur de l'intervalle I.

Preuve. D'après la formule avec reste intégral (2.15), nous avons pour tout $x \in I$

$$|f(x) - \mathcal{T}_n f(x)| \leq \int_{x_0}^{x} \left| \frac{(x-y)^n}{n!} f^{(n+1)}(y) \right| \mathrm{d}y \leq \left(\int_{x_0}^{x} \frac{|(x-y)^n|}{n!} \mathrm{d}y \right) \|f^{(n+1)}\|_{L^\infty(I)}.$$

Pour $x \geq x_0$, nous avons $\int_{x_0}^{x} \frac{|(x-y)^n|}{n!} \mathrm{d}y = \int_{x_0}^{x} \frac{(x-y)^n}{n!} \mathrm{d}y = \frac{(x-x_0)^{n+1}}{(n+1)!}$ et le même calcul s'applique avec un changement de signe pour $x \leq x_0$. On en déduit que

$$|f(x) - \mathcal{T}_n f(x)| \leq \frac{|x-x_0|^{n+1}}{(n+1)!} \|f^{(n+1)}\|_{L^\infty(I)} \leq \frac{|I|^{n+1}}{(n+1)!} \|f^{(n+1)}\|_{L^\infty(I)} \qquad \forall x \in I,$$

ce qui prouve le résultat. □

Remarque 2.1.1 *En appliquant l'estimation (2.16) à des intervalles de la forme $I = [x_0 - r, x_0]$ ou $I = [x_0, x_0 + r]$ pour $r > 0$, on obtient une estimation légèrement meilleure,*

$$\|f - \mathcal{T}_n f\|_{L^\infty(I_r)} \leq \frac{|r|^{n+1}}{(n+1)!} \|f^{(n+1)}\|_{L^\infty(I_r)} \tag{2.17}$$

pour un intervalle de la forme

$$I_r = [x_0 - r, x_0 + r] \quad \textit{avec} \quad r > 0.$$

Les tests numériques présentés dans les figures 6 et 7 permettent de se faire une idée des performances de l'approximation de Taylor sur les fonctions $f_{\sin}$ et f_{Runge} définies en (1.11) et (1.12). Le cas de la fonction sinus est très particulier (toutes les dérivées sont immédiates à calculer) et très favorable, en effet l'estimation (2.16) entraîne que la suite des polynômes de Taylor converge sur tout intervalle borné : la figure 6 (à gauche)

permet vérifier que c'est bien le cas en pratique. Le cas de la fonction de Runge est moins favorable : d'une part son approximation nécessite de calculer ses dérivées d'ordres élevés, d'autre part sa convergence semble confinée à des petits sous-intervalles de I. Ce constat peut être fait sur la figure 6 (à droite) où l'on a tracé les polynômes de Taylor en $x_0 = 0$, et les approximations représentées sur la figure 7 indiquent que c'est toujours le cas lorsque le développement est fait autour des points $x_0 = 0.1$ ou 0.2.

On a ensuite représenté dans les figures 8 et 9 les valeurs (évaluées numériquement) des bornes supérieures apparaissant dans l'estimation a priori (2.17), en fonction du rayon r, pour les approximations tracées dans les figures 6 et 7. L'évolution de ces courbes pour diverses valeurs de n nous renseigne sur les propriétés de convergence des approximations de Taylor : pour la fonction $f_{\sin}$ nous vérifions à nouveau le caractère développable en série entière sur tout intervalle borné, et pour la fonction de Runge ces courbes corroborent l'observation faite plus haut selon laquelle f_{Runge} ne semble développable en série entière que sur un intervalle de rayon ≈ 0.2 autour des différents points x_0 testés. Ces courbes permettent également de comparer numériquement la croissance rapide des dérivées de la fonction de Runge, par rapport à celles d'une fonction sinusoidale.

Remarque 2.1.2 *Si l'on voulait reprendre le cadre théorique introduit dans la section 1.1 pour étudier la convergence des développements de Taylor, on pourrait penser à poser $X = \mathcal{C}^\infty(I)$. Dans cet espace les opérateurs $\mathcal{T}_n$ sont bien définis, mais la norme L^∞ devient problématique car $\mathcal{C}^\infty(I)$ n'est plus complet pour cette norme (le vérifier en exhibant une suite de Cauchy qui ne converge pas dans $\mathcal{C}^\infty(I)$), d'autre part il n'est pas très difficile de voir que pour $n \geq 1$, $\mathcal{T}_n$ n'est pas stable dans L^∞ au sens de* (1.1).

2.2 Interpolation polynomiale

Nous revenons maintenant sur l'approximation polynomiale par interpolation vue dans l'introduction. Nous présenterons plusieurs méthodes d'interpolation, et nous établirons certaines de leurs propriétés. Commençons par un point de terminologie.

- L'**interpolation de Lagrange** désigne l'interpolation des valeurs d'une fonction en des points donnés.
- L'**interpolation de Hermite** désigne l'interpolation des valeurs d'une fonction et de ses dérivées, aux mêmes points.
- L'interpolation de Lagrange est parfois qualifiée d'**interpolation simple** par opposition avec l'interpolation de Hermite.
- Pour chaque méthode (Lagrange ou Hermite), il est nécessaire de spécifier des points (ou noeuds) d'interpolation. On peut ainsi parler d'**interpolation de Tchebychev** pour désigner l'interpolation de Lagrange associée aux points de Tchebychev. De même, on pourra parler d'**interpolation équidistante** pour désigner l'interpolation de Lagrange associée à des noeuds équidistants.

2.2.1 Interpolation de Lagrange

On se place sur un intervalle $I = [a, b]$ et on se donne $n + 1$ points distincts sur cet intervalle :

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

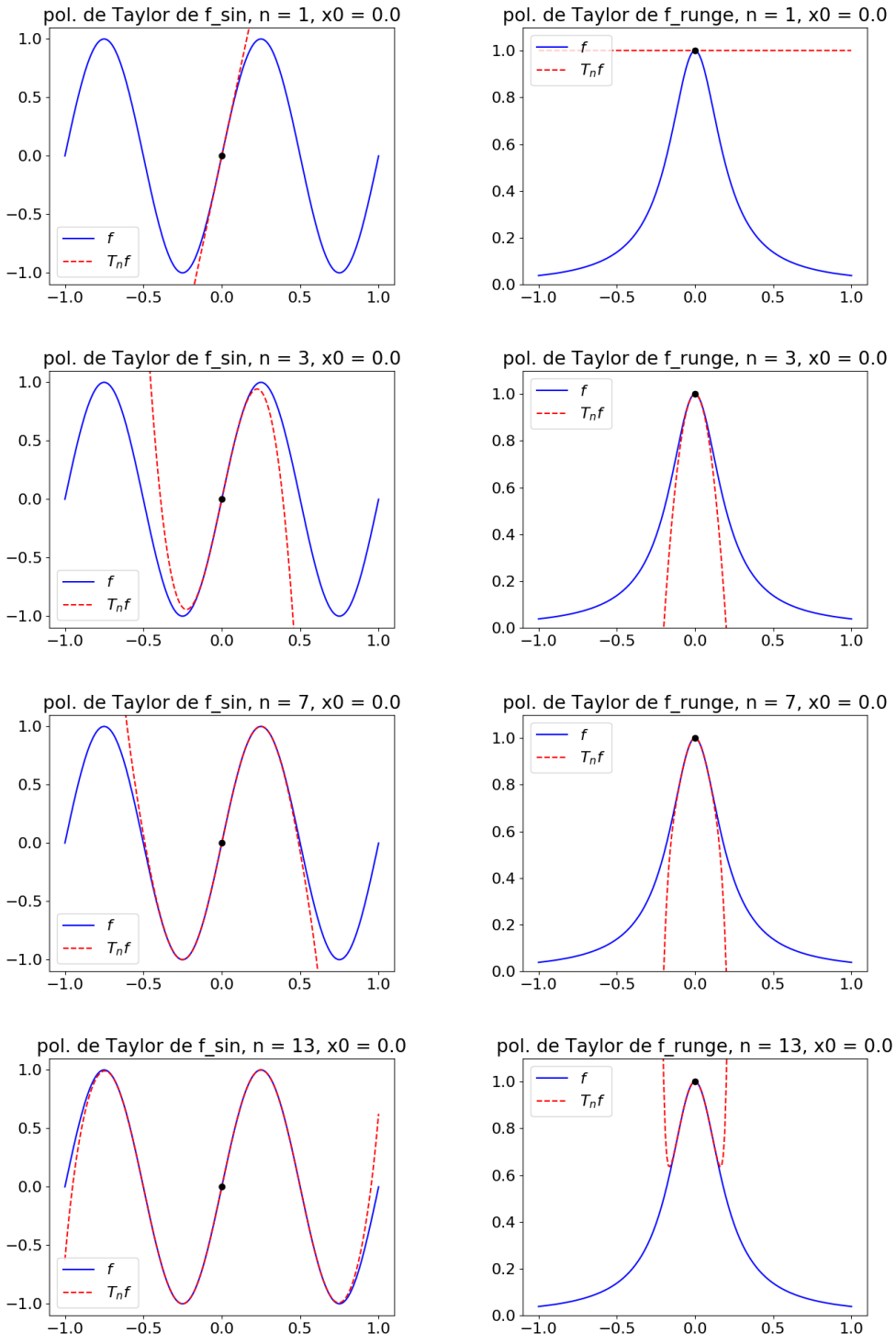

FIGURE 6 – Approximation des fonctions $f_{\sin}$ (à gauche) et f_{Runge} (à droite) définies en (1.11) et (1.12) par leurs polynômes de Taylor en $x_0 = 0$. Les degrés $n = 3$, 5, 9 et 15 sont représentés. Les approximations semblent converger sur $I = [-1, 1]$ pour la fonction sinus, pour la fonction de Runge la précision semble confinée à un sous-intervalle de rayon ≈ 0.2.

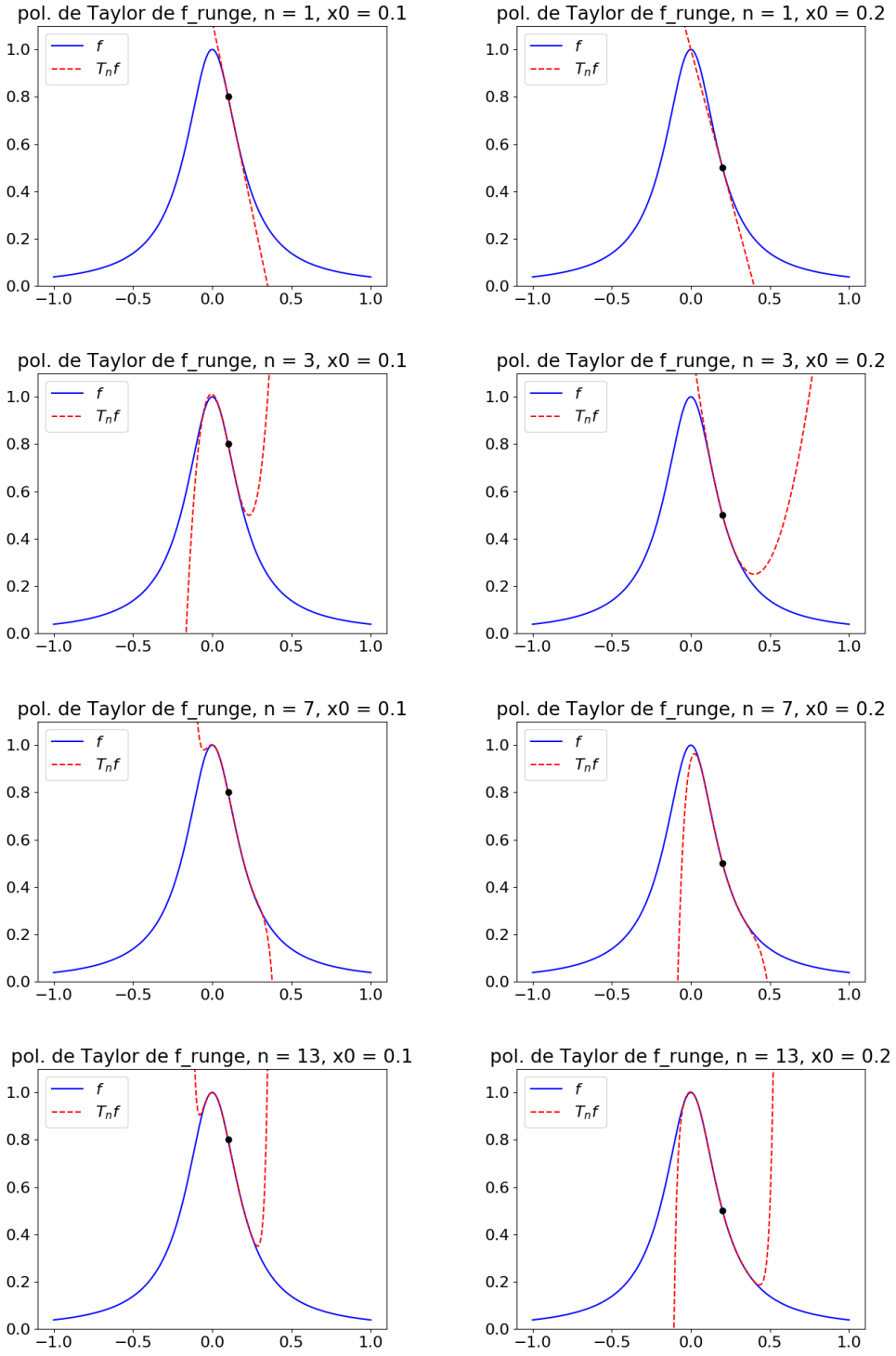

FIGURE 7 – Approximation de la fonction f_{Runge} définie en (1.12) par ses polynômes de Taylor en $x_0 = 0.1$ (à gauche) et $x_0 = 0.2$ (à droite) pour différents degrés n. A nouveau, la précision semble confinée à des sous-intervalles stricts de $I = [-1, 1]$.

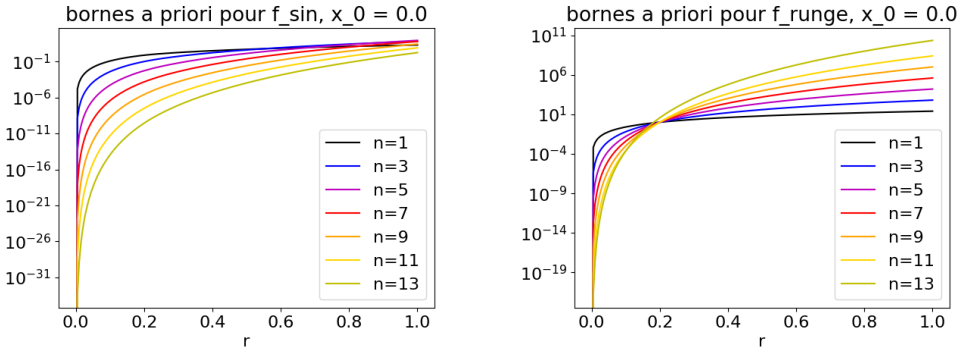

FIGURE 8 – Tracé des bornes $\frac{|r|^{n+1}}{(n+1)!}||f^{(n+1)}||_{L^\infty([x_0-r,x_0+r])}$ figurant dans l'estimation a priori (2.17), en fonction de r, pour différentes valeurs de n. Ici f et x_0 correspondent aux approximations représentées dans la figure 6.

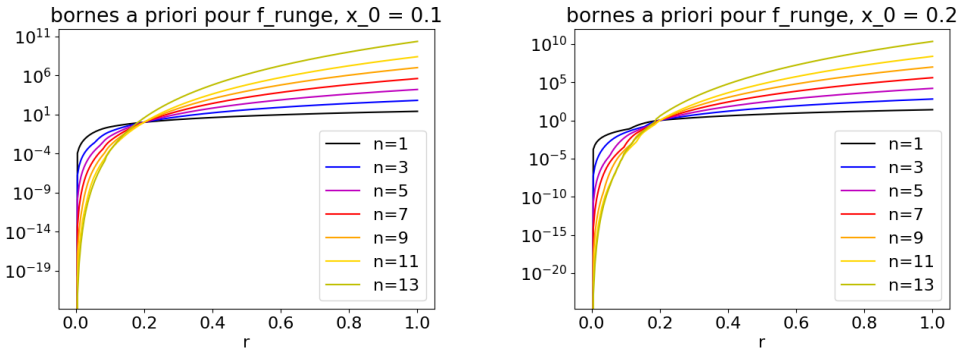

FIGURE 9 – Tracé des bornes $\frac{|r|^{n+1}}{(n+1)!}||f^{(n+1)}||_{L^\infty([x_0-r,x_0+r])}$ figurant dans l'estimation a priori (2.17), en fonction de r, pour différentes valeurs de n. Ici f et x_0 correspondent aux approximations représentées dans la figure 7.

Etant donné un ensemble de $n+1$ réels $\{y_0, \cdots, y_n\}$, 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 2.2.1 *Pour tout ensemble de réels $\{y_0, \cdots, y_n\}$, il existe un unique polynôme $p_n \in \mathbb{P}_n$ tel que $p_n(x_i) = y_i$ pour tout $i = 0, \cdots, n$.*

Preuve. Ceci revient à montrer que l'application $L : \mathbb{P}_n \to \mathbb{R}^{n+1}$ qui à $p \in \mathbb{P}_n$ associe le vecteur de coordonnées $(p(x_0), \cdots, p(x_n))$ est bijective. Cette application est linéaire, et $\dim(\mathbb{P}_n) = \dim(\mathbb{R}^{n+1})$. 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 $x_0, \cdots, x_n$ ce qui n'est possible que si $p = 0$ puisque c'est un polynôme de degré n. □

Remarque 2.2.1 *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 $i = 0, \cdots, n$, sont alors équivalentes au système $(n+1) \times (n+1)$*

$$Va = y \tag{2.18}$$

où a et y sont les vecteurs de coordonnées $(a_0, \cdots, a_n)$ et $(y_0, \cdots, y_n)$ et où $V = (x_i^j)_{i,j=0,\cdots,n}$ est la matrice de Vandermonde associée aux points $x_0, \cdots, x_n$. 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 des valeurs $y_0, \cdots, y_n$ aux points $x_0, \cdots, x_n$. Au vu de l'équation (2.18), on pourrait vouloir inverser la matrice V pour *calculer* les coefficients de p_n. Cette méthode n'est pas recommandée car l'inversion de la matrice V sur un ordinateur entraîne des erreurs d'arrondis très importantes lorsque n augmente, ce qui est lié au (très) mauvais conditionnnement de la matrice de Vandermonde. Une méthode plus efficace consiste à déterminer une base polynomiale dans laquelle les coefficients de p_n sont triviaux.

Plus précisément, en utilisant le théorème 2.2.1 on peut définir pour tout $i = 0, \cdots, n$ un unique polynôme $\ell_i \in \mathbb{P}_n$ tel que

$$\ell_i(x_j) = 0 \text{ si } i \neq j, \qquad \text{et } \ell_i(x_i) = 1.$$

On a alors

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

Pour pouvoir utiliser cette méthode en pratique il ne reste plus qu'à déterminer la forme des polynômes ℓ_i, ce qui est aisé puisqu'on connaît toutes leurs racines : on vérifie en effet que

$$\ell_i(x) = \frac{\prod_{j \in \{0,\cdots,n\}-\{i\}} (x - x_j)}{\prod_{j \in \{0,\cdots,n\}-\{i\}} (x_i - x_j)} = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}.$$

(lorsque $n = 0$ les produits ci-dessus sont vides et on a $\ell_0(x) = 1$). Enfin, en utilisant le fait que $p_n = 0$ si et seulement si tous les y_i sont nuls, on vérifie que la famille $\{\ell_0, \cdots, \ell_n\}$ constitue une base de $\mathbb{P}_n$.

Définition 2.2.1 *Les fonctions ℓ_i sont appelées* **polynômes de base de Lagrange** *de degré n pour les points $\{x_0, \cdots, x_n\}$.*

Bien évidemment, les polynômes ℓ_i dépendent du degré n choisi pour l'interpolation, et plus généralement de la *position* des noeuds $x_0, \ldots, x_n$. Pour ne pas alourdir les notations cette dépendance n'est pas explicite dans la notation, mais elle est rendue évidente dans la figure 10 qui montre le graphe des 4 premières fonctions de Lagrange ℓ_i dans le cas $n = 8$, pour des noeuds equidistants et de Tchebychev. (On pourra tracer à la main les graphes correspondant aux cas simples $n = 1$ et $n = 2$.)

Dans le cas où les y_i sont les valeurs d'une fonction f aux points x_i, on arrive à la définition suivante.

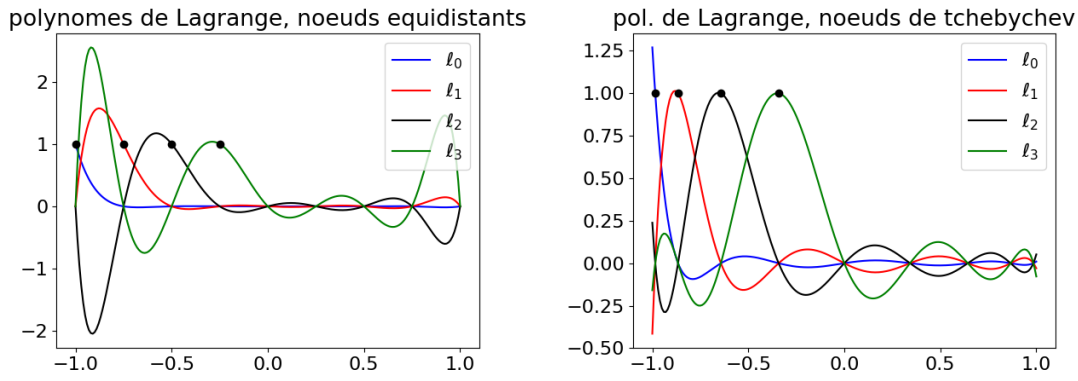

FIGURE 10 – Tracé des 4 premiers polynômes de base de Lagrange ℓ_i, dans le cas $n = 8$, pour des noeuds equidistants (à gauche) et de Tchebychev (à droite). La réduction des oscillations pour les noeuds de Tchebychev, clairement visible ici, correspond au phénomène vu dans l'introduction.

Définition 2.2.2 *Soit f une fonction continue sur I, et $x_0, \dots, x_n$ des points distincts de I. On appelle* **polynôme d'interpolation de Lagrange** *de f en ces points, l'unique $p_n \in \mathbb{P}_n$ tel que $p_n(x_i) = f(x_i)$ pour $i = 0, \cdots, n$.*

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

$$p_n(x) = \sum_{i=0}^{n} f(x_i)\ell_i(x), \tag{2.19}$$

dite *forme de Lagrange*. Cette forme permet de vérifier que l'opérateur d'interpolation

$$\mathcal{I}_n : f \mapsto p_n \tag{2.20}$$

est stable pour la norme $L^\infty(I)$ au sens de (1.1).

Proposition 2.2.1 (stabilité L^∞ des interpolations) *Pour tout $n \in \mathbb{N}$, il existe un réel C_n pour lequel*

$$\|\mathcal{I}_n f\|_{L^\infty(I)} \leq C_n \|f\|_{L^\infty(I)}, \qquad \forall f \in L^\infty. \tag{2.21}$$

Preuve. En utilisant la forme (2.19), on écrit que

$$|\mathcal{I}_n f(x)| \leq \sum_{i=0}^{n} |f(x_i)||\ell_i(x)| \leq C_n \|f\|_{L^\infty(I)}$$

pour tout $x \in I$, avec

$$C_n = \Big\| \sum_{i=0}^{n} |\ell_i| \Big\|_{L^\infty(I)}.$$

□

Remarque 2.2.2 *La constante C_n qui apparaît dans l'estimation* (2.21) *ne dépend pas seulement du nombre des noeuds d'interpolation, mais également de leur position dans l'intervalle I, en effet c'était déjà le cas des polynômes ℓ_i, cf. figure 10.*

On peut d'autre part observer que $\mathcal{I}_n$ est exact sur les polynômes de degré $\leq n$, et ceci quelque soit le choix des noeuds d'interpolation. En effet, il est facile de déduire du théorème 2.2.1 que

$$f \in \mathbb{P}_n \quad \Longleftrightarrow \quad \mathcal{I}_n f = f.$$

Remarque 2.2.3 (stabilité uniforme et phénomène de Runge) *Si la suite des C_n apparaissant dans* (2.21) *était bornée, on pourrait parler de* stabilité uniforme *au sens de* (1.8)*, et la proposition 1.1.1 permettrait d'affirmer le caractère quasi-optimal des interpolations au sens de* (1.7)*. Comme nous l'avons vu dans l'introduction, la situation n'est pas aussi simple. En effet si la stabilité était uniforme la suite des fonctions $\mathcal{I}_n f$ serait bornée dans L^∞ pour toute fonction f, or les résultats numériques présentés dans les figures 1 et 2 montrent que ce n'est a priori pas le cas : l'interpolation de la fonction valeur absolue sur des noeuds équidistants présente des oscillations dont l'amplitude diverge lorsque n augmente. Avec la fonction f_{Runge} définie en* (1.12)*, le mathématicien Runge a montré que ces oscillations s'observaient également avec des fonctions de classe $\mathcal{C}^\infty$. La figure 11 reproduit ce phénomène d'instabilité asymptotique (i.e., pour $n \to \infty$), et montre que les approximations sont bien meilleures lorsque les interpolations sont définies sur les noeuds de Tchebychev. L'étude fine de la stabilité uniforme des interpolations sera abordée dans la section 4.3.2.*

Une autre méthode très efficace pour calculer le polynôme interpolant consiste à l'exprimer sous une forme différente dite *forme de Newton*. On définit par récurrence les **différences divisées** de f en posant pour tout les points $x_0 < \cdots < x_n$,

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

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

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

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

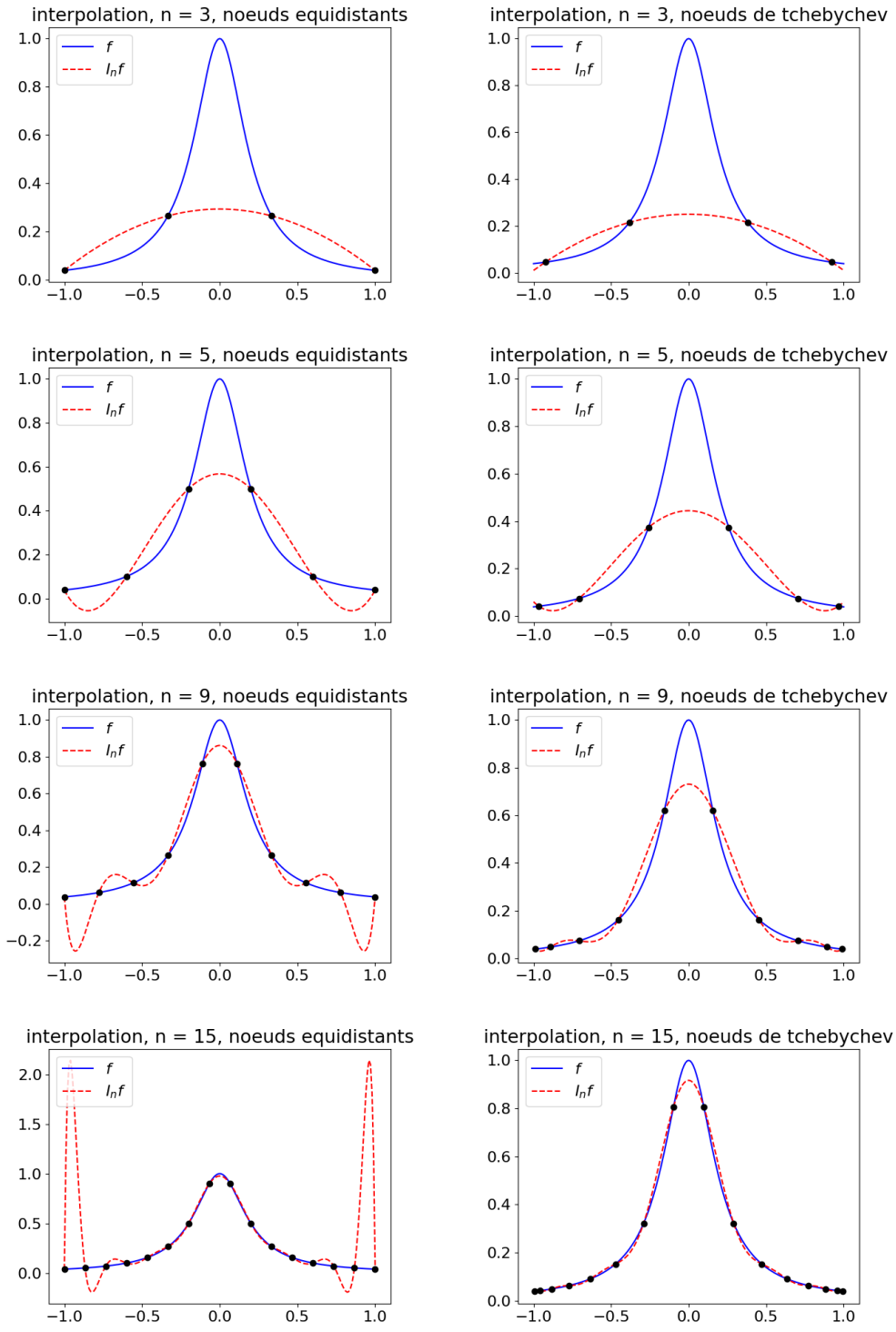

FIGURE 11 – Approximation de la fonction $f = f_{\text{Runge}}$ donnée en (1.12) par ses polynômes interpolateurs de Lagrange de degré $n = 3$, 5, 9 et 15. A gauche, l'interpolation est définie sur les noeuds équidistants, on observe à nouveau des oscillations près des bords qui augmentent avec n : c'est le phénomène de Runge. A droite l'interpolation est définie sur les noeuds de Tchebychev : les grandes oscillations ont disparu.

Proposition 2.2.2 *Le polynôme d'interpolation de Lagrange de f aux points $x_0, \cdots, x_n$ peut aussi s'écrire sous la* **forme de Newton**

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

Preuve. On procède par récurrence sur n. Pour $n = 0$, 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 $n - 1$. Alors p_{n-1} est le polynôme de degré $n-1$ interpolant f en $x_0, \ldots x_{n-1}$, de sorte que $p_n - p_{n-1}$ est un polynôme de degré n qui s'annule aux points $x_0, \cdots, x_{n-1}$. Par conséquent il existe $\tau \in \mathbb{R}$ tel que

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

Il reste à montrer que $\tau = f[x_0, \cdots, x_n]$. Pour cela on remarque que τ 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 $f[x_0, \cdots, x_n]$. 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 $x_1, \cdots, x_n$. On voit ainsi que $q_n \in \mathbb{P}_n$ et on vérifie aisément que

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

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 $f[x_1, \cdots, x_n]$ et $f[x_0, \cdots, x_{n-1}]$. Par conséquent, le coefficient de x^n dans p_n est donné par

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

ce qui conclut la preuve. □

Remarque 2.2.4 *A partir du fait que $f[x, y] = f[y, x]$, on peut établir que la différence divisée $f[x_0, \cdots, x_n]$ est invariante par permutation des indices :*

$$f[x_0, \cdots, x_n] = f[x_{\varphi(0)}, \cdots, x_{\varphi(n)}],$$

pour toute bijection φ de $\{0, \cdots, n\}$ dans lui-même. En particulier la forme de Newton n'exige pas que les x_i soient rangés par ordre croissant.

On renvoie chapitre 4.3 pour une analyse détaillée des erreurs d'interpolation, et des estimations a priori qui permettront de confirmer la qualité des noeuds de Tchebychev sur les noeuds équidistants.

2.2.2 Interpolation de Hermite

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

Théorème 2.2.2 *Pour tout ensemble de réels $\{\alpha_0, \cdots, \alpha_n, \beta_0, \cdots, \beta_n\}$, il existe un unique polynôme $q \in \mathbb{P}_{2n+1}$ tel que*

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

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

Preuve. Comme dans la preuve du théorème 2.2.1, il suffit de montrer que l'application linéaire $L : \mathbb{P}_{2n+1} \to \mathbb{R}^{2n+2}$ qui à $p \in \mathbb{P}_{2n+1}$ associe le vecteur de coordonées $(p(a), \cdots, p^{(n)}(a), p(b), \cdots, p^{(n)}(b))$ est injective. Or si ce vecteur s'annule, cela signifie que p est de la forme

$$p(x) = (x-a)^{n+1}(x-b)^{n+1}r(x),$$

où r est un polynôme, ce qui n'est possible que si $r = 0$ puisque $p \in \mathbb{P}_{2n+1}$. Par conséquent $p = 0$. □

Définition 2.2.3 *Soit f une fonction de classe $\mathcal{C}^n$ sur un intervalle I et soient $a < b$ 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 $q_{2n+1} \in \mathbb{P}_{2n+1}$ tel que*

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

On notera

$$\mathcal{H}_n : f \mapsto q_{2n+1} \tag{2.22}$$

l'opérateur associé à l'interpolation de Hermite d'ordre n. En généralisant l'approche suivie pour l'interpolation de Lagrange, il est possible d'exprimer l'interpolation de Hermite à l'aide de fonctions de base sous la forme

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

Cette écriture permet de démontrer la stabilité de l'interpolation de Hermite dans l'espace $\mathcal{C}^n(I)$ muni d'une norme appropriée, comme l'énonce la proposition suivante dont la preuve est laissée en exercice (on pourra s'inspirer de la preuve de la proposition 2.2.1).

Proposition 2.2.3 *Pour tout $n \in \mathbb{N}$, il existe un réel C_n pour lequel*

$$\|\mathcal{H}_n f\|_{\mathcal{C}^n(I)} \leq C_n \|f\|_{\mathcal{C}^n(I)}, \qquad \forall f \in \mathcal{C}^n(I) \tag{2.23}$$

où la norme de $\mathcal{C}^n$ est définie en (1.6).

On notera que l'expression de ces polynômes de base n'est pas simple à déterminer pour des valeurs quelconques de n. 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.

2.3 Approximation polynomiale des moindres carrés

Dans les sections précédentes on a considéré des approximations polynomiales dont la construction reposait sur un principe d'interpolation (interpolation des dérivées successives en un point pour Taylor, interpolation simple pour Lagrange et interpolation des dérivées successives en deux points pour Hermite), et pouvait donc sembler naturelle. L'étude de leur précision était alors envisagée dans un deuxième temps. Dans cette section on va suivre une approche différente, en *définissant d'abord l'approximation par une propriété de précision*, choisie de telle sorte que sa construction effective soit possible par un algorithme simple. Le principe de base sur laquelle repose cette approche est celui des projections orthogonales. Plus précisément, on rappelle le résultat suivant.

Théorème 2.3.1 *Si X est muni d'un produit scalaire $\langle\cdot,\cdot\rangle_X$ et E_N est un sous-espace de dimension $N \in \mathbb{N}$, alors pour tout $f \in X$ il existe un unique $f_N \in E_N$ tel que*

$$\langle f_N, g\rangle_X = \langle f, g\rangle_X \qquad \forall g \in E_N. \tag{2.24}$$

L'opérateur $P_N : f \to f_N$ est une projection sur E_N, appelée projection orthogonale. *Cet opérateur est également caractérisé par l'égalité*

$$\|f - P_N f\|_X = \inf_{g\in E_N} \|f - g\|_X. \tag{2.25}$$

Preuve. Si l'on note $a_1, \ldots, a_N$ les coefficients de f_N dans une base $\{g_i : i = 1, \ldots, N\}$ de E_N, la relation (2.24) prend la forme $\sum_i a_i \langle g_i, g_j\rangle_X = \langle f, g_j\rangle_X$ pour tout $j = 1, \ldots, N$, qu'on peut écrire comme une égalité dans $\mathbb{R}^N$,

$$Ma = b \tag{2.26}$$

en notant $b_j = \langle f, g_j\rangle_X$ et $M_{i,j} = \langle g_i, g_j\rangle_X$ pour $i, j = 1, \ldots N$. La matrice M (qu'on appelle parfois "matrice de masse") est carrée et inversible, en effet si $Mc = 0$ pour un vecteur $c \in \mathbb{R}^N$ alors par multiplication avec le vecteur transposé c^t on a $0 = c^t Mc$, i.e.

$$0 = \sum_{i,j} c_i \langle g_i, g_j\rangle_X c_j = \Big\langle \sum_i c_i g_i, \sum_j c_j g_j \Big\rangle_X = \Big\| \sum_i c_i g_i \Big\|_X^2$$

ce qui entraîne que la fonction $\sum_i c_i g_i$ est nulle et donc que tout ses coefficients le sont. En particulier l'équalité (2.26) définit bien un unique vecteur a, ce qui signifie que (2.24) définit un unique $f_N \in E_N$. Le fait que P_N soit une projection est clair en effet on a bien $P_N f = f$ pour $f \in E_N$, et (2.24) exprime bien une propriété d'orthogonalité puisque $\langle f_N - f, g\rangle_X = 0$ pour tout $g \in E_N$. Pour montrer que l'égalité (2.25) caractérise la projection f_N, on note $\tilde{g} = f_N - g$ et on calcule ensuite

$$\|f - \tilde{g}\|_X^2 = \langle f - \tilde{g}, f - \tilde{g}\rangle_X = \|f - f_N\|_X^2 + \|f_N - \tilde{g}\|_X^2 \tag{2.27}$$

(c'est le théorème de Pythagore), d'où l'on déduit que

$$\|f - f_N\|_X \le \|f - \tilde{g}\|_X \qquad \forall \tilde{g} \in E_N$$

de sorte que $P_N f = f_N$ vérifie bien (2.25). Pour terminer la preuve on montre l'unicité du meilleur approximant : si $\tilde{f}_N \in E_N$ vérifie également (2.25) alors $\|f - f_N\|_X = \|f - \tilde{f}_N\|_X$, et l'égalité (2.27) appliquée à $\tilde{g} = \tilde{f}_N$ montre que $\|f_N - \tilde{f}_N\|_X = 0$. On a ainsi vérifié que la propriété (2.25) et les relations (2.24) définissaient le même élément de E_N. □

Remarque 2.3.1 *Les conclusions du théorème précédent restent vraies si l'on considère un produit bilinéaire semi-défini positif* $\langle \cdot, \cdot \rangle_X$

2.3.1 Moindres carrés "discrets"

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 $x_0 < \cdots < x_m$ avec $m > n$, on peut chercher à construire un polynôme de degré n qui approche f par un autre procédé.

Plus précisément, le résultat suivant montre qu'il existe un unique polynôme de $\mathbb{P}_n$ dont les valeurs aux points $x_0, \cdots, x_m$ minimisent une distance euclidienne avec des réels quelconques $y_0, \cdots, y_m$. On verra que ce polynôme peut être défini en faisant intervenir la projection orthogonale du vecteur $y \in \mathbb{R}^{m+1}$ sur un sous-espace de dimension $n+1$.

Théorème-Définition 2.3.1 *Soient $x_0, \cdots, x_m$ des réels distincts et $y_0, \cdots, y_m$ des réels quelconques. Si $m \geq n$, alors il existe un unique polynôme $q_n \in \mathbb{P}_n$ qui minimise la quantité*

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

parmi tous les $q \in \mathbb{P}_n$. On l'appelle **polynôme des moindres carrés** *de degré n.*

Preuve. Si l'on écrit $q_n(x) = \sum_{k=0}^{n} \hat{a}_k x^k$, on voit que la recherche de q_n est équivalente à celle d'un vecteur $\hat{a} = (\hat{a}_0, \cdots, \hat{a}_n) \in \mathbb{R}^{n+1}$ qui minimise la norme euclidienne dans $\mathbb{R}^{m+1}$

$$\|V\hat{a} - y\|_2 := \Big(\sum_{i=0}^{m} |(V\hat{a} - y)_i|^2 \Big)^{\frac{1}{2}}$$

où V est la matrice $(m+1) \times (n+1)$ dont les coefficients sont donnés par $v_{i,j} = x_i^j$. Pour étudier ce problème de minimisation, on considère $X = \mathbb{R}^{m+1}$ muni du produit scalaire euclidien $\langle y, z \rangle_2 := \sum_{i=0}^{m} y_i z_i$, et

$$E_N = \mathrm{Im}(V) = \{Va \in \mathbb{R}^{m+1} : a \in \mathbb{R}^{n+1}\}.$$

(On verra plus bas que V est injective, de sorte que la dimension de E_N est ici $N = n+1$.) Le théorème 2.3.1 s'applique alors et nous permet d'affirmer qu'il existe un unique $\hat{y} \in E_N$ tel que

$$\|\hat{z} - y\|_2 = \inf_{z \in E_N} \|z - y\|_2.$$

La forme de l'espace E_N permet alors d'écrire que $\inf_{z \in E_N} \|z - y\|_2 = \inf_{a \in \mathbb{R}^{n+1}} \|Va - y\|_2$, et d'autre part de voir qu'il existe un vecteur $\hat{a} \in \mathbb{R}^{n+1}$ tel que $\hat{z} = V\hat{a}$. Pour montrer que ce vecteur est unique on observe que $\ker V = \{0\}$: en effet si $Va = 0$ pour $a \in \mathbb{R}^{n+1}$ alors

$$0 = \langle Va, Va \rangle = \|Va\|_2^2 = \sum_{i=0}^{m} |(Va)_i|^2 = \sum_{i=0}^{m} |q(x_i)|^2$$

où on a posé $q(x) = \sum_{k=0}^{n} a_k x^k$. Ceci entraîne que q s'annule en $m+1$ points distincts : comme il est dans $\mathbb{P}_n$ avec $n \leq m$, il est forcément nul, et le vecteur a également. On a ainsi montré qu'il existait un unique $\hat{a} \in \mathbb{R}^{n+1}$ tel que

$$\|V\hat{a} - y\|_2 = \inf_{a \in \mathbb{R}^{n+1}} \|Va - y\|_2.$$

ce qui revient au résultat du théorème. □

En reprenant la preuve ci-dessus, on voit que le vecteur $\hat{z} = V\hat{a}$ est caractérisé par les relations de projection orthogonale (2.24), qui s'écrivent ici

$$\langle V\hat{a} - y, Va \rangle_2 = 0, \qquad \forall\, a \in \mathbb{R}^{n+1},$$

de sorte que les coefficients $\hat{a}_0, \cdots, \hat{a}_n$ de p_n sont solutions du système $(n+1) \times (n+1)$

$$V^t V\hat{a} = V^t y, \tag{2.28}$$

avec

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

On appelle *équations normales* le système matriciel (2.28). On observe que l'injectivité de la matrice V entraîne celle de V^tV (car si $V^tVa = 0$ alors $0 = a^tV^tVa = \|Va\|_2^2$) et que celle-ci étant carrée, elle est inversible : on vérifie ainsi bien que le système (2.28) caractérise le vecteur $\hat{a}$.

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 $\{(x_i, y_i),\ i = 0, \cdots, n\}$, et ses coefficients se calculent simplement à partir des valeurs x_k et y_k en résolvant un système 2×2.

Dans le cas où les y_i sont les valeurs d'une fonction f aux points x_i, on arrive à la définition suivante.

Définition 2.3.1 *Soit f une fonction continue sur un intervalle I et $x_0, \cdots, x_m$ des points distincts dans I. Le polynôme $q_n \in \mathbb{P}_n$ qui minimise la quantité*

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

est appelé **approximation des moindres carrés** *de degré n de f, aux points $x_0, \cdots, x_m$.*

2.3.2 Moindres carrés "continus" (projections orthogonales L^2)

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

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

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

$$\frac{b-a}{m+1} \sum_{i=0}^{m} |f(x_i) - q(x_i)|^2,$$

qui est minimisée par le polynôme des moindres carrés aux points $x_0, \cdots, x_m$ est alors une somme de Riemann qui approche l'intégrale ci-dessus lorsque le nombre de points m augmente. On peut vérifier que l'on définit une norme sur $\mathcal{C}^0(I)$ en posant

$$\|g\|_{L^2(I)} := \Big(\int_a^b |g(x)|^2 dx \Big)^{1/2}.$$

Cette norme est appelée norme L^2 sur l'intervalle $I = [a, b]$. On remarque qu'elle dérive du produit scalaire

$$\langle f, g \rangle := \int_a^b f(x) g(x) dx, \tag{2.29}$$

au sens où $\|g\|_{L^2(I)} := \sqrt{\langle g, g \rangle}$. On recherche donc le polynôme $q_n \in \mathbb{P}_n$ solution de

$$\|f - q_n\|_{L^2(I)} = \min_{q \in \mathbb{P}_n} \|f - q\|_{L^2(I)}. \tag{2.30}$$

Afin de prouver l'existence et l'unicité du polynôme q_n solution de (2.30), on introduit la suite des *polynômes de Legendre* qui est définie en appliquant le procédé d'orthogonalisation de Gramm-Schmidt aux fonctions $e_k : x \mapsto x^k$.

Définition 2.3.2 *La suite des* **polynômes de Legendre orthonormés** *sur* $[a, b]$ *est définie par récurrence en posant* $L_0 = \frac{e_0}{\|e_0\|_{L^2(I)}}$ *et*

$$L_n = \frac{e_n - \sum_{k=0}^{n-1} \langle e_n, L_k \rangle L_k}{\| e_n - \sum_{k=0}^{n-1} \langle e_n, L_k \rangle L_k \|_{L^2(I)}},$$

c'est-à-dire $L_0(x) = (b-a)^{-1/2}$ *et* $L_n(x) = \dfrac{x^n - \sum_{k=0}^{n-1} (\int_a^b t^n L_k(t) dt) L_k(x)}{\left(\int_a^b (t^n - \sum_{k=0}^{n-1} (\int_a^b s^n L_k(s) ds) L_k(t))^2 dt \right)^{1/2}}$.

On déduit aisément de cette définition que les polynômes de Legendre forment un ensemble orthonormé au sens où

$$\langle L_i, L_j \rangle = 0 \ \text{ si } \ i \neq j \ \text{ et } \ \langle L_i, L_i \rangle = \|L_i\|^2_{L^2(I)} = 1.$$

La famille $\{L_0, \cdots, L_n\}$ est une base orthonormée de $\mathbb{P}_n$, et il est aussi facile de vérifier que L_n est exactement de degré n.

Théorème-Définition 2.3.2 *Il existe un unique polynôme q_n qui minimise $\|f-q_n\|_{L^2(I)}$ parmi tous les $q \in \mathbb{P}_n$. Ce polynôme est caractérisé par la propriété $\langle f-q_n, q\rangle = 0$ pour tout $q \in \mathbb{P}_n$, et il est donné par*

$$q_n := \sum_{k=0}^{n} \langle f, L_k\rangle L_k \tag{2.31}$$

On l'appelle **projection orthogonale** *de f dans $\mathbb{P}_n$ (au sens du produit scalaire L^2 sur I).*

Preuve. L'unicité du polynôme q_n qui minimise l'erreur L^2 se déduit en utilisant à nouveau le théorème 2.3.1, cette fois avec $X = \mathcal{C}^0(I)$ muni du produit scalaire (2.29), et $E_N = \mathbb{P}_n$ (de dimension $N = n+1$), et nous savons que q_n est bien caractérisé par les relations de projection orthogonale (2.24). Comme les polynômes L_i, $i = 0, \ldots, n$, forment une base orthonormée de $\mathbb{P}_n$, on peut écrire tout g comme une combinaison linéaire des L_i et les équations (2.24) peuvent s'exprimer sous la forme

$$\langle q_n, L_i\rangle = \langle f, L_i\rangle \qquad \forall i \in \{0, \cdots, n\}.$$

Il suffit alors d'écrire $q_n = \sum_{i=0}^{n} c_i L_i$ et de prendre le produit de cette somme par L_j pour vérifier que $c_i = \langle q_n, L_i\rangle = \langle f, L_i\rangle$, ce qui permet de vérifier la forme annoncée pour q_n. □

Remarque 2.3.2 *Il est possible de donner un sens à la projection orthogonale de f sur $\mathbb{P}_n$ lorsque f n'est pas une fonction continue : il suffit en effet que f soit intégrable sur I pour que les produits scalaires $\langle f, L_k\rangle$ soient bien définis.*

Le fait que la projection orthogonale minimise les erreurs en norme L^2 va nous permettre d'utiliser les estimations a priori énoncées dans l'introduction. Comme celles-ci sont écrites en norme L^∞, on commence par vérifier le résultat suivant.

Proposition 2.3.1 *Soit*

$$P_n : \mathcal{C}^0(I) \to \mathbb{P}_n$$

l'opérateur qui associe à toute fonction continue f sur $I = [a, b]$ sa projection orthogonale dans $\mathbb{P}_n$ définie plus haut. Cet opérateur vérifie

$$\|f - P_n f\|_{L^2(I)} \leq (b-a)^{1/2} \inf_{q \in \mathbb{P}_n} \|f - q\|_{L^\infty(I)}, \qquad \forall f \in \mathcal{C}^0(I). \tag{2.32}$$

Preuve. On commence par observer que la norme L^2 sur $I = [a, b]$ peut être majorée par la norme sup sur I suivant

$$\|g\|_{L^2(I)} \leq (b-a)^{1/2} \|g\|_{L^\infty(I)} \tag{2.33}$$

pour toute fonction $g \in \mathcal{C}^0(I)$. La propriété de minimisation vérifiée par la projection orthogonale permet alors d'écrire

$$\|f - P_n f\|_{L^2(I)} \leq \|f - q\|_{L^2(I)} \leq (b-a)^{1/2} \|f - q\|_{L^\infty(I)}$$

pour tout $q \in \mathbb{P}_n$, ce qui entraîne (2.32). □

Remarque 2.3.3 *En observant que les projecteurs P_n, $n \in \mathbb{N}$, vérifient une propriété de stabilité uniforme faisant intervenir les normes L^2 et L^∞,*

$$\|P_n g\|_{L^2(I)} \leq (b-a)^{1/2} \|g\|_{L^\infty(I)}, \qquad \forall g \in \mathcal{C}^0(I)$$

(qui peut se déduire de l'égalité de Pythagore et de (2.33)), on pourrait raisonner comme dans la preuve de la proposition 1.1.1 et en déduire que $\|f - P_n f\|_{L^2} \leq C \inf_{q \in \mathbb{P}_n} \|f - q\|_{L^\infty}$ avec $C = 2(b-a)^{1/2}$. L'inégalité (2.32) a une meilleure constante car elle est obtenue par un argument plus direct, mais sa forme est la même.

En combinant l'inégalité (2.32) avec le théorème 1.2.1 de Weierstrass ou avec le théorème 1.1.1 sur les meilleures approximations polynomiales énoncés dans l'introduction, on obtient immédiatement le résultat suivant sur la vitesse de convergence de q_n vers f.

Proposition 2.3.1 *Pour toute fonction $f \in \mathcal{C}^0(I)$, l'erreur de projection orthogonale sur $\mathbb{P}_n$ vérifie*

$$\lim_{n \to +\infty} \|f - P_n f\|_{L^2(I)} = 0. \tag{2.34}$$

Si f est de classe $\mathcal{C}^m$ sur $I = [a, b]$, on a

$$\|f - P_n f\|_{L^2} \leq C_m \frac{\|f\|_{\mathcal{C}^m(I)}}{n^m}$$

avec une constante C_m dépendant de m et de $|I| = b - a$ mais indépendante de n et f.

La limite (2.34) exprimant la convergence des polynômes $P_n f$ vers f en norme L^2, 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 $[a, b]$. On peut alors établir le résultat suivant qui est classique pour les bases orthonormées en dimension finie.

Proposition 2.3.2 (égalité de Parseval) *La série de terme général $|\langle f, L_k \rangle|^2$ converge, et l'on a*

$$\|f\|^2_{L^2(I)} = \sum_{k=0}^{+\infty} |\langle f, L_k \rangle|^2. \tag{2.35}$$

Preuve. Si $f \in \mathbb{P}_n$, l'égalité (2.35) est une conséquence directe du caractère orthonormé de la base des polynômes de Legendre. Le résultat pour tout $f \in \mathcal{C}^0(I)$ s'obtient en combinant l'égalité de Pythagore

$$\|f\|^2_{L^2(I)} = \|P_n f\|^2_{L^2(I)} + \|f - P_n f\|^2_{L^2(I)} = \sum_{k=0}^{n} |\langle f, L_k \rangle|^2 + \|f - P_n f\|^2_{L^2(I)}$$

avec le fait que $\|f - P_n f\|_{L^2}$ tend vers 0. □

Un autre exemple de base orthonormée sera vu dans la section 2.6, avec les séries de Fourier.

Remarque 2.3.4 *On notera que pour des fonctions à valeurs complexes, le produit scalaire doit faire intervenir une valeur conjuguée,*

$$\langle f, g\rangle := \int_a^b f(x)\overline{g(x)}dx$$

afin que $\langle f, f\rangle$ soit bien un réel positif.

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 $\|L_n\|_{L^2(I)} = 1$ mais que pour tout $n \neq m$ on a $\|L_n - L_m\|_{L^2(I)} = \sqrt{2}$ 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 2.3.5 *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*

$$L_n(x) = \frac{(-1)^n}{2^n n!}\frac{d^n}{dx^n}\Big((1-x^2)^n\Big),$$

et la formule de récurrence

$$L_{n+1}(x) = \frac{2n+1}{n+1}xL_n(x) - \frac{n}{n+1}L_{n-1}(x),$$

initialisée par $L_0(x) = 1$ et $L_1(x) = x$.

Remarque 2.3.6 *En utilisant le changement de variable $x = \cos(t)$, on observe que les polynômes de Tchebychev définis par la formule*

$$T_n(x) = \cos(n \arccos(x)) \in \mathbb{P}_n$$

(ce sont bien des polynômes : voir la preuve du théorème de Weierstrass dans la section 4.2) 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 $w(x) = e^{-x}$).

2.4 Approximation 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 choisit 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 et à utiliser une approximation polynomiale de degré *fixé* sur chacun d'entre eux. On fait ensuite tendre la taille de ces morceaux vers 0.

Remarque 2.4.1 *Dans cette section la lettre n ne désignera plus le degré des approximations polynomiales mais le* **nombre de sous-intervalles** *dans la subdivision, car c'est lui qui sera le paramètre principal qu'on va faire tendre vers ∞. On utilisera la lettre M pour désigner l'ordre des approximations (la notion d'ordre coincidant parfois avec celle de degré, mais pas toujours) qui sera un paramètre fixe du procédé d'approximation. En particulier, les constantes apparaissant dans les estimations a priori pourront dépendre de M, mais pas de n.*

2.4.1 Interpolation polynomiale par morceaux

Une version simple de cette approche consiste à utiliser sur chaque sous-intervalle une interpolation de degré fixé M. Plus précisément, pour $n \geq 0$ on se donne une subdivision de l'intervalle $I = [a, b]$ en n cellules,

$$a = a_0 < a_1 < \cdots < a_{n-1} < a_n = b,$$

et on définit sa finesse par

$$h = \max_{i=0,\cdots,n-1} (a_{i+1} - a_i).$$

On peut alors appliquer un procédé d'interpolation de degré M fixé sur chacun des intervalles $[a_i, a_{i+1}]$. Pour cela on se donne $M + 1$ noeuds dans chaque cellule,

$$a_i = x_{i,0} < x_{i,1} < \cdots < x_{i,M-1} < x_{i,M} = a_{i+1},$$

et un choix classique consiste à prendre des points équidistants, c'est-à-dire $x_{i,j} = a_i + \frac{j}{M}(a_{i+1} - a_i)$ pour $j = 0, \cdots, M$.

Définition 2.4.1 *Soit f une fonction continue sur $I = [a, b]$. On définit son interpolation polynomiale de Lagrange de degré M par morceaux sur la subdivision $a_0, \cdots, a_n$, 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*

$$f_n(x_{i,j}) = f(x_{i,j}), \quad i = 0, \cdots, n-1, \quad j = 0, \cdots, M. \tag{2.36}$$

On pourra noter

$$\mathcal{I}_n^{(M)} : f \mapsto f_n$$

l'opérateur associé. Cet opérateur dépend évidemment de la position des a_i et des $x_{i,j}$. 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 observe 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

$$x_{i,M} = x_{i+1,0} = a_{i+1}.$$

Si l'on considère une subdivision uniforme de I en n intervalles de longueur $h = \frac{b-a}{n}$, i.e.,

$$\begin{cases} a_i = a + \frac{i}{n}(b-a), & i = 0, \cdots, n, \\ x_{i,j} = a_i + \frac{j}{M}(a_{i+1} - a_i), & i = 0, \cdots, n-1, \quad j = 0, \ldots, M, \end{cases} \tag{2.37}$$

on peut montrer le résultat suivant.

Théorème 2.4.1 *Si f est de classe $\mathcal{C}^{M+1}$ sur l'intervalle $I = [a, b]$, son interpolation par morceaux de degré M sur la subdivision uniforme* (2.37) *vérifie*

$$\|f - \mathcal{I}_n^{(M)} f\|_{L^\infty(I)} \leq C_M \frac{\|f\|_{\mathcal{C}^{M+1}(I)}}{n^{M+1}}, \tag{2.38}$$

avec C_M une constante dépendant de $|I| = b - a$ et de M, mais indépendante de f et n.

La preuve n'est pas difficile, nous la laissons en exercice : elle utilise (i) l'estimation a priori établie pour les développements de Taylor (proposition 2.1.1), (ii) la stabilité L^∞ des interpolations de degré M (proposition 2.2.1), et (iii) une propriété d'invariance par changement d'échelle.

Remarque 2.4.2 *On a vu plus haut qu'en interpolant une fonction régulière sur des noeuds équidistants, on pouvait obtenir des polynômes $\mathcal{I}_n f$ dont la norme L^∞ tendait vers l'infini avec n : c'est le phénomène de Runge. Il est important d'observer que l'estimation* (2.38) *empêche ce type de comportement, et il est remarquable que cette propriété soit vérifiée en utilisant des noeuds équidistants. Le point clef est ici que, contrairement au cas des interpolations $\mathcal{I}_n$, le degré des interpolations par morceaux est borné lorsqu'on fait tendre n vers l'infini.*

Remarque 2.4.3 *On peut également comparer l'estimation* (2.38) *ci-dessus avec celle du théorème 1.1.1 : quels sont les avantages respectifs de chaque estimation ?*

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 Γ_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 ℓ_γ constituent une base de l'espace vectoriel des fonctions polynomiales 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 Γ_n coïncide avec $\{a_0, \cdots, a_n\}$ et le graphe de la fonction de base ℓ_{a_i} a la forme d'un "chapeau" à support dans $[a_{i-1}, a_{i+1}]$.

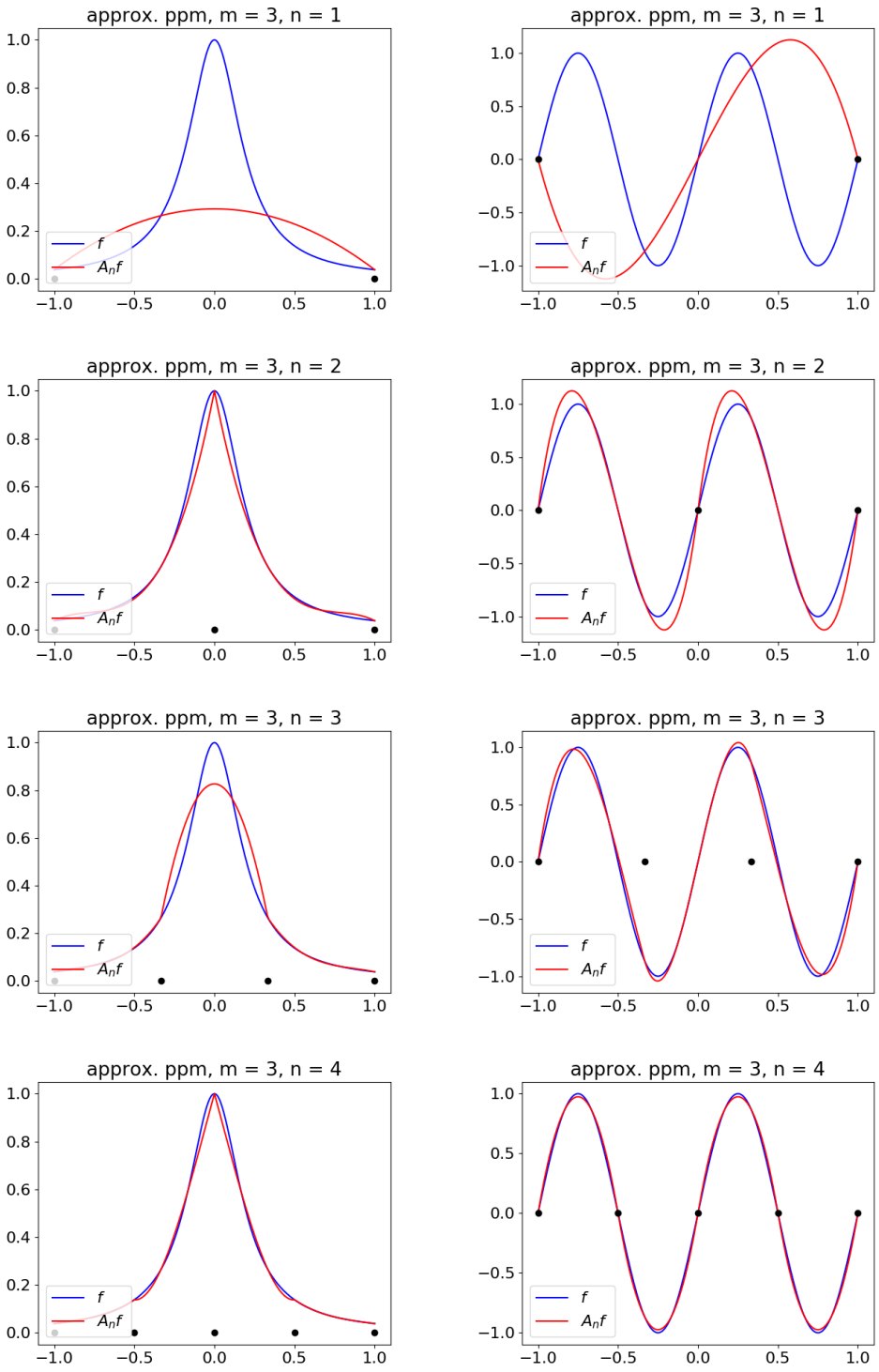

FIGURE 12 – Approximation des fonctions f_{Runge} (à gauche) et $f_{\sin}$ (à droite) par des approximations polynomiales par morceaux sur n sous-intervalles uniformes de $[-1, 1]$, avec $n = 1$, 2, 3 et 4. Sur chaque sous-intervalle (dont les bornes sont représentées par les points noirs), l'approximation est ici une interpolation de degré $M = 3$ avec des noeuds équidistants. Par comparaison avec la figure 11, on vérifie que le phénomène de Runge a disparu bien que les noeuds (non représentés ici) soient équidistants.

2.4.2 Approximation par des splines

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 2.2.2. 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

$$f_n^{(k)}(a_i) = f^{(k)}(a_i), \quad i = 0, \cdots, n, \quad k = 0, \cdots, M$$

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

$$\|f - f_n\|_{L^\infty} \leq C_M \|f^{(2M+2)}\|_{L^\infty} h^{2M+2},$$

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 donnée une subdivision $a = a_0 < \cdots < a_n = b$, 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

$$x \mapsto (x - a_i)_+^M = \Big(\max\{0, (x - a_i)\}\Big)^M, \quad i = 1, \cdots, n - 1,$$

ainsi que des fonctions $x \mapsto x^k$ pour $k = 0, \cdots, M$. L'ensemble de ces fonctions constitue une base de l'espace des splines d'ordre M sur $[a, b]$ pour la subdivision $a_1, \cdots, a_n$, 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 autours des points a_i : pour $i = 1, \cdots, n + M$, 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 $a_1, \cdots, a_n$. 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 2.4.2 *Si M est impair, pour tout ensemble de réels $\{y_0, \cdots, y_n\}$ avec $n \geq M$ et $\{\alpha_1, \cdots, \alpha_{M-1}\}$, il existe une unique fonction spline f_n d'ordre M sur $[a, b]$ pour la subdivision $a_0, \cdots, a_n$ telle que*

$$f_n(a_i) = y_i, \quad i = 0, \cdots, n \ \text{ et } \ f_n^{(k)}(a) = \alpha_k, \quad k = 1, \cdots, M - 1.$$

En particulier si f est une fonction continue, il existe une unique spline d'interpolation f_n d'ordre M définie par

$$f_n(a_i) = f(a_i), \;\; i = 0, \cdots, n \;\text{ et }\; f_n^{(k)}(a) = \alpha_k, \;\; k = 1, \cdots, M-1.$$

Une variante de ce résultat affirme l'existence et l'unicité d'une spline d'interpolation périodique.

Théorème 2.4.3 *Si M est impair, pour tout ensemble de réels $\{y_0, \cdots, y_{n-1}\}$ avec $n \geq M$, il existe une unique fonction spline f_n d'ordre M sur $[a, b]$ pour la subdivision $a_0, \cdots, a_n$ telle que*

$$f_n(a_i) = y_i, \;\; i = 1, \cdots, n \;\text{ et }\; f_n^{(k)}(a) = f_n^{(k)}(b), \;\; k = 0, \cdots, M-1.$$

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

$$f_n(a_i) = f(a_i), \;\; i = 0, \cdots, n \;\text{ et }\; f_n^{(k)}(a) = f_n^{(k)}(b), \;\; k = 1, \cdots, M-1.$$

2.5 Approximation positive par des polynômes de Bernstein

On présente maintenant une méthode d'approximation polynomiale alternative qui qui a la propriété de préserver la positivité des fonctions. En dehors de l'intérêt intrinsèque que cela peut avoir, on note qu'une telle méthode aura a priori moins d'oscillations que l'interpolation de Lagrange sur des noeuds équidistants, puisque dans certains cas (cf. le degré 9 sur la figure 11) ces oscillations ont lieu en dessous de 0 pour une fonction positive. Un autre intérêt de cette méthode sera de fournir une preuve relativement simple du théorème de Weierstrass. En revanche, on verra qu'elle méthode ne converge pas très vite (en particulier, elle ne préserve pas les polynômes de degré ≥ 2). Pour simplifier les notations on se ramène au cas de l'intervalle $I = [0, 1]$.

L'idée de départ est d'utiliser des polynômes positifs sur $[0, 1]$, à savoir

$$x \mapsto x^k(1-x)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

On peut vérifier que ces polynômes forment une base de $\mathbb{P}_n$, qu'on appelle *base de Bernstein*. (Les monômes $x \mapsto x^k$ forment également une base positive, mais ils ont l'inconvénient de ne pas être symétriques sur l'intervalle.) On se propose ensuite d'approcher une fonction f à partir de ses valeurs ponctuelles sur les noeuds réguliers $\frac{k}{n}$, $k = 0, 1, \ldots, n$, par un polynôme de la forme

$$f_n = \sum_{k=0}^{n} c_k f\Big(\frac{k}{n}\Big) x^k (1-x)^{n-k}$$

avec des coefficients c_k à déterminer. On observe alors que cette approximation préserve les constantes si et seulement si $1 = \sum_{k=0}^{n} c_k x^k (1-x)^{n-k}$, de sorte que les c_k doivent être les coefficients du polynôme constant 1 dans la base de Bernstein. Ces coefficients sont donnés par la formule du binôme de Newton, leur valeur est $c_k = \binom{n}{k} := \frac{n!}{k!(n-k)!}$.

Définition 2.5.1 *Soit $f \in \mathcal{C}^0([0,1])$ et $n \in \mathbb{N}$. Son n-ème* **polynôme de Bernstein** *est défini par*

$$\mathcal{B}_n f(x) := \sum_{k=0}^{n} \binom{n}{k} f\Big(\frac{k}{n}\Big) x^k (1-x)^{n-k},$$

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

La figure 13 trace les polynômes de Bernstein des fonctions $f(x) = \sin(2\pi x)$ et $f(x) = |x|$ sur $I = [-1, 1]$ (exercice : écrire le n-ième polynôme de Bernstein sur cet intervalle), pour $n = 5$, 10, 20 et 40. En comparant ces figures avec celles obtenues avec l'interpolation de Lagrange sur des noeuds équidistants, on observe que les oscillations sont complètement supprimées, en revanche la convergence est beaucoup plus lente. On remarquera également qu'elle ne semble pas beaucoup dépendre de la régularité de f.

On liste quelques propriétés de ces approximations dans la proposition suivante, dont la preuve est laissée en exercice.

Proposition 2.5.1 *L'approximation de Bernstein possède les propriétés suivantes :*

- *si f est positive sur $[0,1]$, alors $\mathcal{B}_n f$ l'est également ;*
- *les opérateurs*

$$\mathcal{B}_n : \mathcal{C}^0([0,1]) \to \mathbb{P}_n$$

sont uniformément stables pour la norme L^∞, et plus précisément ils vérifient

$$\min_{y\in[0,1]} f(y) \le \mathcal{B}_n f(x) \le \max_{y\in[0,1]} f(y), \qquad \forall x \in [0,1]; \tag{2.39}$$

- *si $n \ge 1$, $\mathcal{B}_n$ préserve $\mathbb{P}_1$ au sens où $\mathcal{B}_n f = f$ pour $f \in \mathbb{P}_1$;*
- *pour tout $n \in \mathbb{N}$, il existe $f \in \mathbb{P}_2$ tel que $\mathcal{B}_n f \neq f$.*

Remarque 2.5.1 *La propriété* (2.39)*, qui exprime le fait que f est transformée en une fonction dont les valeurs restent comprises entre les extrema (minimum et maximum) de f (ce qui implique que la norme L^∞ ne peut pas croître lors de cette transformation) est parfois appelée "principe du maximum".*

Le résultat suivant fournit une preuve constructive du théorème de Weierstrass. Sa preuve n'est pas très difficile mais on la reporte à la section 4.2.3 car elle met en oeuvre des techniques proches de celles qu'on utilisera pour étudier la convergence des sommes de Fejer, qui seront étudiées dans la section 4.1.

Théorème 2.5.1 *Si f est continue sur $[0,1]$, on a*

$$\lim_{n\to\infty} \|f - \mathcal{B}_n f\|_{L^\infty([0,1])} = 0.$$

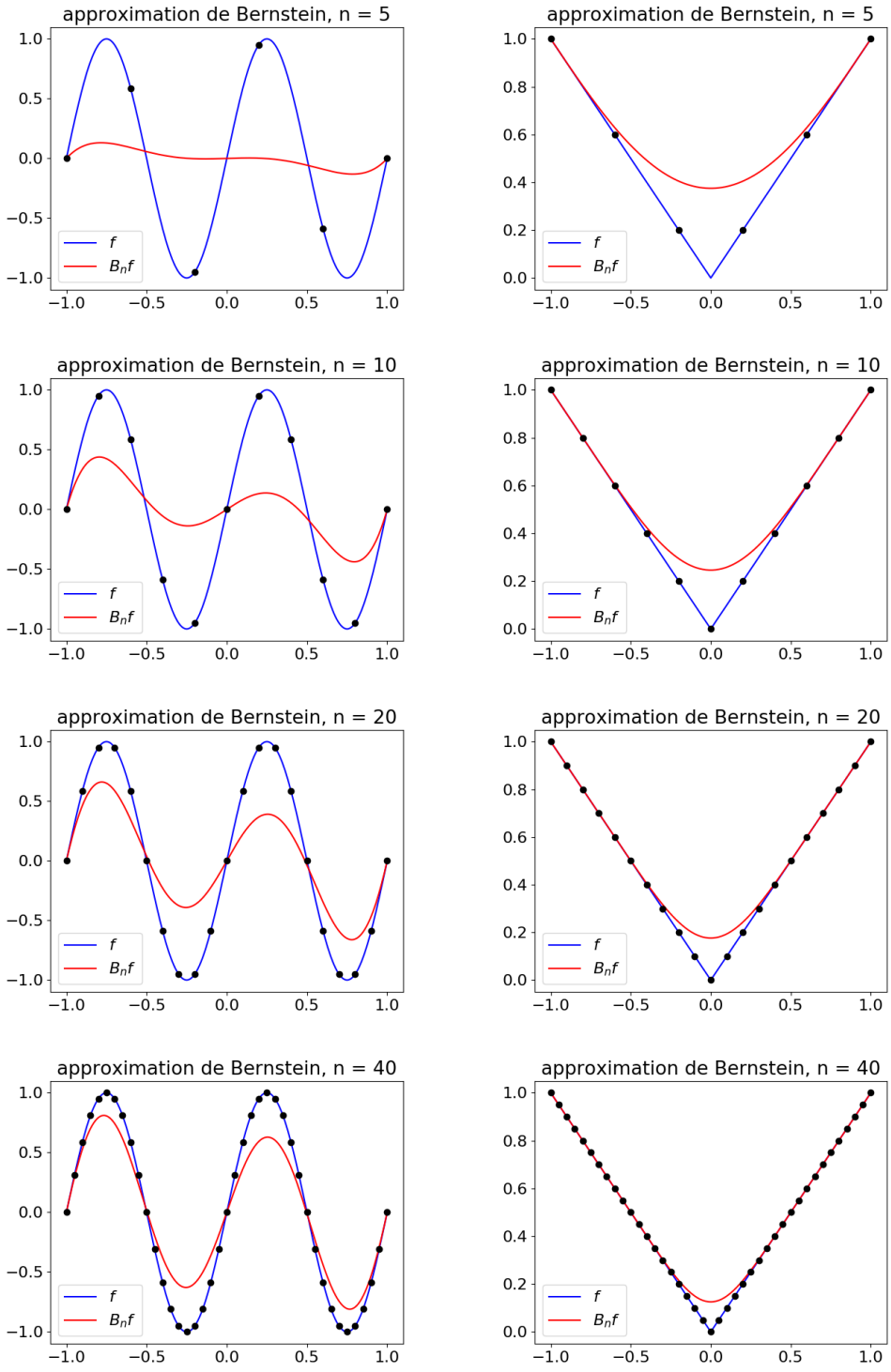

FIGURE 13 – Approximation des fonctions $f = f_{\sin}$ (à gauche) et $f = f_{\text{abs}}$ (à droite) par leurs polynômes de Bernstein $\mathcal{B}_n f$ sur l'intervalle $[-1, 1]$, de degrés $n = 5, 10, 20$ et 40. Les noeuds représentent les valeurs ponctuelles de f utilisés pour l'approximation.

2.6 Approximation par des séries de Fourier

Commencons par quelques rappels sur les séries trigonométriques, qui sont aussi appelées 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 $c_{\pm k} = \frac{1}{2}(a_k \mp i b_k)$ pour $k > 0$. Lorsque ces séries convergent, leurs limites sont des fonctions de période 2π, puisque chacun de leurs termes possède 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 des polynômes trigonométriques de degré n est donc l'espace vectoriel engendré par les exponentielles complexes de période $2\pi/k$ avec $|k| \leq n$,

$$\mathbb{T}_n = \mathrm{Vect}\{e_k : x \mapsto e^{ikx} \ ; \ |k| \leq n\}.$$

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

$$0 = \int_{-\pi}^{\pi} \Big(\sum_{|k|\leq n} c_k e_k(x) \Big) \overline{e_l(x)}dx = 2\pi c_l.$$

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

2.6.1 Sommes partielles de Fourier

Le problème fondamental de la représentation d'une fonction arbitraire 2π-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

$$f(x) = \sum_{k\in\mathbb{Z}} c_k e_k(x),$$

signifie que l'on cherche à approcher la fonction f par la suite de polynômes trigonométriques $\sum_{|k|\leq n} c_k e_k \in \mathbb{T}_n$. Si nous supposons qu'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_k dépend de f suivant la formule

$$c_k(f) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-ikx} dx. \tag{2.40}$$

Définition 2.6.1 *Les coefficients* (2.40) *sont souvent appelés* **coefficients de Fourier** *de f. Le polynôme trigonométrique*

$$\mathcal{S}_n f(x) = \sum_{|k|\leq n} c_k(f) e_k(x)$$

est appelé **somme partielle de Fourier** *de f de degré n.*

Nous rappelons une version simple du *théorème de Dirichlet* qui donne une condition suffisante pour la convergence simple de $\mathcal{S}_n f$ vers f.

Théorème 2.6.1 (de Dirichlet) *Soit f une fonction 2π périodique et continue, telle qu'en un point x il existe une dérivée à gauche et à droite. Alors*

$$\lim_{n\to+\infty} \mathcal{S}_n f(x) = f(x).$$

La figure 14 trace plusieurs sommes partielles de Fourier pour les fonctions f_{Runge} et f_{abs} définies par (1.12) et (1.13) sur $I = [-1, 1]$, complétées par périodicité (de période 2) sur $\mathbb{R}$. (Exercice : comment s'écrivent les approximations de Fourier lorsqu'on considère une période différente de 2π ?). On pourra comparer ces figures avec celles obtenues avec l'interpolation de Lagrange sur des noeuds équidistants et de Tchebychev. On a choisi ici de ne pas représenter l'approximation de la fonction $f_{\sin}$ par ses sommes partielles de Fourier : qu'aurait-on obtenu ?

Du point de vue numérique le théorème de Dirichlet n'est pas satisfaisant car il ne donne aucune estimation sur la façon dont l'erreur $\|f - \mathcal{S}_n f\|_{L^\infty}$ décroit en fonction de n (ici $\|\cdot\|_{L^\infty}$ désigne la norme sup sur $\mathbb{R}$, qui coïncide avec celle sur $[-\pi, \pi]$ puisque l'on considère des fonctions 2π-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 2π périodique de classe $\mathcal{C}^1$ sur $\mathbb{R}$, une intégration par partie permet d'obtenir

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

et par conséquent

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

En itérant l'intégration par partie si f est suffisamment régulière, on trouve

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

et par conséquent (si $f \in \mathcal{C}^{m+1}$),

$$|c_k(f)| \leq \frac{1}{|k|^{m+1}} \|f^{(m+1)}\|_{L^\infty}. \tag{2.41}$$

Comme le théorème de Dirichlet nous permet d'écrire $f(x) = \sum_{k\in\mathbb{Z}} c_k(f)e_k(x)$ pour tout x, l'erreur vaut

$$(f - \mathcal{S}_n f)(x) = \sum_{|k|>n} c_k(f) e_k(x)$$

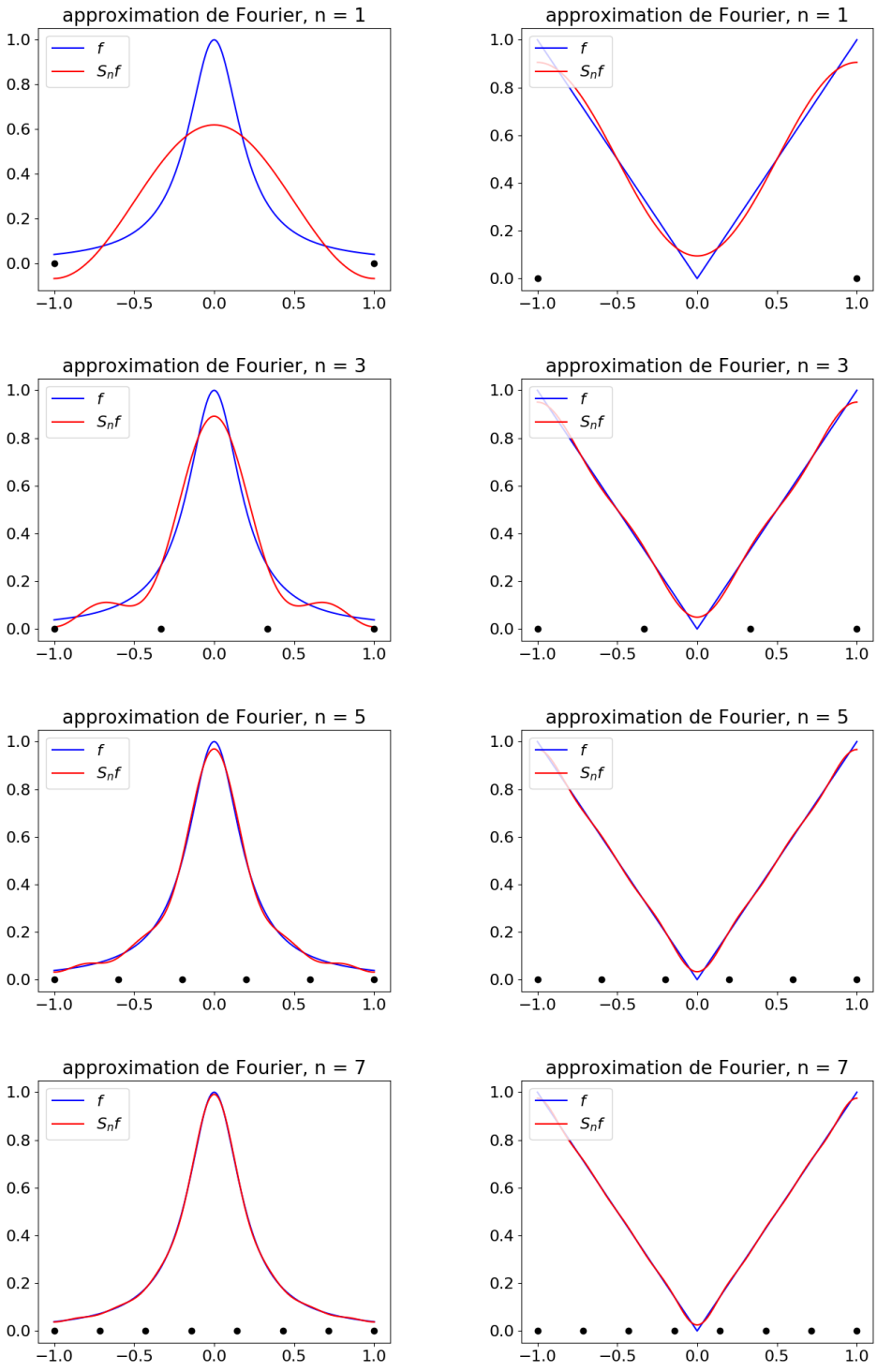

FIGURE 14 – Approximation des fonctions $f = f_{\text{Runge}}$ (à gauche) et $f = f_{\text{abs}}$ (à droite) par leurs sommes partielles de Fourier de degrés $n = 3, 5, 9$ et 15 sur l'intervalle $[-1, 1]$. Les noeuds en abscisse représentent la plus petite période des exponentielles complexes utilisées pour l'approximation. La convergence en tout point est observée, conformément au théorème de Dirichlet.

et l'estimation (2.41) sur les coefficients individuels peut nous donner une estimation sur la décroissance de l'erreur. En effet, pour $m \geq 1$ on peut calculer que

$$|f - \mathcal{S}_n f(x)| \leq \sum_{|k|>n} |c_k(f)| \leq \Big(\sum_{|k|>n} |k|^{-(m+1)} \Big) \|f^{(m+1)}\|_{L^\infty} \leq \frac{2n^{-m}}{m} \|f^{(m+1)}\|_{L^\infty},$$

où l'on a utilisé l'estimation

$$\sum_{|k|>n} |k|^{-(m+1)} = 2 \sum_{k>n} k^{-(m+1)} \leq 2 \int_n^{+\infty} t^{-(m+1)} dt = \frac{2n^{-m}}{m}.$$

On a donc prouvé le résultat suivant.

Théorème 2.6.2 (estimation a priori pour les sommes de Fourier) *Soit f une fonction 2π périodique de classe $\mathcal{C}^{m+1}$ avec $m \geq 1$. On a l'estimation*

$$\|f - \mathcal{S}_n f\|_{L^\infty} \leq C_m(f) \frac{1}{n^m},$$

avec $C_m(f) = \frac{2\|f^{(m+1)}\|_{L^\infty}}{m}$.

Remarque 2.6.1 *En adaptant la preuve, l'estimation ci-dessus reste valide pour des fonctions de classe $\mathcal{C}^m$ telle que $f^{(m+1)}$ est intégrable sur $[-\pi, \pi]$. La valeur de la constante est alors $C_m(f) = \frac{\int_{-\pi}^{\pi} |f^{(m+1)}(x)| dx}{m\pi}$.*

2.6.2 Sommes de Fejer

Lorsque f est seulement supposée continue sur $\mathbb{R}$, le théorème de Dirichlet ne permet pas d'affirmer que $\mathcal{S}_n f$ 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 2π-périodiques telles qu'en certains points $x \in \mathbb{R}$ la série de Fourier $\mathcal{S}_n f(x)$ *diverge* quand $n \to +\infty$.

Il est cependant possible d'approcher les fonctions continues par des polynômes trigonométriques qui diffèrent de $\mathcal{S}_n f$. Une approche consiste à utiliser les *moyennes de Césaro* des sommes partielles de Fourier, ce qui nous ammène à la définition suivante.

Définition 2.6.2 *Soit f une fonction continue sur $[-\pi, \pi]$. Le polynôme trigonométrique*

$$\mathcal{F}_n f = \frac{1}{n+1} \sum_{k=0}^{n} \mathcal{S}_k f \in \mathbb{T}_n \tag{2.42}$$

est appelé **somme de Fejer** *de f de degré n.*

Contrairement aux sommes de Fourier, ces approximations convergent vers toute fonction continue et périodique.

Théorème 2.6.3 *Pour toute fonction f continue et 2π-périodique, on a*

$$\lim_{n \to 0} \|f - \mathcal{F}_n f\|_{L^\infty} = 0.$$

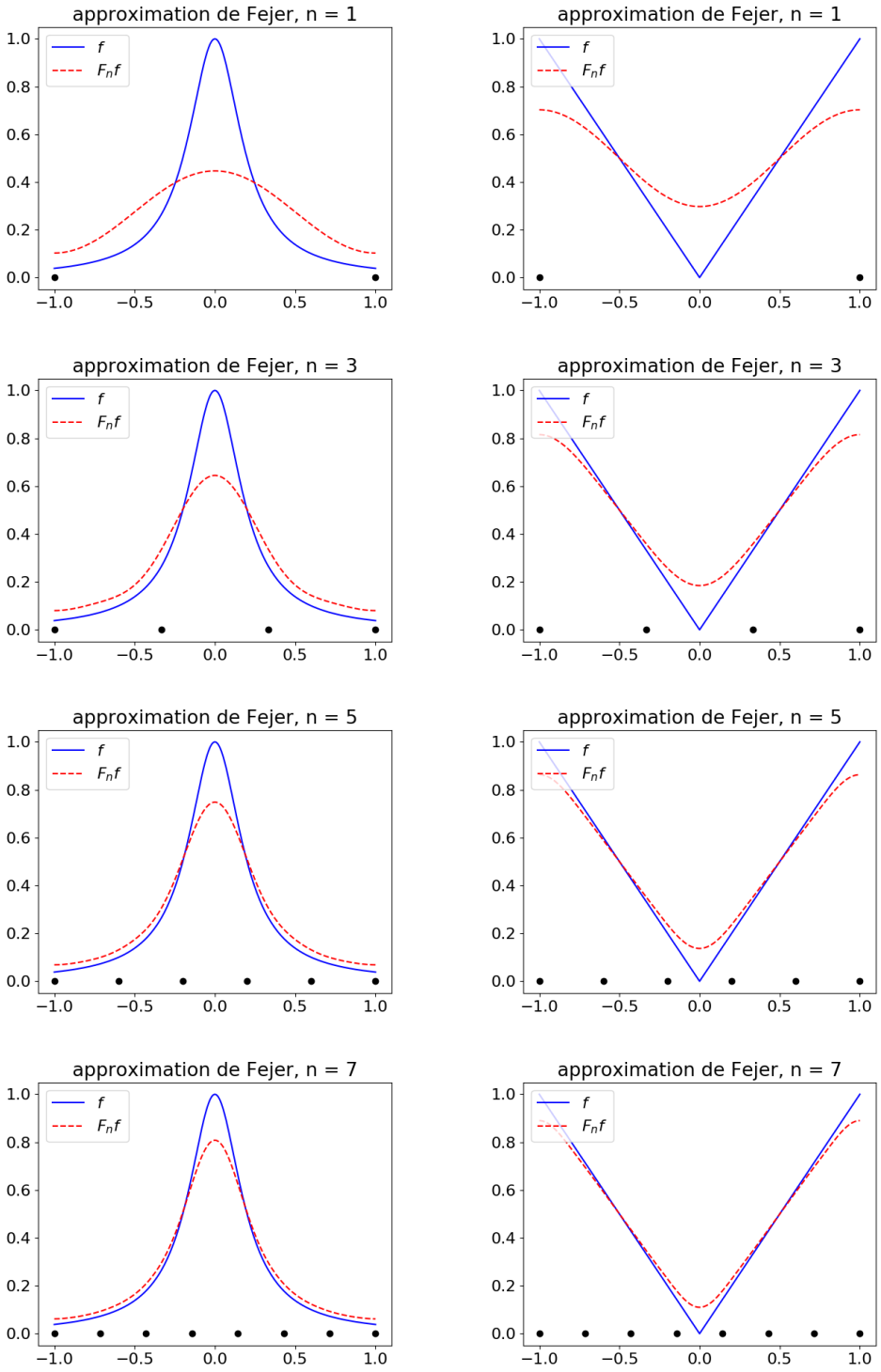

FIGURE 15 – Approximation des fonctions $f = f_{\text{Runge}}$ (à gauche) et $f = f_{\text{abs}}$ (à droite) par leurs sommes de Fejer de degrés $n = 3, 5, 9$ et 15 sur l'intervalle $[-1, 1]$. Les noeuds en abscisse représentent la plus petite période des exponentielles complexes utilisées pour l'approximation. On pourra comparer ces courbes avec celles de la figure 14.

La figure 15 trace plusieurs sommes de Fejer des fonctions f_{Runge} et f_{abs} définies par (1.12) et (1.13) sur $I = [-1, 1]$, et complétées par périodicité sur $\mathbb{R}$ pour considérer leurs approximations de Fejer avec une période 2. Par comparaison avec les sommes partielles de Fourier tracées figure 14 on observe une convergence plus lente. L'intérêt des sommes de Fejer réside dans le fait qu'elles convergent vers f lorsque celle-ci n'est que continue, et dans le fait qu'elles vont nous fournir un prototype d'*approximation à noyau*. Ce type d'approximation sera utilisé dans le chapitre 4 pour analyser certaines propriétés de convergence des polynômes trigonométriques vers des fonctions régulières, qui nous serviront notamment à démontrer le théorème 1.1.1 énoncé dans l'introduction sur les meilleures approximations polynomiales.

3 Approximation de solutions d'équations

On s'intéresse dans cette section à la résolution approchée d'équations de la forme

$$f(x) = 0 \tag{3.43}$$

où f est une fonction continue de $\mathbb{R}^n$ dans $\mathbb{R}^n$. Lorsque $n = 1$, il s'agit d'une équation **scalaire** et lorsque $n \geq 2$ il s'agit d'un **système** de n équations à n inconnues $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$. Lorsque f est de la forme $Ax - b$ avec A une matrice carrée de taille n et $b \in \mathbb{R}^n$ l'équation (ou le système) est dit **linéaire**, sinon il est **non-linéaire**. La résolution exacte de telles équations est aisée 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 (3.43). Ces méthodes sont souvent 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.

Remarque 3.0.1 *Dans toute cette section, la lettre n désigne la dimension de l'espace sur lequel est définie la fonction f. On prendra donc soin de ne pas confondre n avec un indice de suite.*

Remarque 3.0.2 *Dans le cas des systèmes linéaires, la résolution exacte de* (3.43) *est possible en théorie (lorsque A est inversible) mais souvent prohibitive lorsque n est très grand. En pratique, on a donc recours à des méthodes itératives pour calculer des solutions approchées de façon moins coûteuse.*

3.1 La méthode de dichotomie pour des équations scalaires

Une première approche assez inuitive pour résoudre des problèmes scalaires est celle de la recherche par dichotomie, basée sur l'observation suivante : si f est une fonction continue sur un intervalle $I = [a, b]$ et que $f(a)f(b) < 0$, alors d'après le théorème des valeurs intermédiaires, il existe au moins un point $x^* \in]a, b[$ tel que $f(x^*) = 0$. La méthode de dichotomie prend la forme suivante.

Algorithme 3.1.1 (dichotomie) *On considère $f \in \mathcal{C}^0([a, b])$ telle que $f(a)f(b) < 0$.*

- **Initialisation** *: on pose $a_0 := a$ et $b_0 := b$.*
- **Itération** *: pour $n \geq 0$, on pose $c_n := \frac{a_n + b_n}{2}$ et trois cas se présentent :*
  - *si $f(a_n)f(c_n) < 0$, alors on pose $a_{n+1} := a_n$, $b_{n+1} := c_n$*
  - *si $f(a_n)f(c_n) > 0$, alors on pose $a_{n+1} := c_n$, $b_{n+1} := b_n$*
  - *si $f(a_n)f(c_n) = 0$, alors on s'arrête, et on prend $x^* := c_n$.*

Les propriétés de cet algorithme sont faciles à étudier, on peut les résumer par le résultat suivant. (Attention toutefois : si la fonction f possède plusieurs zéros dans l'intervalle I et qu'on est intéressé par l'un d'entre eux, par exemple le premier, il n'y a aucune raison a priori pour que l'algorithme converge vers celui-ci.)

Théorème 3.1.1 *L'algorithme de dichotomie converge vers un $x^* \in I$ tel que $f(x^*) = 0$. De plus, si l'algorithme ne s'arrête pas en un nombre fini d'itérations, on a*

$$|c_n - x^*| \leq \frac{b-a}{2^{n+1}}, \qquad n \in \mathbb{N}.$$

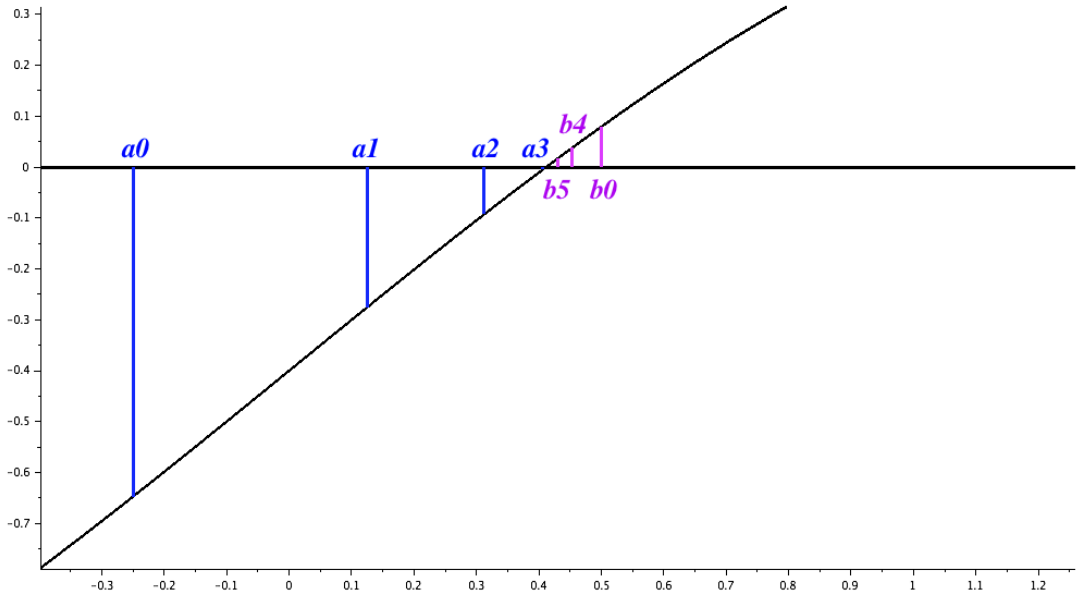

FIGURE 16 – Illustration de la méthode de dichotomie pour la fonction $f(x) = \sin(x) - 0.4$. Les points (a_n en bleu, b_n en mauve) sont indiqués à l'étape où ils apparaissent dans l'algorithme.

3.2 La méthode du point fixe et le théorème de Picard

Un inconvénient de la méthode de dichotomie est qu'il n'est pas immédiat de l'étendre à des fonctions de plusieurs variables. Une approche plus générale est celle de la méthode du point fixe, qui s'écrit de façon très simple dans $\mathbb{R}^n$ avec n quelconque.

Algorithme 3.2.1 (point fixe) *On considère une fonction g continue, de $\mathbb{R}^n$ dans $\mathbb{R}^n$.*
- **Initialisation** *: on choisit $x^0 \in \mathbb{R}^n$.*
- **Itération** *: pour $k \geq 0$, on pose*

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

Le théorème de Picard qu'on va bientôt énoncer est un résultat général qui donne des conditions suffisantes pour que l'algorithme du point fixe converge. Il repose sur une propriété importante : la contraction des fonctions.

Définition 3.2.1 *Soit E un espace vectoriel muni d'une norme $\|\cdot\|$, et soit F un sous-ensemble de E. Une fonction $g : E \to E$ définie sur F est dite* **contractante** *sur F si et seulement si il existe une constante $a < 1$ 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 3.2.1 (du point fixe de Picard) *Soit E un espace vectoriel (sur $\mathbb{R}$ ou $\mathbb{C}$) 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 $g(x^*) = x^*$,*
- *pour tout $x^0 \in F$, l'algorithme du point fixe 3.2.1 converge vers x^*.*

Preuve. On va démontrer que la suite x^k converge, puis que sa limite est un point fixe dans F, et enfin que ce point fixe est unique. Comme $g(F) \subset F$, on observe que tous les x^k sont dans F. La propriété de contraction de g sur F entraîne alors

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

et par récurrence on obtient donc

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

Comme $a < 1$ ceci montre que la série $\sum_{j=0}^{k-1}\|x^{j+1} - x^j\|$ converge, ce qui entraine que la suite $\sum_{j=0}^{k-1}(x^{j+1} - x^j)$, et donc x^k, est de Cauchy. Comme E est un espace de dimension finie sur un corps complet il est complet, d'où l'on déduit 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

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

donc la convergence de la suite vers un point fixe est démontrée. Enfin 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 entraine $x^* = y^*$ puisque $a < 1$, d'où l'unicité du point fixe. □

Remarque 3.2.1 *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 3.2.2 *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 $g : \mathbb{R}^2 \to \mathbb{R}^2$ 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 ℓ^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 3.2.3 *Le théorème de Picard est "constructif" au sens où sa preuve établit également la convergence de la 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 $\mathbb{R}^2$.*

3.3 Etude de la méthode du point fixe dans le cas scalaire

Dans cette section comme dans la suivante, on suppose que $n = 1$, i.e. g est une fonction à variables et à valeurs réelles. Pour examiner plus en détail le comportement de l'algorithme du point fixe, nous allons considérer le cas où g est de classe $\mathcal{C}^1$. La proposition suivante est une conséquence immédiate du théorème des accroissement finis.

Proposition 3.3.1 *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 un intervalle ouvert I et $x^* \in I$ un point fixe de g (pas nécessairement unique dans I), avec $g \in \mathcal{C}^1(I)$. 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 3.2.1 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** :

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

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 par le théorème des accroissements finis

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

ce qui montre que la suite x^n tend à s'éloigner de x^* lorsqu'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 $g(x) = \mathrm{sh}(x) = (e^x - e^{-x})/2$ 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 $|\mathrm{sh}(x)| > |x|$ 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 $x^0 \in F$,

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

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**.

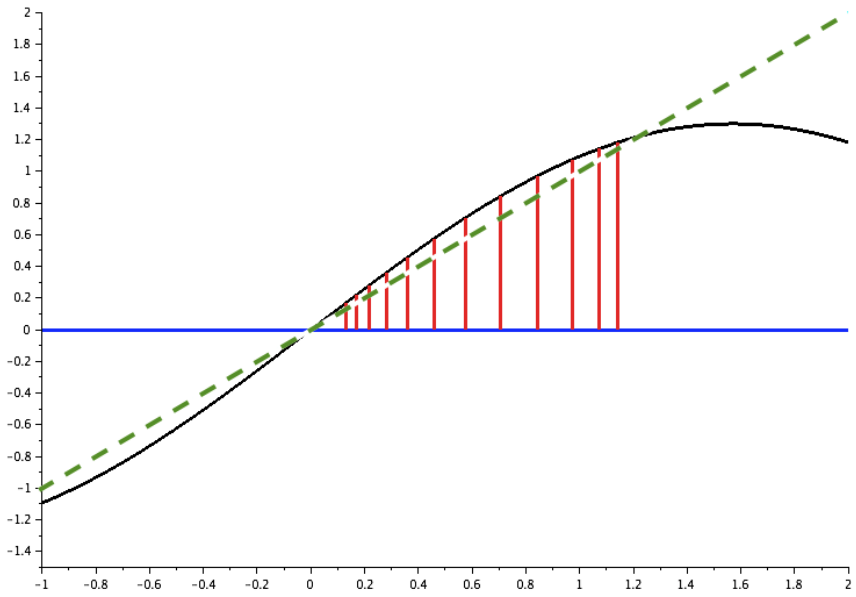

FIGURE 17 – Illustration de la méthode du point fixe pour la fonction $f(x) = 1.3\sin(x)$ dont le graphe est représenté en noir. La ligne verte en pointillés représente la diagonale $y = x$, et chaque segment rouge indique la position d'une itération x^n. En reportant à la main les valeurs des itérations successives, on observe qu'un des points fixes est répulsif et l'autre attractif, conformément à la discussion menée dans la section 3.3.

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 au problème 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 particuliers 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 τ 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 τ 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 τ. 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 connait 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.

3.4 La méthode de Newton pour des équations scalaires

La méthode de Newton est une approche systématique pour approcher rapidement les solutions d'une équation de la forme

$$f(x) = 0$$

dans le cas où f est dérivable (on considère ici le cas scalaire, i.e. $n = 1$). 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 $y = x - \frac{f(x)}{f'(x)}$. Cette idée conduit à l'algorithme suivant.

Algorithme 3.4.1 (de Newton) *On considère une fonction f dérivable.*

— **Initialisation** *: on choisit x^0 tel que $f'(x^0) \neq 0$.*

— **Itération** *: pour $k \geq 0$, on pose*

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

Cette méthode itérative peut donc s'interprèter 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 3.4.1 *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 $x \in I$

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

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^k converge donc quadratiquement vers x^* si x^0 en est suffisamment proche. □

Remarque 3.4.1 *La localité du résultat ci-dessus peut représenter un handicap important en pratique, surtout dans les cas où la position du zéro x^* n'est pas bien estimée. En particulier, on vérifiera sur un dessin que si la dérivée de f est proche de 0 en x^0, la méthode peut calculer des itérées qui s'éloignent arbitrairement de x^*. Le choix de l'initialisation a donc une importance cruciale sur le comportement de la méthode de Newton.*

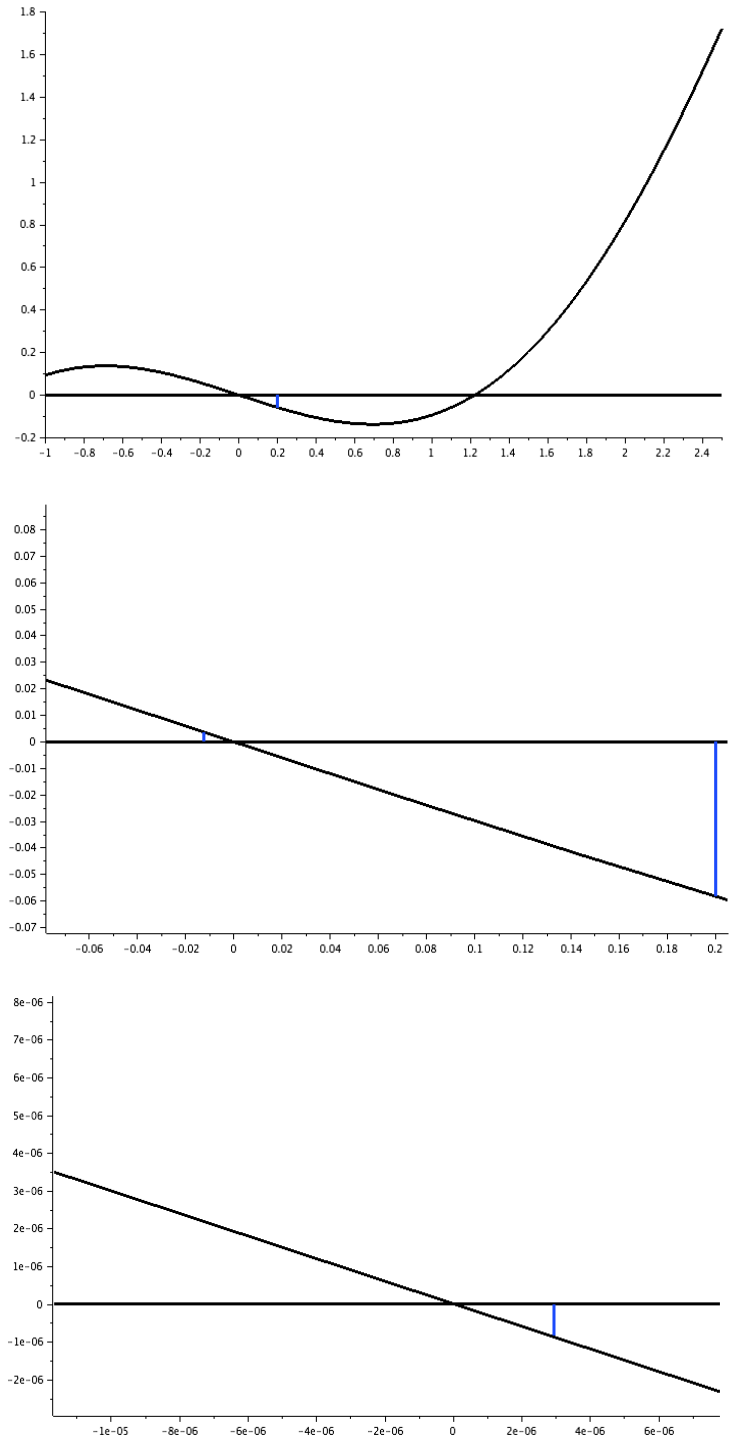

FIGURE 18 – Convergence de la méthode de Newton appliquée à la fonction $f(x) = x - 1.3\sin(x)$: on indique 3 itérations successives partant de $x^0 = 0.2$, avec des zooms. Noter la convergence quadratique visible sur les zooms successifs des axes.

3.5 Etude de la méthode du point fixe dans le cas vectoriel

Nous généralisons ici l'étude précédente au cas vectoriel, i.e. avec n quelconque. Les notions de calcul matriciel et les normes vectorielles utiles à la bonne compréhension de cette section sont rappelées plus bas.

On considère donc ici $g : \mathbb{R}^n \to \mathbb{R}^n$ dont les composantes $(g_1, \cdots, g_n)$ sont chacune des fonctions de $\mathbb{R}^n$ dans $\mathbb{R}$.

Pour étudier le comportement de l'algorithme du point fixe 3.2.1, on va supposer g possède au moins un point fixe x^* dans un ouvert $U \subset \mathbb{R}^n$, et est qu'elle est de classe $\mathcal{C}^1$ sur cet ouvert. On rappelle que g est de classe $\mathcal{C}^1$ sur un ouvert U 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 x est une application linéaire notée $dg_x \in \mathcal{L}(\mathbb{R}^n)$, dont la matrice dans la base canonique de $\mathbb{R}^n$ est

$$dg_x = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(x) & \cdots & \frac{\partial g_1}{\partial x_n}(x) \\ \vdots & & \vdots \\ \frac{\partial g_n}{\partial x_1}(x) & \cdots & \frac{\partial g_n}{\partial x_n}(x) \end{pmatrix}.$$

Cette matrice est appelée la *matrice Jacobienne* de g au point x, et on rappelle que le 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 ε est une fonction vectorielle telle que $\lim_{h\to 0} \varepsilon(h) = 0$. Le résultat suivant est une généralisation de la proposition 3.3.1.

Proposition 3.5.1 *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 valeurs dans $\mathbb{R}^n$. Si g est a-lipschitzienne sur U pour cette norme (c'est-à-dire, si $\|g(x) - g(y)\| \leq a\|x - y\|$ pour tout $x, y \in U$) alors on a $\|dg_x\| \leq a$ 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

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

Puisque $\|g(x + tv) - g(x)\| \leq a|t|\|v\|$, on en déduit $\|dg_x v\| \leq a\|v\|$ par définition de la norme subordonnée. 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

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

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

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

où on a utilisé l'inégalité $\| \int_a^b h(t)dt\| \leq \int_a^b \|h(t)\|dt$ 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. On suppose donc g de classe $\mathcal{C}^1$ sur un ouvert $U \subset \mathbb{R}^n$, à valeurs dans $\mathbb{R}^n$, et on suppose qu'elle possède au moins un point fixe $x^* \in U$. Comme dans le cas des fonctions réelles, on peut distinguer plusieurs cas selon les propriétés de dg_{x^*}. En s'inspirant de notre discussion dans le cas scalaire, on pourrait vouloir utiliser la norme de la matrice dg_{x^*} : d'après la proposition ci-dessus la fonction g sera contractante sur un voisinage de x^* et on pourra en déduire la convergence de la méthode. Le problème est que ce critère est ambigu, voir la remarque 3.5.1.

Pour avoir un critère moins ambigu, on va utiliser une quantité très importante en calcul numérique matriciel : le **rayon spectral** $\varrho(A)$ d'une matrice carrée A (cf. la définition 3.8.7 plus bas). La discussion ci-dessous va s'appuyer sur deux propriétés importantes du rayon spectral : il minore toute les normes matricielles subordonnées, et il est presque un majorant de certaines normes bien choisies. Pour des énoncés précis de ces propriétés, nous renvoyons aux propositions 3.8.5 et 3.8.6 dans la section 3.8, où ces propriétés seront également démontrées.

Cas 1. $\varrho(dg_{x^*}) < 1$. D'après la proposition 3.8.6, pour tout $\varepsilon > 0$ il existe une norme (matricielle) subordonnée à une norme $\|\cdot\|$ sur $\mathbb{C}^n$ telle que

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

et en choisissant ε 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 3.3.1, ceci entraine 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 3.5.1 *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 entraine $\varrho(dg_{x^*}) < 1$ d'après la proposition 3.8.5, mais elle n'est pas nécessaire : on peut avoir $\|dg_{x^*}\| > 1$ pour certaines normes subordonnées et néammoins $\varrho(dg_{x^*}) < 1$ 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 on peut même montrer qu'elle a tendance à s'éloigner de x^* si l'on démarre à proximité et dans une direction bien choisie. Le point fixe est dit **répulsif**. Notons cependant que la méthode du point fixe peut converger pour certains choix particulier de x^0. Par exemple si $g : \mathbb{R}^2 \to \mathbb{R}^2$ 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 $\varphi : \mathbb{R}^n \to \mathbb{R}$ 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

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

est la matrice des dérivées secondes ou *hessienne* au point z, et $\varepsilon : \mathbb{R}^n \to \mathbb{R}$ est telle que $\lim_{h\to 0} \varepsilon(h) = 0$. En particulier, si l'on fixe $z, h \in \mathbb{R}^n$, la fonction $\varphi_{z,h} : \mathbb{R} \to \mathbb{R}$ 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_l, avec $z = x^*$ et $h = x - x^*$ pour $x \in F$, on obtient (en utilisant l'hypothèse $\nabla g_l(x^*) = 0$)

$$g_l(x) - x_l^* = g_l(x) - g_l(x^*) = \frac{1}{2}\sum_{i,j=1,\cdots,n} \frac{\partial g_l}{\partial x_i \partial x_j}(x^* + s_l(x-x^*))(x_i - x_i^*)(x_j - x_j^*),$$

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

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

on obtient ainsi

$$|g_l(x) - x_l^*| \leq \frac{M_2}{2}\Big(\max_{i=1,\cdots,n} |x_i - x_i^*|\Big)^2,$$

pour tout $l = 1, \cdots, n$ 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 (voir le théorème 3.8.1) 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

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

$b := \frac{M_2\tilde{r}}{2}$ 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.

3.6 La méthode de Newton-Raphson pour des fonctions vectorielles

La méthode de Newton introduite dans la section 3.4 pour résoudre une équation scalaire peut aussi être généralisée au cas des fonctions vectorielles $f : \mathbb{R}^n \to \mathbb{R}^n$. Elle repose à nouveau sur l'idée de remplacer f au voisinage de x par la fonction "tangente"

$$\tilde{f}_x : y \mapsto f(x) + df_x(y - x),$$

qui s'annule au point $y = x - (df_x)^{-1}(f(x))$. Cherchant x^{k+1} tel que $\tilde{f}_{x^k}$ s'annule, on arrive à l'algorithme suivant.

Algorithme 3.6.1 (de Newton-Raphson) *On considère une fonction $f : \mathbb{R}^n \to \mathbb{R}^n$ de classe $\mathcal{C}^1$.*

— **Initialisation** *: on choisit $x^0 \in \mathbb{R}^n$ tel que la matrice df_{x^0} soit inversible.*

— **Itération** *: pour $k \geq 0$, on pose*

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

Cette méthode peut être vue comme une méthode de point fixe pour la fonction

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

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

Théorème 3.6.1 *Soit f une fonction de classe $\mathcal{C}^2$. 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

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

On rappelle le développement limité de Taylor-Lagrange pour une fonction $\varphi : \mathbb{R}^n \to \mathbb{R}$, 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_l de f avec $z = x$ et $h = x^* - x$, on obtient

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

où la l-ème coordonnée du reste δ est

$$\delta_l = \frac{1}{2}\sum_{i,j=1,\cdots,n} \frac{\partial^2 f_l}{\partial x_i \partial x_j}(x + s_l(x^* - x))(x_i - x_i^*)(x_j - x_j^*),$$

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

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

et donc

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

où

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

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^k converge donc quadratiquement vers x^* si x^0 en est suffisamment proche. □

3.7 La méthode de la sécante

Revenons en dimension 1. La méthode de Newton décrite dans la section 3.4 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 $f'(x^k)$ par le quotient $\frac{f(x^k)-f(x^{k-1})}{x^k - x^{k-1}}$. C'est la méthode de la *sécante* qui s'écrit

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

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 $f(x^k) \neq f(x^{k-1})$.

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 $F = [x^* - \delta, x^* + \delta]$. Prouvons que sous ces conditions, on a

$$|z - x^*| \leq K|y - x^*| \max\{|x - x^*|, |y - x^*|\}, \tag{3.44}$$

où l'on a posé $K = \frac{3M_2}{2a}$ avec $M_2 := \max_{t \in F} |f''(t)|$, et

$$z := y - \frac{f(y)(y-x)}{f(y) - f(x)} \qquad \text{pour} \quad \text{ă } x, y \in F \text{ tel que } x \neq y.$$

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

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

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 entraine l'estimation (3.44). Ce résultat entraine immédiatement que si δ 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^k et celle-ci est bien définie pour tout k (sauf si elle atteint x^* pour un k fini auquel cas il n'y a plus lieu de continuer l'algorithme). On a de plus

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

Par récurrence, on en déduit

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

où s_k est la suite de Fibonacci définie par $s_0 = s_1 = 1$ et $s_{k+1} = s_k + s_{k-1}$. La suite de Fibonacci est asymptotiquement proportionelle à c^k 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).

3.7.1 Analyse de la méthode de la sécante grâce à la théorie de la méthode de Newton-Raphson

Ceci est le texte d'un examen de 2010 pour vous entrainer.

On se donne une fonction $f : \mathbb{R} \to \mathbb{R}$ de classe $\mathcal{C}^2$. On suppose que $f'(x) > 0$ pour tout $x \in \mathbb{R}$ et que f possède un unique zéro noté $x^\star$: $f(x^\star) = 0$. Dans cet exercice nous

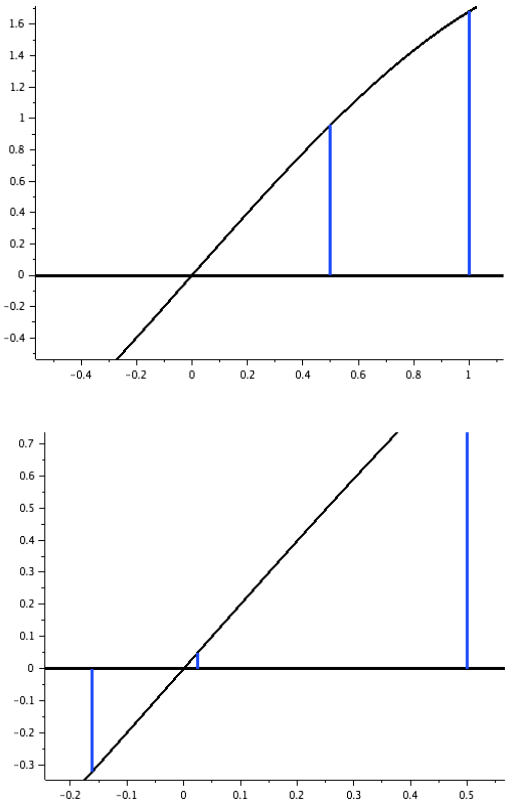

FIGURE 19 – Convergence de la méthode de la sécante : 3 itérés successifs avec zoom sur Noter que la première figure montre une initialisation à deux points, ce qui est une indication que la méthode de la sécante peut s'analyser comme une méthode de Newton-Raphson dans $\mathbb{R}^2$. Voir le sujet d'examen plus bas pour plus d'explications.

considérons la méthode de la sécante pour un calcul itératif de $x^\star$. Tout d'abord on se donne deux nombres réels, x^0 quelconque et $x^1 \neq x^0$. Ensuite on définit x^{n+1} par

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

Par récurrence cela construit la suite $n \mapsto x^n$. L'analyse de cette méthode fait l'objet des questions qui suivent.

1) Montrer que x^{n+1} est bien défini si $x^n \neq x^{n-1}$.

Rappeler pourquoi cette méthode peut se concevoir comme une modification de la méthode de Newton.

2) On suppose qu'il existe un indice $p \geq 2$ tel que $f(x^p) - f(x^{p-1}) = 0$, et tel que $f(x^q) - f(x^{q-1}) \neq 0$ pour tout $1 \leq q \leq p-1$. Montrer que $x^p = x^{p-1} = x^\star$, ce qui fait que la suite a convergé vers le point fixe avant qu'il soit besoin de calculer x^{p+1}.

Par la suite on supposera au contraire que $f(x^p) - f(x^{p-1}) \neq 0$ pour tout $p \geq 1$: par récurrence x^{n+1} est correctement défini pour tout n.

3) Soit $\varphi : \mathbb{R}^2 \to \mathbb{R}$ la fonction à valeur réelle définie par

$$\varphi(X_1, X_2) = \frac{f(X_1) - f(X_2)}{X_1 - X_2} \text{ pour } X_1 \neq X_2,$$

$$\varphi(X_1, X_2) = f'(X_1) \text{ pour } X_1 = X_2.$$

Soit $G : \mathbb{R}^2 \to \mathbb{R}^2$ la fonction à valeur vectorielle définie par

$$G(X) = \begin{pmatrix} X_1 - \dfrac{f(X_1)}{\varphi(X_1, X_2)} \\ \\ X_2 \end{pmatrix}, \quad X = (X_1, X_2).$$

Montrer que l'algorithme de la sécante se récrit $X^{n+1} = G(X^n)$ avec $X^n = (x^n, x^{n-1})$. Déterminer les points fixes de G.

4) Calculer $\frac{\partial}{\partial X_1}\varphi$ et $\frac{\partial}{\partial X_2}\varphi$ d'abord pour $X_1 \neq X_2$, et ensuite pour $X_1 = X_2$.
En déterminer $dG(X)$ pour $X_1 \neq X_2$, puis pour $X_1 = X_2 \neq x^\star$.

5) Montrer que $dG(X^\star) = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$. En déduire que la méthode de la sécante converge vers $x^\star$ pour x^0 et x^1 suffisamment proche de $x^\star$.
Quel est le taux de convergence ?

3.7.2 Proposition de correction

1) Pour $x^n \neq x^{n-1}$ alors $f(x^n) \neq f(x^{n-1})$ car $f > 0$ partout. Donc il n'y a pas de division par zéro.
Le terme $\frac{x^n - x^{n-1}}{f(x^n) - f(x^{n-1})}$ est une approximation d'ordre un en $|x^n - x^{n-1}|$ de $\frac{1}{f'(x^n)}$ quand $x^n - x^{n-1}$ est petit. En ce sens, la méthode proposée est une approximation de la méthode de Newton qui s'écrit $x^{n+1} = x^n - \frac{f(x^n)}{f'(x^n)}$.

2) Si $f(x^p) - f(x^{p-1}) = 0$ alors $x^p = x^{p-1}$ car un développement de Taylor au premier ordre montre que

$$0 = f(x^p) - f(x^{p-1}) = f'(c)(x^p - x^{p-1})$$

avec $f'(c) \neq 0$.

3) Les points fixes de G sont définis par $G(X^\star) = X^\star$, soit

$$X_1^\star = X_1^\star - \frac{f(X_1^\star)}{\varphi(X_1^\star, X_2^\star)} \text{ et } X_2^\star = X_1^\star.$$

Donc

$$X_1^\star = X_1^\star - \frac{f(X_1^\star)}{\varphi(X_1^\star, X_1^\star)} = X_1^\star - \frac{f(X_1^\star)}{f'(X_1^\star)}$$

donc $f(X_1^\star) = 0$. Donc $X_1^\star = x^\star$ l'unique point fixe de f. De même $X_2^\star = x^\star$.

4) Pour $X_1 \neq X_2$ on a

$$\partial_{X_1}\varphi = \frac{f'(X_1)}{X_1 - X_2} - \frac{f(X_1) - f(X_2)}{(X_1 - X_2)^2}$$
$$= \frac{f'(X_1)(X_1 - X_2) - f(X_1) + f(X_2)}{(X_1 - X_2)^2}$$

et

$$\partial_{X_2}\varphi = -\frac{f'(X_2)}{X_1 - X_2} + \frac{f(X_1) - f(X_2)}{(X_1 - X_2)^2}$$
$$= \frac{-f'(X_2)(X_1 - X_2) + f(X_1) - f(X_2)}{(X_1 - X_2)^2}.$$

Pour $X_1 = X_2$, la dérivée peut se calculer en passant à la limite dans les expressions ci-dessus. On a

$$f'(X_1)(X_1 - X_2) - f(X_1) + f(X_2)$$
$$= \frac{1}{2}f''(X_1)(X_1 - X_2)^2 + O((X_1 - X_2)^3).$$

Donc on obtient à la limite

$$\partial_{X_1}\varphi \to_{X_2 \to X_1} = \frac{1}{2}f''(X_1).$$

De même

$$\partial_{X_2}\varphi \to_{X_2 \to X_1} = -\frac{1}{2}f''(X_1).$$

D'autres modes de calcul sont possibles.
Pour $X_1 \neq X_2$, one a

$$dG(X) = \left(\begin{array}{c|c} 1 - \frac{f'(X_1)}{\varphi(X_1,X_2)} + \frac{f(X_1)\partial_{X_1}\varphi(X_1,X_2)}{\varphi(X_1,X_2)^2} & \frac{f(X_1)\partial_{X_2}\varphi(X_1,X_2)}{\varphi(X_1,X_2)^2} \\ \hline 1 & 0 \end{array}\right).$$

Pour $X_1 = X_2 \neq x^\star$ on obtient

$$dG(X) = \left(\begin{array}{c|c} 1 - \frac{f'(X_1)}{f'(X_1)} + \frac{f(X_1)\frac{1}{2}f'(X_1)}{f'(X_1)^2} & -\frac{f(X_1)\frac{1}{2}f'(X_1)}{f'(X_1)^2} \\ \hline 1 & 0 \end{array}\right)$$
$$= \left(\begin{array}{c|c} \frac{f(X_1)\frac{1}{2}f'(X_1)}{f'(X_1)^2} & -\frac{f(X_1)\frac{1}{2}f'(X_1)}{f'(X_1)^2} \\ \hline 1 & 0 \end{array}\right).$$

5) Enfin pour $X_1 = X_2 = x^\star$, on trouve

$$dG(X^\star) = \left(\begin{array}{c|c} 0 & 0 \\ \hline 1 & 0 \end{array}\right)$$

Les valeurs propres de $dG(X^\star)$ sont nulles car les polynome caractéristique est

$$p_{X^\star} = det\,(dG(X^\star) - \lambda I) = \lambda^2.$$

Il s'ensuit que $\rho(dG(X^\star)) = 0$. La fonction G étant $\mathcal{C}^2$, il s'ensuit que par un théorème du cours, X^n tend vers $X^\star$ pour X^0 choisit suffisamment proche de $X^\star$.

6) En reprenant un théorème du cours, on trouve l'extimation

$$||X^n - X^\star||_\infty \leq \frac{2}{M_2} b^{2^k}.$$

La constante b est $0 \leq b < 1$. La constante M_2 est > 0. On parle d'une convergence quadratique.

3.8 Résultats de calcul matriciel

On présente ici en détail les résultats d'algèbre linéaire qui sont utilisés dans la discussion de la section 3.5.

3.8.1 Rappels élémentaires

$\mathbb{K}$ désigne ici 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 élements 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 $(e_1, \cdots, e_n)$ 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 $(x_1, \cdots, x_n) \in \mathbb{K}^n$ tels que

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

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

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

Une famille non-libre est dite liée ou linéairement dépendante. La famille $(e_1, \cdots, e_n)$ 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 $(x_1, \cdots, x_n) \in \mathbb{K}^n$ tels que $x = \sum_{i=1}^n x_i e_i$. Les x_i sont les coordonnées de x dans la base $(e_1, \cdots, e_n)$.

L'espace E est de dimension finie si et seulement si il existe une base $(e_1, \cdots, e_n)$ 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 $(e_1, \cdots, e_n)$ où e_i est le vecteur $(0, \cdots, 0, 1, 0, \cdots, 0)$ 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 $k = 0, \cdots, n$. Il existe des espaces de dimension infinie, par exemple l'espace des polynômes de degré quelconque.

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

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

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 $L : E \to F$ 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

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

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

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

On rappelle que L est surjective si et seulement si $\mathrm{Im}(L) = F$, ce qui équivaut à $\mathrm{rg}(L) = \dim(F)$ lorsque F est de dimension finie, et que L injective si et seulement si $\mathrm{Ker}(L) = \{0\}$, ce qui équivaut à $\mathrm{rg}(L) = \dim(E)$ 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 $(e_1,\cdots,e_n)$ et $(f_1,\cdots,f_m)$ 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 $(f_1,\cdots,f_m)$:

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

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 $(e_1,\cdots,e_n)$ et $(f_1,\cdots,f_m)$. Dans le cas où $E = F$ et $e_i = f_i$ on dit que A est la matrice de L dans la base $(e_1,\cdots,e_n)$. 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 $\mathrm{Im}(A)$, $\mathrm{Ker}(A)$ et $\mathrm{rg}(A)$ 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 $a_{i,j} = 0$ si $i \neq j$ auquel cas on la note parfois $A = \mathrm{diag}(\mathrm{a_{1,1}},\cdots,\mathrm{a_{n,n}})$. La matrice identité est $I = \mathrm{diag}(1,\cdots,1)$. La *trace* d'une matrice carrée $A = (a_{i,j})$ est la quantité

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

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 $(e_1,\cdots,e_n)$ et $(f_1,\cdots,f_m)$ et si $U \in \mathcal{L}(F,G)$ a pour matrice $B = (b_{i,j}) \in \mathcal{M}_{p,m}$ dans les bases $(f_1,\cdots,f_m)$ et $(g_1,\cdots,g_p)$, 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 $(e_1,\cdots,e_n)$ et $(g_1,\cdots,g_p)$, où $C = BA$ est le produit matriciel défini par

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

Les matrices transposée et adjointe de $A = (a_{i,j}) \in \mathcal{M}_{m,n}(\mathbb{K})$ sont les matrices $A^t = (a^t_{i,j})$ et $A^* = (a^*_{i,j})$ de $\mathcal{M}_{n,m}(\mathbb{K})$ définies par $a^t_{i,j} = a_{j,i}$ 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 $\mathrm{Ker}(A^*)$ est le supplémentaire orthogonal de $\mathrm{Im}(A)$ 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 $\mathrm{Im}(A) = \mathbb{K}^n$ c'est-à-dire $\mathrm{rg}(A) = n$, ainsi qu'à $\mathrm{Ker}(A) = \{0\}$. 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 carré 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 $(a_1, \cdots, a_n)$ 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 $(e_1, \cdots, e_n)$ et $(f_1, \cdots, f_n)$ 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 $(f_1, \cdots, f_n)$ on obtient le vecteur de coordonnées de x dans la base $(e_1, \cdots, e_n)$. Si A est la matrice de $L \in \mathcal{L}(E)$ dans la base $(e_1, \cdots, e_n)$, alors

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

est la matrice de L dans la base $(f_1, \cdots, f_n)$. 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.

3.8.2 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 3.8.1 *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 λ.*

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

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

Il est appelé espace propre pour la valeur λ. 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 λ 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 λ_0 est une racine multiple de P c'est-à-dire $(\lambda - \lambda_0)^k$ se factorise dans $P(\lambda)$, on dit que λ_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 $\{\lambda_1, \cdots, \lambda_p\}$ sont les valeurs propres distinctes de A, les espaces propres E_{λ_i} ont la propriété de somme directe : pour tout vecteurs $u_i \in E_{\lambda_i}$

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

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 λ_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 3.8.2 *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*

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

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 $(\lambda_1, \cdots, \lambda_n)$ 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

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

Les valeurs propres de A sont donc exactement les $(\lambda_1, \cdots, \lambda_n)$. 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 3.8.1 *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 $Ae_1 = \lambda e_1$. On complète e_1 par des vecteurs $(e_2, ..., e_n)$ pour obtenir une base de $\mathbb{C}^n$. Si on introduit la matrice de passage P_1 de la base canonique dans la base $(e_1, ..., e_n)$, la représentation de A dans cette base est donc de la forme

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

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 $P_2^{-1} B P_2 = T_2$ 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 $(e_1, e_2, \cdots, e_n)$ et une nouvelle base $(e_1, f_2, \cdots, f_n)$

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

La matrice de passage de la base canonique dans la base $(e_1, f_2, \cdots, f_n)$ est $P = P_1 P_3$ et la représentation de A dans cette base est donc de la forme

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

où T est une matrice triangulaire supérieure. □

Remarque 3.8.1 *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 3.8.3 *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 3.8.1 *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 3.8.1 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 $(e_1, \cdots, e_n)$ 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 $(e_1, \cdots, e_n)$ 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 $(f_1, \cdots, f_{j-1})$, qui est aussi celui engendré par $(e_1, \cdots, e_{j-1})$. 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 $(f_1, \cdots, f_n)$, qui est unitaire puisque cette base est orthonormale. □

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 3.8.4 *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 3.8.2 *Une matrice $A \in \mathcal{M}_n(\mathbb{C})$ est normale si et seulement si 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 3.8.1 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. □

Théorème 3.8.3 *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 λ est une de ces valeurs propre et $x \neq 0$ un vecteur tel que $Ax = \lambda x$, on a

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

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

Remarque 3.8.2 *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 3.8.1, 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 3.8.3 *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 $(e_1, \cdots, e_n)$ 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 λ_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 $i = 1, \cdots, n$. Cette remarque s'étend aux matrices auto-adjointes complexes avec $q(x) = \langle Ax, x\rangle = \sum_{i=1}^n \sum_{j=1}^n a_{i,j}x_i\overline{x}_j$.

3.8.3 Normes matricielles

Rappelons la définition d'une norme.

Définition 3.8.5 *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

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

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

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

dont la norme euclidienne est un cas particulier ($p = 2$) et la norme "sup" ou ℓ^∞

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

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 3.8.4 *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 ℓ^p sur les coefficients matriciels définies pour $A = (a_{i,j})$ par

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

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

Définition 3.8.6 *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 3.8.4 *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 3.8.2 *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 ℓ^p est notée

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

et parfois appelée norme ℓ^p de A. Elle ne doit pas être confondue avec la norme ℓ^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 ℓ^1 et ℓ^∞.

Proposition 3.8.3 *Soit $\|A\|_1$ et $\|A\|_\infty$ les normes matricielles subordonnées aux normes ℓ^1 et ℓ^∞ sur $\mathbb{K}^n$. On a*

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

Preuve. Pour la norme ℓ^1 on écrit

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

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 ℓ^∞ on écrit

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

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 $u_j = \frac{\overline{a}_{i_0,j}}{|a_{i_0,j}|}$ 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,

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

d'où le résultat. □

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

Définition 3.8.7 *Le rayon spectral de $A\in\mathcal{M}_n(\mathbb{K})$ est défini par*

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

où $(\lambda_1,\cdots,\lambda_p)$ 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 3.8.4 *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 3.8.3 affirme qu'il existe une base orthonormée de vecteurs propres $(e_1,\cdots,e_n)$ de A. Pour $x\in\mathbb{K}^n$ décomposé suivant $x = \sum_{i=1}^n x_i e_i$ dans cette base on a

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

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 3.8.5 *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 λ 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.

Proposition 3.8.6 *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 $\mathbb{C}^n$ telle que*

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

Cette norme dépend en général de A et de ε.

Preuve. D'après la proposition 3.8.1, 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 $D_\delta = \mathrm{diag}(1, \delta, \delta^2, \cdots, \delta^{n-1})$ de telle sorte qu'en posant $U_\delta := UD_\delta$, la matrice T_δ 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 δ suffisamment petit pour que les éléments extra-diagonaux de T_δ 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_δ qui est semblable à A, on en déduit que $\|T_\delta\|_\infty \leq \varrho(A) + \varepsilon$. L'application

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

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

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

d'où le résultat. □

4 Estimations a priori pour l'approximation de fonctions

Dans cette section nous revenons sur les méthodes d'approximation de fonctions, et nous nous proposons de démontrer plusieurs estimations a priori dont l'analyse n'est pas immédiate. En utilisant la notion d'approximation *à noyau*, nous allons d'abord établir des estimations a priori d'ordre élevé pour les erreurs de meilleure approximation par des séries trigonométriques. Nous allons ensuite montrer comment ces résultats peuvent être appliqués à l'analyse des approximations polynomiales, ce qui nous permettra de démontrer le théorème 1.1.1 énoncé dans l'introduction. Nous terminerons par l'étude de la stabilité asymptotique de l'interpolation de Lagrange (i.e., l'étude de la stabilité des interpolations de degré n lorsque $n \to \infty$), avec pour objectif principal d'établir des estimations a priori pour les erreurs d'interpolation.

4.1 Approximations trigonométriques à noyau

4.1.1 Convergence des sommes de Fejer

On rappelle que les sommes de Fejer sont définies par la formule 2.42, $\mathcal{F}_n f(x) = \frac{1}{n+1}\sum_{k=0}^{n} \mathcal{S}_k f(x)$. Le fait que $\mathcal{F}_n f \in \mathbb{T}_n$ est facile à vérifier.

Pour étudier la convergence des sommes de Fejer, on observe qu'on peut les mettre sous la forme

$$\mathcal{F}_n f(x) = \frac{1}{2\pi(n+1)} \sum_{k=0}^{n} \sum_{l=-k}^{k} (\int_{-\pi}^{\pi} f(y) e^{-ily} dy) e^{ilx} = \int_{-\pi}^{\pi} f(y) \phi_n(x-y) dy$$

avec

$$\phi_n(z) := \frac{1}{2\pi(n+1)} \sum_{k=0}^{n} \sum_{l=-k}^{k} e^{ilz}. \tag{4.45}$$

La fonction ϕ_n, qui est un polynôme trigonométrique de degré n, est appelée *noyau de Fejer*. On remarque que l'on peut factoriser ϕ_n suivant

$$\begin{aligned} \phi_n(z) &= \frac{1}{2\pi(n+1)}(1 + e^{iz} + e^{2iz} + \cdots + e^{inz})(1 + e^{-iz} + e^{-2iz} + \cdots + e^{-inz}) \\ &= \frac{1}{2\pi(n+1)} \frac{1 - e^{i(n+1)z}}{1 - e^{iz}} \frac{1 - e^{-i(n+1)z}}{1 - e^{-iz}} \\ &= \frac{1}{2\pi(n+1)} \frac{\sin^2\left(\frac{(n+1)z}{2}\right)}{\sin^2(\frac{z}{2})}. \end{aligned} \tag{4.46}$$

On voit ainsi que le noyau de Fejer est positif, et possède les propriétés suivantes :

$$\int_{-\pi}^{\pi} \phi_n(z) dz = 1,$$

$\phi_n(0) = \frac{n+1}{2\pi}$ et pour tout $\alpha > 0$ et $\varepsilon > 0$ il existe n_0 tel que pour tout $n \geq n_0$ on a

$$\alpha \leq |z| \leq \pi \implies |\phi_n(z)| \leq \varepsilon$$

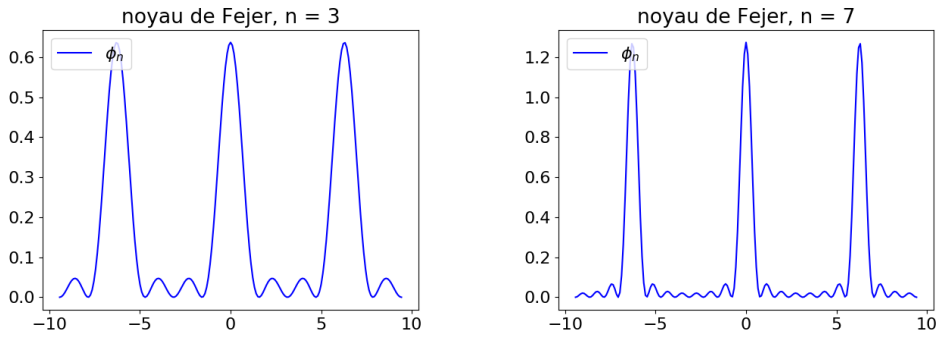

FIGURE 20 – Allure des noyaux de Fejer ϕ_n pour $n = 3$ (à gauche) et $n = 7$ (à droite), sur 3 périodes.

(voir figure 20). On peut alors démontrer le théorème de convergence 2.6.3, dont on rappelle ici l'énoncé.

Théorème 4.1.1 *Pour toute fonction f continue et 2π-périodique, on a*

$$\lim_{n\to 0} \|f - \mathcal{F}_n f\|_{L^\infty} = 0.$$

Preuve. La fonction f étant périodique, elle est uniformément continue. Pour $\varepsilon > 0$, on fixe α tel que pour tout z

$$|z| \leq \alpha \Rightarrow |f(x) - f(x-z)| \leq \varepsilon/2,$$

et on écrit ensuite

$$|f(x) - \mathcal{F}_n f(x)| = |\int_{-\pi}^{\pi} (f(x) - f(y))\phi_n(x-y)dy| = |\int_{-\pi}^{\pi} (f(x) - f(x-z))\phi_n(z)dz|,$$

où on a utilisé $\int_{-\pi}^{\pi} \phi_n = 1$, un changement de variable et la périodicité. Il vient que

$$|f(x) - \mathcal{F}_n f(x)| \leq \int_{|z|\leq\alpha} |f(x) - f(x-z))|\phi_n(z)dz + \int_{\alpha\leq|z|\leq\pi} |f(x) - f(x-z))|\phi_n(z)dz.$$

Le premier terme est inférieur à $\varepsilon/2$, et le second à $2\|f\|_{L^\infty} \int_{\alpha\leq|z|\leq\pi} \phi_n(z)dz$ et par conséquent inférieur à $\varepsilon/2$ pour n suffisamment grand, ce qui prouve que

$$\|f - \mathcal{F}_n f\|_{L^\infty} \leq \varepsilon,$$

pour n suffisamment grand, autrement dit $\mathcal{F}_n f$ converge uniformément vers f. □

4.1.2 Estimations a priori pour les sommes de Fejer

A nouveau, le théorème 4.1.1 n'est pas très satisfaisant d'un point de vue pratique car il ne donne aucune estimation sur la vitesse avec laquelle les erreurs $\|f - \mathcal{F}_n f\|_{L^\infty}$ tendent vers 0 lorsque $n \to \infty$.

Pour quantifier cette vitesse, on va utiliser la proposition suivante qui précise quelques propriétés du noyau de Fejer.

Proposition 4.1.1 *Le noyau de Fejer* (4.46) *possède les propriétés suivantes :*

$$\int_{-\pi}^{\pi} \phi_n(z)dz = 1, \quad \int_{-\pi}^{\pi} z\phi_n(z) = 0 \quad \text{et} \quad \int_{-\pi}^{\pi} z^2|\phi_n(z)|dz \leq \frac{K}{n}, \qquad \forall n \geq 1,$$

pour une constante $K > 0$ indépendante de n.

Preuve. La première propriété a déjà été observée (elle se voit facilement sur (4.45)), la deuxième se déduit de la parité de ϕ_n, et pour la troisième on écrit que

$$\int_{-\pi}^{\pi} z^2|\phi_n(z)|dz = \frac{1}{\pi(n+1)} \int_0^{\pi} \frac{z^2 \sin^2\left(\frac{(n+1)z}{2}\right)}{\sin^2(\frac{z}{2})} dz \leq \frac{1}{\pi(n+1)} \int_0^{\pi} \frac{z^2}{\sin^2(\frac{z}{2})} dz \leq \frac{\pi^2}{n+1}$$

où on a utilisé le fait que $\sin\left(\frac{z}{2}\right) \geq \frac{z}{\pi}$ sur $[0, \pi]$. Le résultat s'en déduit avec $K = \pi^2$. □

On peut alors démontrer le résultat suivant.

Théorème 4.1.2 *Si f est 2π-périodique et de classe $\mathcal{C}^2$, les sommes de Fejer vérifient*

$$\|f - \mathcal{F}_n f\|_{L^\infty} \leq C \frac{\|f''\|_{L^\infty}}{n}$$

pour $n \geq 1$, et une constante C indépendante de n.

Preuve. On démarre comme dans la preuve du theorème 4.1.1, en écrivant

$$|f(x) - \mathcal{F}_n f(x)| = \Big| \int_{-\pi}^{\pi} (f(x) - f(x-z))\phi_n(z)dz \Big|$$

et on observe que si f est une fonction périodique et de classe $\mathcal{C}^2$ sur $\mathbb{R}$, elle admet un développement de Taylor-Lagrange de la forme

$$f(x-z) = f(x) - zf'(x) + \frac{z^2}{2} f''(x_z) \quad \text{avec} \quad x_z \in [-\pi, \pi].$$

On déduit alors de la proposition 4.1.1 que

$$|f(x) - \mathcal{F}_n f(x)| = \Big| \int_{-\pi}^{\pi} \frac{z^2}{2} f''(x_z)\phi_n(z)dz \Big| \leq \frac{\|f''\|_{L^\infty}}{2} \frac{K}{n},$$

ce qui est l'estimation annoncée, avec $C = \frac{1}{2}K$. □

4.1.3 Estimations a priori pour des approximations d'ordre élevé

Il est possible de généraliser le principe ci-dessus en introduisant des procédés d'approximation par des polynômes trigonométriques qui permettent de converger plus rapidement. Ces procédés ont une forme générale

$$\mathcal{A}_n f(x) = \int_{-\pi}^{\pi} f(y)\psi_n(x-y)dy,$$

similaire aux sommes de Fejer, où ψ_n est une fonction choisie dans $\mathbb{T}_n$ ce qui entraîne que $\mathcal{A}_n f \in \mathbb{T}_n$. La fonction ψ_n est appelée *noyau de sommabilité*. Il est possible (exercice difficile) de mettre au point, pour un entier $m > 0$ arbitraire que l'on s'est fixé et pour n suffisamment grand, la fonction ψ_n de façon à ce qu'elle vérifie les propriétés suivantes :

$$\int_{-\pi}^{\pi} \psi_n(z)dz = 1 \text{ et } \int_{-\pi}^{\pi} z^k \psi_n(z) = 0, \ \ k = 1, \ldots, m-1,$$

et

$$\int_{-\pi}^{\pi} |z|^m |\psi_n(z)|dz \leq K_m n^{-m}, \ \ n \geq 0$$

pour une constante K_m indépendante de n. On pourra à titre d'exemple vérifier que la fonction $\psi_n(z) = \alpha_n(\phi_{\overline{n}}(z))^2$ avec $\overline{n}$ la partie entière de $n/2$ et $\alpha_n = (\int_{-\pi}^{\pi}(\phi_{\overline{n}}(z))^2 dz)^{-1}$ est un choix possible pour la valeur $m = 2$.

On peut alors reprendre la preuve du Theorème 4.1.2 en utilisant un développement de Taylor-Lagrange d'ordre m, de la forme

$$f(x) - f(x-z) = \sum_{k=1}^{m-1} a_k z^k + r(z)$$

où $|r(z)| \leq \frac{|z|^m}{m!} \|f^{(m)}\|_{L^\infty}$ lorsque $z \in [-\pi, \pi]$. On en déduit que

$$|f(x) - \mathcal{A}_n f(x)| = \Big| \int_{-\pi}^{\pi} r(z)\psi_n(z)dz \Big| \leq \frac{\|f^{(m)}\|_{L^\infty}}{m!} \int_{-\pi}^{\pi} |z|^m |\psi_n(z)|dz \leq C_m \frac{\|f^{(m)}\|_{L^\infty}}{n^m},$$

où $C_m = \frac{K_m}{m!}$. En résumé, on a obtenu le résultat suivant.

Théorème 4.1.3 *Si f est 2π-périodique et de classe $\mathcal{C}^m$, on a*

$$\inf_{g \in \mathbb{T}_n} \|f - g\|_{L^\infty} \leq C_m \frac{\|f^{(m)}\|_{L^\infty}}{n^m}$$

pour n suffisamment grand.

Par comparaison avec le théorème 2.6.2, on observe que la condition "$f \in \mathcal{C}^m$" 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*

4.2 Application aux approximations polynomiales

Dans cette section, on montre comment il est possible d'obtenir des bornes a priori sur la meilleure approximation polynomiale, à partir des bornes établies dans les sections précédentes sur la meilleure approximation par des polynômes trigonométriques.

4.2.1 Une preuve du théorème de Weierstrass

On commence par montrer comment la convergence uniforme des sommes de Fejer vers toute fonction continue permet de prouver le théorème 1.2.1 de Weierstrass qui affirme que toute fonction continue peut être approchée en distance uniforme par une suite de polynômes. Rappelons son énoncé.

Théorème 4.2.1 (de Weierstrass) *Si f est continue sur $I = [a, b]$, alors*

$$\lim_{n\to+\infty} \inf_{g\in\mathbb{P}_n} \|f - g\|_{L^\infty(I)} = 0.$$

Autrement dit, il existe une suite $(f_n)_{n\geq 0}$ de polynômes $f_n \in \mathbb{P}_n$ qui converge uniformément vers f sur l'intervalle I.

Preuve. La preuve se fait en 3 temps :

(i) on identifie une fonction périodique F qui prend les mêmes *valeurs* que f, et qui est continue comme f,

(ii) on utilise un résultat connu pour exhiber une suite de polynômes trigonométriques $F_n \in \mathbb{T}_n$ approchant F,

(ii) on montre qu'on peut en déduire l'existence d'une suite de polynômes $f_n \in \mathbb{P}_n$ qui approchent f.

On commence donc par observer 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 $g_n \in \mathbb{P}_n$, alors les fonctions $f_n = g_n \circ \phi^{-1}$ sont aussi dans $\mathbb{P}_n$ puisque ϕ^{-1} est affine et elles approchent uniformément f sur $[a, b]$:

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

On suppose donc à présent que f est une fonction continue sur $I = [-1, 1]$. On peut alors définir, pour tout $t \in \mathbb{R}$,

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

La fonction F est bien continue et 2π-périodique. D'après le Théorème 4.1.1, il existe une suite de polynômes trigonométrique $F_n \in \mathbb{T}_n$ qui converge uniformément vers F. On remarque qu'il est toujours possible de supposer que F_n est de la forme

$$F_n(t) = \sum_{k=0}^{n} c_k \cos(kt). \tag{4.47}$$

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 $T_k \in \mathbb{P}_k$, appelés *polynômes de Tchebychev*, tels que pour tout $k \geq 0$ et $t \in \mathbb{R}$,

$$\cos(kt) = T_k(\cos(t)). \tag{4.48}$$

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 $T_n \in \mathbb{P}_n$. 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 $f_n \in \mathbb{P}_n$ par

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

On conclut en écrivant

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

ce qui montre que f_n converge uniformément vers f. □

4.2.2 Estimations a priori pour les approximations polynomiales d'ordre élevé

En examinant la preuve du théorème de Weierstrass dans la section précédente, on constate que si f est de classe $\mathcal{C}^1$ sur $I = [-1, 1]$, alors F est aussi de classe $\mathcal{C}^1$ sur $\mathbb{R}$ avec

$$\|F'\|_{L^\infty} \leq \|f'\|_{L^\infty(I)}. \tag{4.49}$$

Par conséquent, en utilisant le Théorème 4.1.3 pour la valeur $m = 1$, on obtient le résultat suivant.

Proposition 4.2.1 *Si f est de classe $\mathcal{C}^1$ sur $[-1, 1]$ on a*

$$\inf_{g\in\mathbb{P}_n} \|f - g\|_{L^\infty(I)} \leq C_1 \|f'\|_{L^\infty(I)} n^{-1}.$$

Remarque 4.2.1 *Par changement de variable affine, on obtient les mêmes estimations sur l'intervalle $[a, b]$, avec la constante C_1 multipliée par $b - a$.*

On peut maintenant établir une preuve du théorème 1.1.1, dont on rappelle ici l'énoncé.

Théorème 4.2.2 *Si f est de classe $\mathcal{C}^m$ sur $I = [a, b]$, on a*

$$\inf_{g\in\mathbb{P}_n} \|f - g\|_{L^\infty(I)} \leq C_m \frac{\|f\|_{\mathcal{C}^m(I)}}{n^m} \tag{4.50}$$

avec une constante C_m dependant de m et de $|I| = b - a$, mais indépendante de n et f.

Preuve. On utilise le même raisonnement que dans la preuve du théorème de Weierstrass pour se ramener sur l'intervalle $[-1,1]$, puis à l'approximation d'une fonction 2π-périodique $F(t) = f(\cos(t))$. Le théorème 4.1.3 nous garantit alors l'existence d'une suite $F_n \in \mathbb{T}_n$ qui vérifie

$$\|F - F_n\|_{L^\infty} \leq C_m \|F^{(m)}\|_{L^\infty} n^{-m}$$

pour n suffisamment grand, et l'argument de parité déjà utilisé nous permet à nouveau de supposer que F_n est de la forme (4.47). La propriété (4.48) vérifiée par les polynômes de Tchebychev T_k nous permet ensuite d'exhiber un polynôme $f_n(x) = \sum_{k=0}^n c_k T_k(x) \in \mathbb{P}_n$ tel que

$$\|f - f_n\|_{L^\infty([-1,1])} \leq \|F - F_n\|_{L^\infty},$$

de sorte que la preuve est terminée si nous pouvons généraliser l'inégalité (4.49) aux ordres supérieurs. Pour cela nous pouvons démontrer par récurrence sur m que

$$F^{(m)}(t) = \sum_{k=0}^m G_{m,k}(t) f^{(k)}(\cos(t)) \tag{4.51}$$

où

$$G_{m,k}(t) = \sum_{j=0}^k \alpha_{m,k,j} (\cos(t))^j (\sin(t))^{k-j} \tag{4.52}$$

est un polynôme homogène de degré k en $\cos(t)$ et $\sin(t)$ dont les coefficients ne dépendent pas de f : en effet si $G_{m,k}$ s'écrit sous la forme (4.52), alors on a

$$G'_{m,k}(t) = \sum_{j=0}^k \Big(\alpha_{m,k,j-1}(k-(j-1)) - \alpha_{m,k,j+1}(j+1)\Big)(\cos(t))^j (\sin(t))^{k-j}$$

où l'on a posé $\alpha_{m,k,-1} = \alpha_{m,k,k+1} = 0$, de sorte que l'hypothèse de récurrence (4.51) entraîne

$$F^{(m+1)}(t) = \sum_{k=0}^m \Big(G'_{m,k}(t) f^{(k)}(\cos(t)) - \sin(t) G_{m,k}(t) f^{(k+1)}(\cos(t))\Big)$$

qui établit la propriété pour tout m. Ceci entraîne que

$$\|F^{(m)}\|_{L^\infty} \leq \Big(\sum_{k=0}^m \|G_{m,k}\|_{L^\infty}\Big) \|f\|_{\mathcal{C}^m([-1,1])},$$

et la périodicité des fonctions $G_{m,k}$ permet de montrer l'estimation (4.50) pour n suffisamment grand. Un petit raisonnement basé sur le fait que la norme de $\mathcal{C}^m$ majore la norme L^∞ permet enfin de montrer que cette estimation est valable pour tout $n \geq 1$. □

4.2.3 Convergence des approximations de Bernstein

Avant d'attaquer l'étude de l'interpolation polynomiale, on donne dans cette section une preuve du Théorème 2.5.1, dont on rappelle l'énoncé :

Théorème 4.2.3 *Si f est continue sur $[0,1]$, on a*

$$\lim_{n\to\infty} \|f - \mathcal{B}_n f\|_{L^\infty([0,1])} = 0.$$

Ce résultat nous fournit une preuve directe du théorème de Weierstrass, et sa démonstration emploie une technique assez proche de celle utilisée pour étudier la convergence des sommes de Fejer (théorème 4.1.1).

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

$$\mathcal{B}_n e_0(x) = \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} = (x + 1 - x)^n = 1. \tag{4.53}$$

Pour e_1, on obtient

$$\begin{array}{rl} \mathcal{B}_n e_1(x) & = \sum_{k=0}^n \frac{k}{n}\binom{n}{k} x^k(1-x)^{n-k} \\ & = \sum_{k=1}^n \binom{n-1}{k-1} x^k (1-x)^{n-k} \\ & = x\Big(\sum_{k=1}^n \binom{n-1}{k-1} x^{k-1}(1-x)^{n-1-(k-1)}\Big) \\ & = x\mathcal{B}_{n-1}e_0(x) = x. \end{array} \tag{4.54}$$

Pour e_2, on obtient par des considérations similaires

$$\begin{array}{rl} \mathcal{B}_n e_2(x) & = \sum_{k=0}^n (\frac{k}{n})^2 \binom{n}{k} x^k(1-x)^{n-k} \\ & = \frac{n-1}{n}\Big(\sum_{k=0}^n \frac{k(k-1)}{n(n-1)}\binom{n}{k}x^k(1-x)^{n-k} + \sum_{k=0}^n \frac{k}{n(n-1)}\binom{n}{k}x^k(1-x)^{n-k}\Big) \\ & = \frac{n-1}{n}\Big(x^2\mathcal{B}_{n-2}e_0(x) + \frac{1}{n-1}x\mathcal{B}_n e_0(x)\Big) \\ & = \frac{n-1}{n}x^2 + \frac{1}{n}x. \end{array} \tag{4.55}$$

Pour $x \in [0,1]$, on peut écrire

$$\begin{array}{rl} |f(x) - \mathcal{B}_n f(x)| & = |f(x) - \sum_{k=0}^n f(\frac{k}{n})\binom{n}{k}x^k(1-x)^{n-k}| \\ & = |\sum_{k=0}^n \Big(f(x) - f(\frac{k}{n})\Big)\binom{n}{k}x^k(1-x)^{n-k}|. \end{array}$$

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

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

et Σ_F la somme similaire pour $k \in F$, on a donc $|f(x) - \mathcal{B}_n f(x)| \le \Sigma_E + \Sigma_F$. On estime le premier terme en écrivant

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

Pour le second terme, on peut écrire

$$\begin{array}{rl} \Sigma_F & \le 2\|f\|_{L^\infty(0,1)}\sum_{k\in F}\binom{n}{k}x^k(1-x)^{n-k} \\ & \le \frac{2\|f\|_{L^\infty(0,1)}}{\delta^2}\sum_{k\in F}\Big(x - \frac{k}{n}\Big)^2\binom{n}{k}x^k(1-x)^{n-k} \\ & = \frac{2\|f\|_{L^\infty(0,1)}}{\delta^2}(x^2\mathcal{B}_n e_0(x) - 2x\mathcal{B}_n e_1(x) + \mathcal{B}_n e_2(x)) \\ & = \frac{2\|f\|_{L^\infty(0,1)}}{\delta^2}\frac{1}{n}(x - x^2) \le \frac{2\|f\|_{L^\infty(0,1)}}{n\delta^2}. \end{array}$$

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

$$\|f - \mathcal{B}_n f\|_{L^\infty(0,1)} \leq \omega(f,\delta) + \frac{2\|f\|_{L^\infty(0,1)}}{n\delta^2}.$$

Pour tout $\varepsilon > 0$, on peut choisir $\delta > 0$ tel que $\omega(f,\delta) \leq \varepsilon/2$, puis n_0 tel que $\frac{2\|f\|_{L^\infty(0,1)}}{n\delta^2} \leq \varepsilon/2$ pour $n \geq n_0$, ce qui entraîne $\|f - \mathcal{B}_n f\|_{L^\infty(0,1)} \leq \varepsilon$. On a ainsi montré que $\mathcal{B}_n f$ converge uniformément vers f sur $[0,1]$. □

Remarque 4.2.2 *En s'inspirant de l'étude des sommes de Fejer, on peut démontrer que lorsque $f \in \mathcal{C}^2$ les approximations de Bernstein vérifient*

$$\|f - \mathcal{B}_n f\|_{L^\infty(0,1)} \leq C\frac{\|f''\|_{L^\infty}}{n}$$

pour une constante C indépendante de n. Pour cela on écrit

$$\mathcal{B}_n f(x) = \sum_{k=0}^{n} f\Big(\frac{k}{n}\Big) b_{n,k}(x)$$

avec $b_{n,k}(x) = \binom{n}{k} x^k (1-x)^{n-k}$, et on observe que les égalités (4.53)-(4.55) *nous donnent*

$$\sum_{k=0}^{n} b_{n,k}(x) = 1, \quad \sum_{k=0}^{n} \Big(x - \frac{k}{n}\Big) b_{n,k}(x) = 0, \quad \sum_{k=0}^{n} \Big(x - \frac{k}{n}\Big)^2 b_{n,k}(x) \leq \frac{1}{n}.$$

On peut alors raisonner comme dans la preuve du théorème 4.1.2.

4.3 Analyse d'erreur pour l'interpolation polynomiale

On se propose à présent d'établir des estimations a priori pour l'erreur entre f et son polynôme d'interpolation de Lagrange introduit dans la section 2.2.1. On rappelle que pour un choix arbitraire de points $a \leq x_0 < \cdots < x_n \leq b$ dans un intervalle $I = [a,b]$, on désigne par $\mathcal{I}_n$ l'opérateur d'interpolation qui à une fonction f continue sur $[a,b]$ associe son polynôme d'interpolation p_n de degré n en ces points.

4.3.1 Estimations directes de l'erreur d'interpolation

Dans cette section on va raisonner de façon directe, et on commence par énoncer un résultat qui généralise le théorème de Rolle.

Lemme 4.3.1 *Soit f une fonction de classe $\mathcal{C}^n$ sur un intervalle I et qui s'annule en $n+1$ points distincts $a_0 < \cdots < a_n$ 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 en $a_0, \cdots, a_n$, alors f' s'annule en n points $b_0, \cdots, b_{n-1}$ avec $a_i < b_i < a_{i+1}$. Par l'hypothèse de récurrence, $f^{(n)} = (f')^{(n-1)}$ s'annule en un point $z \in]b_0, b_{n-1}[\subset]a_0, a_n[$. □

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

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

Théorème 4.3.1 *Soit f une fonction de classe $\mathcal{C}^{n+1}$ sur I et p_n son polynôme d'interpolation de Lagrange en $n+1$ points distincts $x_0 < \cdots < x_n$ contenus dans I, alors pour tout $x \in I$ il existe $y_x \in I$ qui dépend de x, tel que*

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

Le point y_x 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 se fixe un x différent de tous les x_i, ce qui entraîne $\Pi_n(x) \neq 0$. Par conséquent, il existe un nombre $\mu_x \in \mathbb{R}$ (qui dépend de x) tel que

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

La fonction $g(t) = f(t) - p_n(t) - \mu_x\Pi_n(t)$ s'annule aux $n+2$ points distincts $x_0, \cdots, x_n$ et x. Par conséquent, d'après le Lemme 4.3.1 il existe un point $y_x \in]\min\{x, x_0\}, \max\{x, x_n\}[$ (qui dépend de x) tel que

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

et par conséquent $\mu_x = f^{(n+1)}(y_x)$ ce qui donne le résultat. □

Remarque 4.3.1 *Lorsque f n'est pas de classe $\mathcal{C}^{n+1}$ mais seulement continue, on peut établir la formule d'erreur*

$$f(x) - p_n(x) = f[x_0, x_1, \cdots, x_n, x]\prod_{i=0}^{n}(x - x_i)$$

à partir de la forme de Newton du polynôme d'interpolation, en remarquant qu'au point x la fonction f coïncide avec le polynôme d'interpolation p_{n+1} de f aux points $\{x_0, \cdots, x_n, x\}$. Combinée au résultat précédent, cette formule d'erreur entraîne la propriété suivante : si f est de classe $\mathcal{C}^n$, alors pour tout ensemble de points $\{y_0, \cdots, y_n\}$ il existe $y \in]\min x_i, \max x_i[$ tel que

$$f[y_0, \cdots, y_n] = \frac{f^{(n)}(y)}{n!}.$$

A partir du Théorème 4.3.1 on peut estimer l'erreur d'interpolation sur un intervalle $[a, b]$ qui contient $[x_0, x_n]$. Une première conséquence est l'estimation en norme sup sur $[a, b]$.

$$\|f - p_n\|_{L^\infty([a,b])} \leq \|f^{(n+1)}\|_{L^\infty([a,b])}\|\Pi_n\|_{L^\infty([a,b])}$$

On peut estimer la norme sup de Π_n en écrivant

$$\|\Pi_n\|_{L^\infty([a,b])} = \frac{1}{(n+1)!}\max_{x\in[a,b]}\prod_{i=0}^{n}|x - x_i| \leq \frac{(b-a)^{n+1}}{(n+1)!},$$

ce qui entraine l'estimation a priori suivante.

Proposition 4.3.1 *Soit $f \in \mathcal{C}^{(n+1)}(I)$ et $\mathcal{I}_n$ l'opérateur d'interpolation de Lagrange défini sur $n+1$ points distincts dans $I = [a, b]$. On a*

$$\|f - \mathcal{I}_n f\|_{L^\infty(I)} \leq \frac{(b-a)^{n+1}}{(n+1)!} \|f^{(n+1)}\|_{L^\infty(I)}. \tag{4.56}$$

Remarque 4.3.2 *Dans le cas de points $a = x_0 < \cdots < x_n$ 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

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

asymptotiquement meilleure que (4.56) *quand $n \to +\infty$.*

On peut observer que l'estimation (4.56) a la même forme que celle qu'on a établie dans la proposition 2.1.1 pour les développements de Taylor. On sait donc qu'elle permet de démontrer une convergence très rapide pour des fonctions très régulières, en revanche elle ne nous apportera aucune information sur le comportement des interpolations lorsque f est de régularité modérée (i.e., de classe $\mathcal{C}^m$ avec $m < \infty$) ou même dans $\mathcal{C}^\infty$ mais avec une croissance rapide des normes $\|f^{(n)}\|_{L^\infty}$ lorsque $n \to \infty$. Cette observation n'est pas étonnante puisqu'on a vu que les interpolations pouvaient diverger pour de telles fonctions lorsque les points d'interpolation n'étaient pas bien choisis (voir par exemple la figure 11).

Pour avoir des estimations qui soient applicables à des fonctions de régularité modérée, on va reprendre dans la section 4.3.2 le cadre théorique présenté dans l'introduction, en analysant la stabilité asymptotique (i.e., pour $n \to \infty$) des opérateurs d'interpolation associés à des noeuds bien choisis. Avant cela, on peut démontrer que les interpolations (avec des noeuds arbitraires) convergent vers f lorsque celle-ci est très régulière au sens où elle admet un développement en série entière sur un intervalle contenant l'intervalle $[a, b]$.

Théorème 4.3.2 *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 $0 \leq u \leq \frac{b-a}{2}$, on a $\frac{d^n}{du^n}(u^k) \geq 0$ et on peut donc écrire

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

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

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

On a par conséquent

$$\|f^{(n+1)}\|_{L^\infty([a,b])} \leq C(r)(n+1)! r |r - \frac{b-a}{2}|^{-(n+2)},$$

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

$$\|f - p_n\|_{L^\infty([a,b])} \leq r C'(r) \rho^{n+1} \quad \text{avec} \quad \rho := \frac{b-a}{|r - \frac{b-a}{2}|} \quad \text{et} \quad C'(r) = \frac{C(r)}{|r - \frac{b-a}{2}|}$$

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

Remarque 4.3.3 *Dans le cas où les points x_i sont équidistants avec $x_0 = a$ et $x_n = b$, on peut utiliser l'estimation $\|f - p_n\|_{L^\infty([a,b])} \leq \frac{1}{n^{n+2}} \|f^{(n+1)}\|_{L^\infty([a,b])} (b-a)^{n+1}$ 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 $n \to +\infty$).*

La preuve du Théorème 4.3.2 nous indique que la convergence de p_n vers f est très rapide, puisque $\|f - p_n\|_{L^\infty([a,b])} \leq C\rho^{n+1}$ avec $0 < \rho < 1$, mais à nouveau 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 a priori et nous avons vu sur plusieurs exemples numériques qu'en pratique elle pouvait ne pas avoir lieu, que f soit de régularité faible (figure 1) ou même de classe $\mathcal{C}^\infty$ (figure 11).

Dans la section suivante nous abordons donc l'analyse de la stabilité asymptotique qui nous permettra de nous appuyer sur le résultat optimal d'approximation polynomiale du théorème 1.1.1, valable pour des fonctions de classe $\mathcal{C}^m$. Comme on s'y attend, le choix des points d'interpolation va jouer un rôle important dans cette étude.

4.3.2 Stabilité asymptotique des interpolations polynomiales

On a vu dans la section 2.2.1 que l'opérateur d'interpolation pouvait se mettre sous la forme

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

où les ℓ_i sont les fonctions de bases de Lagrange aux points $x_0, \cdots, x_n$. Il est immédiat de vérifier que $\mathbb{P}_n$ est une application linéaire c'est-à-dire un élément de $\mathcal{L}(\mathcal{C}^0([a,b]), \mathbb{P}_n)$.

On appelle *constante de Lebesgue* du procédé d'interpolation de Lagrange aux points $x_0, \cdots, x_n$ la norme de l'opérateur $\mathcal{I}_n$ subordonnée à la norme sup sur $[a, b]$ au sens de (1.10), c'est-à-dire

$$\Lambda_n := \sup_{\substack{f \in \mathcal{C}^0([a,b]) \\ f \neq 0}} \frac{\|\mathcal{I}_n f\|_{L^\infty([a,b])}}{\|f\|_{L^\infty([a,b])}} = \sup_{\substack{f \in \mathcal{C}^0([a,b]) \\ \|f\|_{L^\infty([a,b])} \leq 1}} \|\mathcal{I}_n f\|_{L^\infty([a,b])}.$$

La constante de Lebesgue joue un rôle central dans l'étude de la stabilité du procédé d'interpolation puisque pour toute paire de fonctions f et g on a

$$\|\mathcal{I}_n f - \mathcal{I}_n g\|_{L^\infty([a,b])} \leq \Lambda_n \|f - g\|_{L^\infty([a,b])}.$$

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

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

Plus précisément, on a

$$\Lambda_n = \sup_{\substack{y \in \mathbb{R}^{n+1} \\ y \neq 0}} \frac{\|\mathcal{J}_n y\|_{L^\infty([a,b])}}{\|y\|_\infty} = \sup_{\substack{y \in \mathbb{R}^{n+1} \\ \|y\|_\infty \leq 1}} \|\mathcal{J}_n y\|_{L^\infty([a,b])}.$$

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

On a vu dans l'introduction que la stabilité asymptotique d'un procédé d'approximation jouait également un rôle fondamental dans l'étude de la convergence. On pense en particulier à la proposition 1.1.1 : comme l'opérateur $\mathcal{I}_n$ préserve les polynômes de degré $\leq n$, il lui suffirait d'être uniformément stable pour être quasi-optimal au sens de (1.7), ce qui en vertu du théorème 1.1.1 entraînerait une convergence d'ordre m pour toute fonction de classe $\mathcal{C}^m$.

En ce qui concerne les interpolations nous ne serons pas en mesure de démontrer la stabilité uniforme, mais nous pourrons utiliser le résultat suivant.

Théorème 4.3.3 *Pour tout $f \in \mathcal{C}^0([a,b])$, on a*

$$\|f - \mathcal{I}_n f\|_{L^\infty([a,b])} \leq (1+\Lambda_n) \inf_{g\in\mathbb{P}_n} \|f-g\|_{L^\infty([a,b])}.$$

Preuve. Pour tout $g \in \mathbb{P}_n$, on peut écrire

$$\begin{aligned}\|f - \mathcal{I}_n f\|_{L^\infty([a,b])} &\leq \|f-g\|_{L^\infty([a,b])} + \|\mathcal{I}_n f - g\|_{L^\infty([a,b])} \\ &= \|f-g\|_{L^\infty([a,b])} + \|\mathcal{I}_n f - \mathcal{I}_n g\|_{L^\infty([a,b])} \leq (1+\Lambda_n)\|f-g\|_{L^\infty([a,b])},\end{aligned}$$

en utilisant le fait que $\mathcal{I}_n g = g$. Comme g est arbitraire on obtient le résultat annoncé. □

En combinant ce résultat avec ceux qui décrivent l'erreur de meilleure approximation par des polynômes, on obtient des estimations sur l'erreur d'interpolation. Par exemple, en utilisant le Théorème 4.2.2, on obtient le résultat suivant.

Corollaire 4.3.1 *Si f est de classe $\mathcal{C}^m$ sur $[a,b]$, alors*

$$\|f - \mathcal{I}_n f\|_{L^\infty([a,b])} \leq C_m \|f\|_{\mathcal{C}^m([a,b])}(1+\Lambda_n) n^{-m}$$

pour une constante C_m dépendant de m et de $b-a$ mais indépendante de n et f.

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

Proposition 4.3.2 *On a*

$$\Lambda_n = \max_{x\in[a,b]} \sum_{i=0}^n |\ell_i(x)| = \Big\| \sum_{i=0}^n |\ell_i| \Big\|_{L^\infty([a,b])}$$

où les ℓ_i sont les fonctions de base de Lagrange.

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

$$|\mathcal{I}_n f(x)| = |\sum_{i=0}^n f(x_i)\ell_i(x)| \leq \Big(\max_{i=0,\cdots,n} |f(x_i)|\Big) \sum_{i=0}^n |\ell_i(x)| \leq \|f\|_{L^\infty([a,b])} \Big\| \sum_{i=0}^n |\ell_i| \Big\|_{L^\infty([a,b])},$$

ce qui entraine

$$\|\mathcal{I}_n f\|_{L^\infty([a,b])} \leq \|f\|_{L^\infty([a,b])} \Big\| \sum_{i=0}^n |\ell_i| \Big\|_{L^\infty([a,b])},$$

et par conséquent $\Lambda_n \leq \Big\| \sum_{i=0}^n |\ell_i| \Big\|_{L^\infty([a,b])}$. Pour démontrer l'inégalité inverse, on considère le point x^* tel que

$$\sum_{i=0}^n |\ell_i(x^*)| = \max_{x\in[a,b]} \sum_{i=0}^n |\ell_i(x)| = \Big\| \sum_{i=0}^n |\ell_i| \Big\|_{L^\infty([a,b])},$$

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 $\|f\|_{L^\infty([a,b])} = 1$ (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

$$\|\mathcal{I}_n f\|_{L^\infty([a,b])} \geq |\mathcal{I}_n f(x^*)| = |\sum_{i=0}^n y_i \ell_i(x^*)| = \sum_{i=0}^n |\ell_i(x^*)| = \Big\| \sum_{i=0}^n |\ell_i| \Big\|_{L^\infty([a,b])},$$

et par conséquent $\Lambda_n \geq \Big\| \sum_{i=0}^n |\ell_i| \Big\|_{L^\infty([a,b])}$. □

Considérons à present 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 4.3.3 *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

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

Et par conséquent

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

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

Remarque 4.3.5 *On peut aussi majorer* Λ_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 Λ_n est plus forte que toute décroissance en n^{-m}, et par conséquent les estimations telles que celles du Corollaire 4.3.1 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

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

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 $T_{n+1} \in \mathbb{P}_{n+1}$ 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 $u \mapsto x = \frac{a+b}{2} + \frac{b-a}{2}u$. On pose donc

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

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

Proposition 4.3.4 *Pour les points de Tchebychev on a $\Lambda_n \leq C\log(n)$ pour tout $n > 1$, 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

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

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

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

D'autre part, on peut écrire

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

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

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

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

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

Corollaire 4.3.2 *Si f est de classe $\mathcal{C}^m$ sur $[a,b]$, alors pour $\mathcal{I}_n$ l'opérateur d'interpolation avec les points de Tchebychev, on a*

$$\|f - \mathcal{I}_n f\|_{L^\infty([a,b])} \leq C_m \|f\|_{\mathcal{C}^m([a,b])} n^{-m}\log(n)$$

avec une constante C_m dépendant de m et de $b-a$ mais indépendante de n et f.

Remarque 4.3.6 *Cette estimation garantit que l'interpolation sur les noeuds de Tchebychev ne produit pas d'effet de Runge. En particulier, si f est une fonction de régularité modérée (par exemple de classe $\mathcal{C}^1$), cette estimation garantit que ses interpolations $\mathcal{I}_n f$ convergent uniformément vers f : ils ne peuvent donc pas osciller de façon importante.*

5 Calcul approché des intégrales

On présente dans cette section une application importante des méthodes d'approximations vues plus haut, à savoir le calcul approché des intégrales. En effet, 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 mais 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 $a = a_0 < \cdots < a_k = b$ 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.

5.1 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

$$\int_a^b f(x)dx = \int_a^b \Big(f(a) + \int_a^x f'(t)dt\Big)dx = (b - a)f(a) + \int_a^b \int_a^x f'(t)dtdx,$$

d'où

$$\left|\int_a^b f(x)dx - (b - a)f(a)\right| \leq \int_a^b \int_a^x |f'(t)|dtdx \leq \|f'\| \int_a^b \int_a^x dtdx = \frac{(b - a)^2}{2}\|f'\|,$$

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

$$\left|\int_a^b f(x)dx - \Sigma(f)\right| \leq \frac{1}{2}\|f'\|\Big(\sum_{i=0}^{k-1}(a_{i+1}-a_i)\Big)h = \frac{b-a}{2}\|f'\|h.$$

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

$$\begin{aligned}\textstyle\int_a^b f(x)dx &= \textstyle\int_a^b\Big(f(\frac{a+b}{2}) + (x-\frac{a+b}{2})f'(\frac{a+b}{2}) + \int_{\frac{a+b}{2}}^x (x-t)f''(t)dt\Big)dx\\ &= (b-a)f(\tfrac{a+b}{2}) + \textstyle\int_a^b\int_{\frac{a+b}{2}}^x (x-t)f''(t)dtdx,\end{aligned}$$

d'où

$$\left|\int_a^b f(t)dt - (b-a)f(\frac{a+b}{2})\right| \leq \|f''\|\int_a^b\int_{\frac{a+b}{2}}^x (x-t)dtdx = \frac{(b-a)^3}{24}\|f''\|,$$

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 $h \to 0$. 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 $a = a_0 < \cdots < a_k = b$.

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

$$\int_a^b p_1(x)dx = \int_a^b\Big(f(a) + (t-a)\frac{f(b)-f(a)}{b-a}\Big)dt = (b-a)\frac{f(a)+f(b)}{2}.$$

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 §4.4, 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 remplacant dans la quadrature simple la fonction affine p_1 par le polynôme d'interpolation $p_n \in \mathbb{P}_n$ pour la subdivision équidistante $x_j = a + \frac{j}{n}(b-a)$ pour $j = 0, \cdots, n$. La règle de quadrature simple peut s'écrire

$$\int_a^b p_n(x)dx = \sum_{j=0}^n \Big(\int_a^b \ell_j(x)dx\Big) f(x_j).$$

où les ℓ_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 $\tilde{p}_n \in \mathbb{P}_n$ est le polynôme d'interpolation de $\tilde{f} = f \circ \phi$ aux points équidistants $y_i = -1 + \frac{2j}{n}$ 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 ℓ_j désigne ici la fonction de base de Lagrange sur associée au point y_j de la subdivision uniforme de $[-1,1]$. Les poids ω_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*

$$\int_a^b p_2(x)dx = (b-a)\Big(\frac{1}{6}f(a) + \frac{2}{3}f(\frac{a+b}{2}) + \frac{1}{6}f(b)\Big),$$

dont la version composée est

$$S(f) = \sum_{i=0}^{k-1}(a_{i+1}-a_i)\Big(\frac{1}{6}f(a_i) + \frac{2}{3}f(\frac{a_i+a_{i+1}}{2}) + \frac{1}{6}f(a_{i+1})\Big).$$

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 entraine que l'on a nécessairement

$$\sum_{j=0}^n \omega_j = 2.$$

5.2 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 les x_j sont de la forme $x_j = \phi(y_j) = \frac{a+b}{2} + \frac{b-a}{2}y_j$ avec $\{y_0, \cdots, y_n\}$ fixés dans l'intervalle $[-1, 1]$, et où les poids ω_j vérifient la relation $\sum_{j=0}^{n}\omega_j = 2$. Toutes les méthodes de la section précédente sont de ce type.

Définition 5.2.1 *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 $p \in \mathbb{P}_m$ on a*

$$\int_a^b p(x)dx = \frac{b-a}{2}\sum_{j=0}^{n}\omega_j p(x_j),$$

Remarque 5.2.1 *Afin de vérifier qu'une quadrature sur $[a, b]$ est d'ordre m, il suffit de le vérifier 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 $p \in \mathbb{P}_m$, en posant $\tilde{p} = p \circ \phi \in \mathbb{P}_m$,*

$$\frac{b-a}{2}\sum_{j=0}^{n}\omega_j p(x_j) = \frac{b-a}{2}\sum_{j=0}^{n}\omega_j \tilde{p}(y_j) = \frac{b-a}{2}\int_{-1}^{1}\tilde{p}(y)dy = \int_a^b p(x)dx.$$

D'autre part, par linéarité, il est suffisant de vérifier que la quadrature est exacte pour les fonctions $x \mapsto x^k$ pour $k = 0, \cdots, m$. Une quadrature est donc d'ordre m si on a

$$\sum_{j=0}^{n}\omega_j y_j^k = \int_{-1}^{1} y^k dy = \frac{1+(-1)^k}{k+1}, \quad k = 0, \cdots, m.$$

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 de quadrature donnée, on note

$$E(f) = \int_a^b f(x)dx - \frac{b-a}{2}\sum_{j=0}^{n}\omega_j f(x_j),$$

Théorème 5.2.1 *Si la méthode de quadrature est d'ordre m, on a pour toute fonction $f \in \mathcal{C}^{m+1}([a, b])$*

$$|E(f)| \leq C(b-a)^{m+2}\|f^{(m+1)}\|$$

où $C := \frac{1+\frac{1}{2}(\sum_{j=0}^{n}|\omega_j|))}{2^{m+1}(m+1)!}$.

Preuve. On considère le développement de Taylor-Lagrange de f autour du point $c = \frac{a+b}{2}$ qui est de la forme

$$f(x) = q_m(x) + r(x) \;\; q_m \in \mathbb{P}_m \;\text{ et }\; r(x) = \frac{(x-c)^{m+1}}{(m+1)!} f^{(m+1)}(y),$$

avec y compris entre c et x. La quadrature étant d'ordre m on a $E(q_m) = 0$ et par conséquent

$$|E(f)| = |E(r)| \leq \int_a^b |r(x)|dx + \frac{b-a}{2} \sum_{j=0}^{n} |\omega_j r(x_j)| \leq (b-a)(1 + \frac{1}{2}(\sum_{j=0}^{n} |\omega_j|))\|r\|.$$

On conclut en observant que

$$\|r\| \leq \frac{(b-a)^{m+1}}{2^{m+1}(m+1)!} \|f^{(m+1)}\|$$

□

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 5.2.1 *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 la constante du théorème 5.2.1 et $h = \max |a_{i+1} - a_i|$.

Terminons en indiquant qu'il est possible d'obtenir une meilleure constante C dans les résultats ci-dessus au moyen d'une analyse plus fine. A cet effet, on introduit pour tout $t \in \mathbb{R}$ et $m \in \mathbb{N}$ la fonction

$$g_{t,m}(x) := (x-t)_+^m = \Big(\max\{x-t, 0\}\Big)^m$$

Théorème 5.2.2 *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 la 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 entraine d'une part que

$$\int_a^b f(x)dx = \int_a^b p(x)dx + \frac{1}{m!}\int_a^b \Big(\int_a^b g_{t,m}(x)dx\Big) f^{(m+1)}(t)dt,$$

et d'autre part que

$$\frac{b-a}{2}\sum_{j=0}^n \omega_j f(x_j) = \frac{b-a}{2}\sum_{j=0}^n \omega_j p(x_j) + \frac{1}{m!}\int_a^b \Big(\frac{b-a}{2}\sum_{j=0}^n \omega_j g_{t,m}(x_j)\Big) f^{(m+1)}(t)dt.$$

En soustrayant ces deux identités et en utilisant le fait que la quadrature est exacte pour p, on obtient le résultat annoncé. □

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

$$K_m(t) = \frac{b-a}{2}\Big(\int_{-1}^1 \tilde{g}_{t,m}(y)dy - \sum_{j=0}^n \omega_j \tilde{g}_{t,m}(y_j)\Big).$$

On remarque ensuite que

$$\tilde{g}_{\phi(t),m}(y) = g_{\phi(t),m}(\phi(y)) = (\phi(y)-\phi(t))_+^m = \Big(\frac{b-a}{2}(y-t)\Big)_+^m = \Big(\frac{b-a}{2}\Big)^m g_{t,m}(y),$$

et par conséquent

$$K_m(\phi(t)) = \Big(\frac{b-a}{2}\Big)^{m+1} k_m(t),$$

c'est-à-dire $K_m \circ \phi = \Big(\frac{b-a}{2}\Big)^{m+1} k_m$. Ceci entraine immédiatement

$$\int_a^b |K_m(t)|dt = \Big(\frac{b-a}{2}\Big)^{m+2}\int_{-1}^1 |k_m(t)|dt.$$

Nous avons donc établi le résultat suivant.

Corollaire 5.2.2 *Si la méthode de quadrature est d'ordre m, on a pour toute fonction $f \in \mathcal{C}^{m+1}([a,b])$*

$$|E(f)| \leq \tilde{C}(b-a)^{m+2}\|f^{(m+1)}\|,$$

avec

$$\tilde{C} := \frac{1}{m!\,2^{m+2}} \int_{-1}^{1} |k_m(t)|dt.$$

Le calcul de la nouvelle constante $\tilde{C}$ (qui est toujours plus petite que la constante C du Théorème 5.2.1) peut être facilité lorsque le noyau de Peano $k_m(t)$ est de signe constant sur $[-1,1]$. On a dans ce cas

$$\tilde{C} = \frac{1}{m!\,2^{m+2}} \left| \int_{-1}^{1} k_m(t)dt \right|,$$

et d'autre part, pour la fonction $e_{m+1}(x) := x^{m+1}$, la formule d'erreur du Théorème 5.2.2 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

$$\tilde{C} = \frac{1}{(m+1)!\,2^{m+2}}|E(e_{m+1})| = \frac{1}{(m+1)!\,2^{m+2}} \left| \frac{1+(-1)^{m+1}}{m+2} - \sum_{j=0}^{n} \omega_j y_j^{m+1} \right|$$

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

$$\tilde{C} := \frac{1}{16}\left|\frac{2}{3} - 2\right| = \frac{1}{12}.$$

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

$$\tilde{C} = \frac{1}{768}\left|\frac{2}{5} - \frac{2}{3}\right| = \frac{1}{2880}.$$

5.3 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 5.3.1 *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* $f \in \mathbb{P}_{2n+1}$*. Les poids* ω_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 $q \in \mathbb{P}_n$, le produit $p_{n+1}q$ est de degré $2n+1$ et par conséquent

$$\int_{-1}^{1} p_{n+1}(x)q(x)dx = \sum_{j=0}^{n} \omega_j p_n(x_j)q(x_j) = 0.$$

Ceci montre que p_{n+1} est orthogonal à $\mathbb{P}_n$ 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 §4.6 et α_{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}. Il est possible de vérifier que ces racines $\{x_0, \cdots, x_n\}$ sont toutes contenues dans l'intervalle $[-1,1]$. Plus précisément on montre en utilisant la formule de récurrence donnée dans la remarque 3.6.2, qu'avec la normalisation $L_k(1) = 1$ on a $|L_k(x)| \geq 1$ pour tout $k \geq 0$ et $|x| \geq 1$, ce qui implique cette propriété. 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 $\mathbb{P}_n$ et donc a-fortiori dans $\mathbb{P}_{2n+1}$, 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 ω_j sont aussi uniquement déterminés. L'unicité est donc établie. Montrons à présent que pour ce choix des points x_j et des poids ω_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 $q, r \in \mathbb{P}_n$. Comme q et p_{n+1} sont orthogonaux, on a

$$\int_{-1}^{1} p(x)dx = \int_{-1}^{1} r(x)dx.$$

Puisque $r \in \mathbb{P}_n$ 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 $\ell_i^2 \in \mathbb{P}_{2n}$ 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 ω_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.

On peut ainsi calculer la constante C intervenant dans les estimations d'erreur obtenues en appliquant le théorème 5.2.1 et son corrollaire à la méthode de Gauss-Legendre. 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 $\tilde{C}$.

Remarque 5.3.1 *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 §4.6.