Florian DUSSAP
CY Cergy Paris Université
Site de Saint-Martin
Bureau E538
florian.dussap@cyu.fr

Cours d'analyse numérique

L3 Mathématiques

2026 – 2027

Sommaire

Préambule

Introduction

L'analyse numérique est un domaine à l'interface des mathématiques et de l'informatique qui s'intéresse à la résolution approchée des problèmes d'analyse : résolution d'équations algébriques et différentielles, calcul d'intégrales, optimisation de fonctions, approximation de fonctions. En dehors de cas particuliers, les solutions de ces problèmes ne peuvent pas s'exprimer à l'aide d'une formule close. On a alors recours à des méthodes numériques, c'est-à-dire des algorithmes, pour calculer des solutions approchées. L'analyse numérique étudie ces méthodes numériques, leur conception, l'évaluation de l'erreur, jusqu'à leur implémentation en machine. C'est un domaine d'une grande importance en mathématiques appliquées, en physique et en ingénierie (et bien d'autres sciences).

Ce cours se concentre sur les problèmes en dimension 1. En réalité, beaucoup de méthodes décrites dans ce cours, et les résultats les concernant, se généralisent en dimension supérieure. Cependant, on se limite à la dimension 1 pour faciliter certaines preuves, et mieux se concentrer sur les principes de l'étude des méthodes numériques. L'étudiant·e qui souhaite approfondir ces sujets pourra consulter les références ci-après.

Ce cours est constitué de quatre chapitres. Le chapitre 1 traite de la résolution approchée des équations. On passe en revue les méthodes les plus courantes : dichotomie, méthode de Newton, méthodes de la sécante et les méthodes itératives de point fixe. Le chapitre 2 traite de l'approximation polynomiale, en mettant l'accent sur l'interpolation. On y aborde également l'approximation uniforme (théorème de Weierstrass) et l'approximation quadratique (théorie des polynômes orthogonaux). Le chapitre 3 traite des méthodes d'intégration numérique des fonctions d'une variable réelle sur un segment. On étudie les méthodes de Newton–Cotes usuelles (rectangles, trapèzes, Simpson) et la méthode de Gauss–Legendre. Enfin, le chapitre 4 traite des méthodes à un pas pour la résolution des équation différentielles d'ordre 1, en dimension 1.

L'analyse numérique matricielle, les schémas numériques pour les EDP, et l'optimisation ne seront pas abordés dans ce cours. Ces sujets sont étudiés en master de mathématiques appliquées à CY Cergy Paris Université.

En plus des séances de CM et de TD, des séances de TP sont prévues pour implémenter les méthodes vues en cours. Le langage de programmation utilisé est Python.

Évaluation

L'évaluation de cette UE comporte un examen et du contrôle continu. Le contrôle continu sera composé d'un devoir écrit et d'un TP noté. La note de contrôle continu est la moyenne du devoir écrit et du TP. La note finale est le maximum entre l'examen seul, et la moyenne pondérée entre l'examen (coeff. 2) et le contrôle continu (coeff. 1) :

$$\text{Note finale} = \max\Big(\text{CT}; \tfrac{2}{3}\,\text{CT} + \tfrac{1}{3}\,\text{CC}\Big),$$

où CT (contrôle terminal) est la note de l'examen et CC le contrôle continu.

En session 1, l'examen est un devoir écrit. En session 2, ce sera un devoir écrit ou une interrogation orale, selon l'effectif des concerné·es.

Organisation de ce document de cours

Le contenu du cours est divisé en différents items numérotés, répartis en 7 catégories : *définition*, *remarque*, *exemple*, *proposition*, *théorème*, *lemme* et *corollaire*. Rappelons le sens de ces quatre derniers termes :

- **Proposition :** nom générique pour désigner un résultat mathématique démontré.
- **Théorème :** proposition importante, centrale, d'une théorie mathématique.
- **Lemme :** proposition, souvent technique, qui vise à préparer la démonstration d'une autre proposition ou d'un théorème.
- **Corollaire :** proposition qui est la conséquence directe d'une proposition ou d'un théorème plus général ou plus important.

Les différents items de ce cours sont numérotés selon le chapitre et la section auxquels ils appartiennent. Par exemple, l'item « Proposition 2.4.8 » est une proposition se trouvant au chapitre 2, section 4, et constitue le 8^{e} item de cette section.

Expliquons enfin la signification des deux symboles □ et ◄ employés dans ce document :

- Le symbole □ marque la fin d'une démonstration. C'est un symbole standard qui remplace le sigle CQFD (« ce qu'il fallait démontrer ») traditionnellement utilisé pour indiquer la fin d'une démonstration.
- Le symbole ◄ marque la fin d'un item. Ce n'est pas un symbole standard. Dans ce cours, il sert à mieux délimiter les différents items et faciliter la lecture.

La mise en page de ce cours est inspirée du livre numérique *An Infinite Descent into Pure Mathematics* de Clive Newstead.

Références

Éric CANON : *Analyse numérique*. Vuibert, 2012.

Jean-Pierre DEMAILLY : *Analyse numérique et équations différentielles*. Collection Grenoble Sciences. EDP sciences, 4^{e} édition, 2016.

Alain YGER et Jacques-Arthur WEIL : *Mathématiques appliquées L3*. Pearson éducation France, 2009.

Chapitre 1

Résolution approchée d'une équation

Dans ce chapitre, on s'intéresse à la résolution approchée de l'équation $f(x) = 0$, où f est une fonction continue d'une variable réelle. Ceci inclut toute équation de la forme $f(x) = g(x)$ où f et g sont continues, car cette dernière est équivalente à résoudre $f(x) - g(x) = 0$. De nombreux théorèmes d'analyse réelle affirment l'existence d'un x_* satisfaisant une équation (par exemple le théorème des valeurs intermédiaires, ou le théorème de Rolle), mais sans donner d'expression pour x_*. Sauf cas particulier (par exemple les équations polynômiales de degré inférieur à 4), il n'existe pas de formule close permettant de calculer les solutions de $f(x) = 0$ pour une fonction f générique. On doit donc se tourner vers des méthodes numériques pour calculer des valeurs approchées des solutions.

Soient I un intervalle de $\mathbb{R}$ et $f\colon I \to \mathbb{R}$ une fonction. Dans tout le chapitre, on supposera que f est continue sur I. Si x_* est un zéro de f dans I, notre objectif est de construire une suite $(x_n)_{n\in\mathbb{N}}$ qui converge rapidement vers x_*.

1.1 Méthode de la dichotomie

Commençons par un résultat d'existence d'une solution de $f(x) = 0$.

Théorème 1.1.1. Soit $f\colon [a,b] \to \mathbb{R}$ continue tel que $f(a)f(b) < 0$. Alors il existe $c \in]a,b[$ tel que $f(c) = 0$.

Démonstration. L'hypothèse $f(a)f(b) < 0$ signifie que $f(a)$ et $f(b)$ sont de signes opposés, donc 0 est compris entre $f(a)$ et $f(b)$. On applique le théorème des valeurs intermédiaires. □

Soit $f\colon [a,b] \to \mathbb{R}$ une fonction continue telle que $f(a)f(b) < 0$. D'après le théorème précédent, f possède un zéro dans l'intervalle $[a,b]$. L'idée de la dichotomie est de calculer $f(m)$, l'image du milieu de l'intervalle $m = \frac{a+b}{2}$, et de comparer son signe avec celui de $f(a)$ et $f(b)$. Si $f(m) = 0$ (très improbable), alors on a trouvé ce qu'on cherchait. Sinon, on a $f(m)f(a) < 0$ ou $f(m)f(b) < 0$: dans le premier cas, f possède un zéro dans l'intervalle $[a,m]$; dans le second cas, f possède un zéro dans l'intervalle $[m,b]$. On recommence la procédure sur le nouvel intervalle. À chaque étape, la longueur de l'intervalle contenant un zéro de f est divisée par 2.

La **méthode de la dichotomie** définit trois suites $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ et $(x_n)_{n\in\mathbb{N}}$ par récurrence.

1. On initialise $a_0 := a$, $b_0 := b$ et $x_0 := \frac{a+b}{2}$.
2. Pour tout $n \geq 0$:
   - Si $f(x_n) = 0$, on pose $a_{n+1} = b_{n+1} = x_{n+1} := x_n$. (En pratique, on a trouvé un zéro, on arrête l'algorithme.)
   - Si $f(x_n)f(a_n) < 0$, on pose $a_{n+1} := a_n$ et $b_{n+1} := x_n$.
   - Sinon, on pose $a_{n+1} := x_n$ et $b_{n+1} := b_n$.

   Dans tous les cas, on pose $x_{n+1} := \frac{a_{n+1}+b_{n+1}}{2}$.

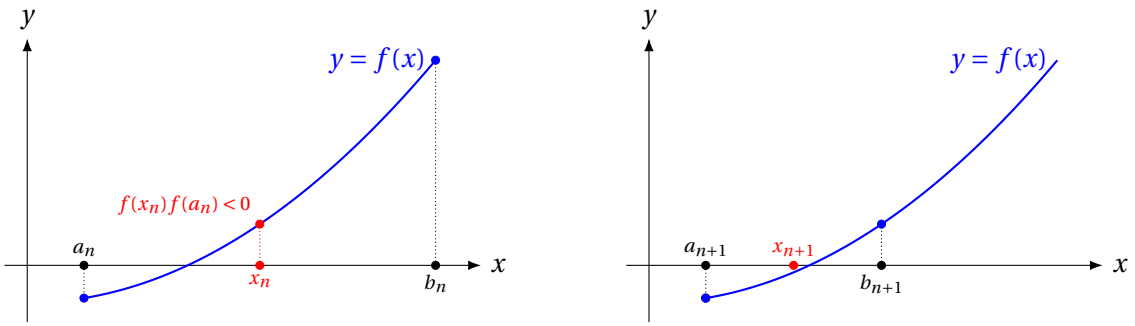


Figure 1.1 – Une itération de la méthode de la dichotomie

Les suites $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ sont adjacentes, donc elles convergent vers une même limite que l'on note x_*. Par encadrement, $(x_n)_{n\in\mathbb{N}}$ converge également vers x_*.

Par continuité de f, puisque $f(a_n)f(b_n) \le 0$ pour tout n, on a $f(x_*)^2 \le 0$, donc $f(x_*) = 0$. Enfin, on montre par récurrence que :

$$\forall n \in \mathbb{N}, \quad b_n - a_n = \frac{b-a}{2^n}.$$

Puisque $x_n, x_* \in [a_n, b_n]$, on en déduit que $|x_n - x_*| \le \frac{b-a}{2^n}$.

Critère d'arrêt. On ne peut pas utiliser directement la condition $|x_n - x_*| < \varepsilon$ comme critère d'arrêt, puisqu'on ne connait pas x_*. Mais on sait que $|x_n - x_*| < \frac{b-a}{2^n}$, donc si on veut que $|x_n - x_*| < \varepsilon$, il suffit de choisir n tel que $\frac{b-a}{2^n} \le \varepsilon$, c'est-à-dire :

$$n \ge \log_2\left(\frac{b-a}{\varepsilon}\right),$$

où $\log_2$ désigne le logarithme binaire[1]. Avec $n(\varepsilon) := \left\lceil \log_2\left(\frac{b-a}{\varepsilon}\right)\right\rceil$ itérations de la méthode de la dichotomie, on garantit que $|x_{n(\varepsilon)} - x_*| \le \varepsilon$. Un autre critère est d'itérer jusqu'à ce que $|f(x_n)| \le \varepsilon$.

Remarque 1.1.2. La méthode de la dichotomie a l'avantage de toujours converger vers une solution. De plus, la méthode produit un encadrement : à la n-ième itération, on sait que l'intervalle $[a_n, b_n]$ contient une solution. Enfin, cet intervalle est de longueur $\frac{b-a}{2^n}$, ce qui fournit une majoration explicite de l'erreur. ◀

1.2 Vitesse de convergence

Définition 1.2.1. Soit $(x_n)_{n\in\mathbb{N}}$ une suite qui converge vers une limite x_*. On note $e_n = |x_n - x_*|$. On dit que la convergence est :

- **linéaire** s'il existe $C \in [0,1[$ tel qu'à partir d'un certain rang, $e_{n+1} \le Ce_n$.
- **superlinéaire** si $e_{n+1} = o(e_n)$.
- **d'ordre** $p > 1$ si $e_{n+1} = O(e_n^p)$. Pour $p = 2$, on parle de **convergence quadratique**.

Remarque 1.2.2. Ces définitions donnent une convergence au moins linéaire ou d'ordre p. Une autre convention, plus restrictive, consiste à demander l'existence d'une limite :

$$\frac{e_{n+1}}{e_n^p} \xrightarrow[n\to+\infty]{} c \in]0, +\infty[,$$

ce qui définit l'ordre exact de convergence ainsi que sa constante asymptotique c. L'existence d'une telle limite implique les définitions ci-dessus, mais la réciproque est fausse en général. ◀

1. on note $\log_2 : \mathbb{R}_+^* \to \mathbb{R}$ le logarithme de base 2, défini par $\log_2(x) = \frac{\ln(x)}{\ln(2)}$. C'est l'application réciproque de la fonction exponentielle de base 2 : $x \mapsto 2^x := e^{x\ln(2)}$.

Remarque 1.2.3. L'ordre de convergence p contrôle la vitesse à laquelle le nombre de décimales correctes augmente. Notons $d_n = -\log_{10}(e_n)$ le nombre de décimales exactes de x_n. Si $e_{n+1} = O(e_n^p)$, c'est-à-dire $e_{n+1} \leq C\, e_n^p$ à partir d'un certain rang, alors en passant au logarithme :

$$d_{n+1} = -\log_{10}(e_{n+1}) \geq p\, d_n - \log_{10}(C).$$

Ainsi, à une constante additive près, le nombre de décimales correctes est au moins multiplié par p à chaque itération. En particulier :

- Pour une convergence linéaire de constante C, le nombre de décimales correctes croît au moins linéairement : on gagne au moins $-\log_{10}(C)$ décimales par itération.
- Pour une convergence quadratique, le nombre de décimales correctes double à chaque itération : si x_n est correct à d décimales, x_{n+1} l'est au moins à environ $2d$ décimales. ◂

Exemple 1.2.4. La méthode de la dichotomie a une vitesse de convergence linéaire. La méthode de Newton étudiée en section 1.4 a quant à elle une vitesse de convergence quadratique, comme on le verra. ◂

1.3 Méthodes itératives

Soit $f : I \to \mathbb{R}$ une fonction dont on cherche un zéro, c'est-à-dire un réel $x_* \in I$ tel que $f(x_*) = 0$. En posant $F(x) = x - \lambda(x) f(x)$ où λ est une fonction qui ne s'annule pas sur I (par exemple $\lambda := 1$, ou encore $\lambda := 1/f'$ dans la méthode de Newton, voir section 1.4), on a :

$$F(x_*) = x_* \iff \lambda(x_*) f(x_*) = 0 \iff f(x_*) = 0.$$

Ainsi, x_* est un zéro de f si et seulement si x_* est un **point fixe** de F. Le choix de la fonction λ (donc de F) est déterminant : il conditionne la vitesse de convergence, voire l'existence même d'une suite convergente construite par itération de F, comme on va le voir dans la suite.
Ceci motive l'étude générale de la recherche de points fixes d'une fonction F, que l'on va approcher par la suite des itérées.

Définition 1.3.1. Soit I un intervalle stable par F, c'est-à-dire tel que $F(I) \subset I$. On appelle **suite des itérées** par F une suite $(x_n)_{n\in\mathbb{N}}$ définie par :

$$\begin{cases} x_{n+1} = F(x_n) \\ x_0 \in I \text{ donné.} \end{cases} \tag{1.1}$$

Remarque 1.3.2. La fonction F sera supposée définie sur un intervalle I et à valeurs dans ce même intervalle I. Cette condition assure que la suite récurrente $(x_n)_{n\in\mathbb{N}}$ est bien définie. En effet, si F n'est pas à valeurs dans I, il pourrait exister $n \in \mathbb{N}$ tel que $x_{n+1} = F(x_n) \notin I$, et $x_{n+2} = F(x_{n+1})$ ne serait pas défini. ◂

1.3.a Étude des suites itératives

L'idée d'itérer la fonction F pour approcher le point fixe vient de la proposition suivante.

Proposition 1.3.3. Soit $(x_n)_{n\in\mathbb{N}}$ définie par (1.1) avec $F : I \to I$ continue. Si $(x_n)_{n\in\mathbb{N}}$ converge vers une limite $x_* \in I$, alors x_* est un point fixe de F. ◂

Démonstration. Supposons que F est continue et que $x_n \to x_*$. La suite $(x_n)_{n\in\mathbb{N}}$ vérifie :

$$\forall n \in \mathbb{N}, \quad x_{n+1} = F(x_n).$$

Le membre de gauche converge vers x_*, et le membre de droite converge vers $F(x_*)$ par continuité de F en x_*. Par unicité de la limite, $x_* = F(x_*)$. □

Remarque 1.3.4.

1. Attention, la proposition précédente n'affirme pas que la suite $(x_n)_{n\in\mathbb{N}}$ est toujours convergente.
2. Il faut que la limite x_* appartienne à I pour appliquer le résultat. Cette condition est automatiquement satisfaite si I est un intervalle fermé. ◀

Pour montrer la convergence de $(x_n)_{n\in\mathbb{N}}$, on pourra parfois s'appuyer sur le résultat suivant afin de montrer que $(x_n)_{n\in\mathbb{N}}$ est monotone.

Proposition 1.3.5. Soient $F\colon I \to I$ et $(x_n)_{n\in\mathbb{N}}$ définie par (1.1).

(i) Si F est croissante sur I, alors $(x_n)_{n\in\mathbb{N}}$ est monotone et son sens de variation est déterminé par le signe de $F(x_0) - x_0$.

(ii) Si F est décroissante sur I, alors les suites $(x_{2n})_{n\in\mathbb{N}}$ et $(x_{2n+1})_{n\in\mathbb{N}}$ sont monotones, et de sens de variation contraires. ◀

Démonstration.
(i) Supposons F croissante. Montrons que le signe de $x_{n+1} - x_n$ est constant. Par croissance de F, $x_{n+1} - x_n = F(x_n) - F(x_{n-1})$ a le même signe que $x_n - x_{n-1}$. Par récurrence, on obtient que $x_{n+1} - x_n$ a le même signe que $x_1 - x_0 = F(x_0) - x_0$, d'où le résultat.

(ii) Supposons F décroissante. En notant $y_n := x_{2n}$ et $z_n := x_{2n+1}$, les suites $(y_n)_{n\in\mathbb{N}}$ et $(z_n)_{n\in\mathbb{N}}$ satisfont la relation de récurrence : $y_{n+1} = (F\circ F)(y_n)$ et $z_{n+1} = (F\circ F)(z_n)$, avec $F\circ F$ croissante. D'après le résultat précédent, ces suites sont donc monotones. Enfin, on a :

$$z_1 - z_0 = x_3 - x_1 = F(x_2) - F(x_0) = F(y_1) - F(y_0),$$

donc par décroissance de F, $z_1 - z_0$ et $y_1 - y_0$ sont de signes opposés. Par conséquent, les suites $(y_n)_{n\in\mathbb{N}}$ et $(z_n)_{n\in\mathbb{N}}$ sont de sens de variation contraires. □

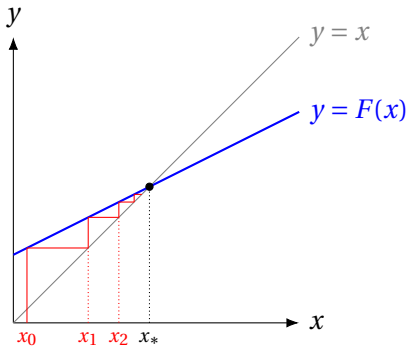


Figure 1.2 – F croissante

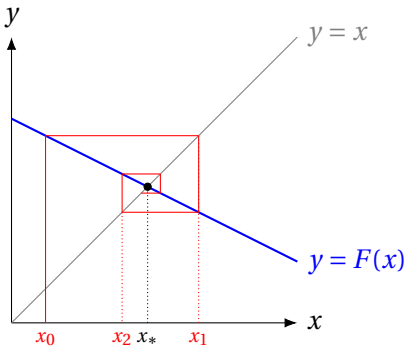


Figure 1.3 – F décroissante

1.3.b Théorème de Banach–Picard

On cherche des conditions qui assurent la convergence de la suite des itérés.

Définition 1.3.6. Soit $F\colon I \to \mathbb{R}$. On dit que F est **contractante** si :

$$\exists K \in [0,1[,\ \forall x, y \in I,\ \big|F(x) - F(y)\big| \leq K|x - y|.$$

Remarque 1.3.7. Une application contractante est une application K**–lipschitzienne** avec $K < 1$. En particulier, elle est uniformément continue sur I. ◀

Proposition 1.3.8. Si F est dérivable sur I, alors :

$$F \text{ est contractante} \iff \sup_{t\in I}\left|F'(t)\right| < 1. \qquad \blacktriangleleft$$

Démonstration. Soit $F\colon I \to I$ dérivable.
($\Longrightarrow$) Supposons que F est contractante. Alors pour tous $x, y \in I$ distincts :

$$\left|\frac{F(x)-F(y)}{x-y}\right| \leq K.$$

En faisant tendre $y \to x$, on obtient que : $\forall x \in I,\ \left|F'(x)\right| \leq K$. Ainsi, on a montré que :

$$\sup_{t\in I}\left|F'(t)\right| \leq K < 1.$$

($\Longleftarrow$) Supposons que $\sup_{t\in I}\left|F'(t)\right| < 1$. On pose $K = \sup_{t\in I}|F'(t)|$, on a bien $K \in [0,1[$ par hypothèse. Alors l'inégalité des accroissements finis affirme que :

$$\forall x, y \in I, \quad \left|F(x)-F(y)\right| \leq K|x-y|.$$

Donc F est contractante. □

Théorème 1.3.9 (point fixe de Banach–Picard)**.** Soit I un intervalle fermé de $\mathbb{R}$. Si $F\colon I \to I$ est contractante, alors F possède un unique point fixe x_* dans I. De plus, pour tout $x_0 \in I$, la suite des itérées $(x_n)_{n\in\mathbb{N}}$ converge vers x_* et :

$$|x_n - x_*| \leq \frac{K^n}{1-K}|x_1 - x_0|.$$

Remarque 1.3.10. Pour utiliser ce théorème, il faut trouver un intervalle fermé I stable par F, c'est-à-dire $F(I) \subset I$, et tel que F soit contractante sur I. D'après la proposition 1.3.8, dans le cas où F est dérivable, cela revient à $\sup_{t\in I}|F'(t)| < 1$. ◀

Démonstration. On suppose que I est fermé et que $F\colon I \to I$ est contractante. On considère $(x_n)_{n\in\mathbb{N}}$ définie par (1.1).
• *Unicité.* Soient $x_*, y_* \in I$ des points fixes de F. Alors :

$$|x_* - y_*| = |F(x_*) - F(y_*)| \leq K|x_* - y_*|,$$

donc :

$$(1-K)|x_* - y_*| \leq 0.$$

Puisque $1-K > 0$, on a $|x_* - y_*| = 0$, c'est-à-dire $x_* = y_*$.

• *Existence.* Montrons que la suite $(x_n)_{n\in\mathbb{N}}$ est une suite de Cauchy. Pour tous $n, p \in \mathbb{N}$, on a :

$$|x_{n+1} - x_{n+p+1}| = |F(x_n) - F(x_{n+p})| \leq K|x_n - x_{n+p}|.$$

Par récurrence sur n, on en déduit que :

$$\forall n, p \in \mathbb{N}, \quad |x_n - x_{n+p}| \leq K^n|x_0 - x_p|. \tag{1.2}$$

Majorons la distance $|x_0 - x_p|$:

$$\begin{aligned}
|x_0 - x_p| &\leq \sum_{i=0}^{p-1}|x_i - x_{i+1}| && \text{inégalité triangulaire}\\
&\leq \sum_{i=0}^{p-1} K^i|x_0 - x_1| && \text{d'après (1.2)}\\
&\leq \frac{1-K^p}{1-K}|x_0 - x_1| && \text{somme géométrique}\\
&\leq \frac{1}{1-K}|x_0 - x_1|.
\end{aligned}$$

Finalement, on a montré que :

$$\forall n, p \in \mathbb{N}, \quad |x_n - x_{n+p}| \leq \frac{K^n}{1-K}|x_0 - x_1| \tag{1.3}$$

Le membre de droite tend vers 0 quand $n \to +\infty$, uniformément en p, d'où :

$$\lim_{n\to+\infty} \sup_{p\in\mathbb{N}} |x_n - x_{n+p}| = 0.$$

La suite $(x_n)_{n\in\mathbb{N}}$ est donc de Cauchy. Par conséquent, elle converge dans $\mathbb{R}$ vers une limite x_*. De plus, les termes de la suite $(x_n)_{n\in\mathbb{N}}$ appartiennent à I qui est fermé, donc $x_* \in I$. D'après la proposition 1.3.3, x_* est un point fixe de F.

• *Majoration de l'erreur.* En faisant tendre $p \to +\infty$ dans (1.3), on a :

$$|x_n - x_*| \leq \frac{K^n}{1-K}|x_0 - x_1|. \qquad \square$$

Critères d'arrêt

1. On sait que l'erreur commise est majorée par :

   $$|x_n - x_*| \leq \frac{K^n}{1-K}|x_0 - x_1|.$$

   Pour avoir $|x_n - x_*| \leq \varepsilon$, il suffit de choisir n tel que $\frac{K^n}{1-K}|x_0 - x_1| \leq \varepsilon$, c'est-à-dire :

   $$n \geq \frac{\ln(\varepsilon) + \ln(1-K) - \ln(|x_0 - x_1|)}{\ln(K)}. \tag{1.4}$$

2. Un autre critère est d'itérer jusqu'à ce que $|x_{n+1} - x_n|$ soit assez petit. Plus précisément, on a la majoration :

   $$|x_n - x_{n+p}| \leq \sum_{i=0}^{p-1} |x_{n+i} - x_{n+i+1}| \leq \sum_{i=0}^{p-1} K^i |x_n - x_{n+1}| \leq \frac{1-K^p}{1-K}|x_n - x_{n+1}|.$$

   Pour $p \to +\infty$, on obtient :

   $$|x_n - x_*| \leq \frac{1}{1-K}|x_n - x_{n+1}|.$$

   Par conséquent, si on itère jusqu'à ce que $|x_{n+1} - x_n| \leq (1-K)\varepsilon$, alors on aura $|x_n - x_*| \leq \varepsilon$.

Dans les deux cas, plus la constante K est proche de 1, plus le nombre d'itérations devra être élevé.

Exemple 1.3.11 (Cosinus itéré)**.** On considère $F(x) = \cos(x)$ sur $I = [0,1]$. Puisque cos est décroissante sur I, on a :

$$F(I) = [\cos(1), \cos(0)] = [\cos(1), 1] \subset [0,1] = I,$$

car $\cos(1) > 0$. L'intervalle fermé I est donc stable par F. De plus, F est dérivable sur I, de dérivée $F'(x) = -\sin(x)$. Comme sin est positive et croissante sur I :

$$K = \sup_{x\in I} |F'(x)| = \sin(1) \approx 0{,}84 < 1.$$

D'après la proposition 1.3.8, F est donc contractante sur I, de constante $K = \sin(1)$. Le théorème de Banach–Picard s'applique : F possède un unique point fixe $x_* \in I$ et pour tout $x_0 \in I$, la suite $(x_n)_{n\in\mathbb{N}}$ définie par $x_{n+1} = \cos(x_n)$ converge vers x_*.
La constante K étant plutôt proche de 1, la convergence est assez lente. En partant de $x_0 = 1$, pour obtenir une approximation de x_* à 10^{-3} près, on calcule 47 itérations (d'après (1.4)). On obtient alors $x_* \approx 0{,}739$. ◂

1.3.c Étude locale

Théorème 1.3.12. Supposons que F soit de classe $\mathscr{C}^1$ sur I et soit x_* un point fixe de F.

(i) Si $|F'(x_*)| < 1$, alors il existe $J \subset I$ un voisinage fermé de x_* tel que $F(J) \subset J$ et $F_{|J}$ est contractante. Par conséquent, pour tout $x_0 \in J$, la suite des itérées $(x_n)_{n\in\mathbb{N}}$ converge vers x_*.

(ii) Si $|F'(x_*)| > 1$, alors il existe $J \subset I$ un voisinage de x_* tel que, pour tout $x_0 \in J \setminus \{x_*\}$, il existe $n \in \mathbb{N}$ tel que $x_n \notin J$. Par conséquent, la suite des itérées $(x_n)_{n\in\mathbb{N}}$ ne peut converger vers x_* que si elle est stationnaire égale à x_*.

Démonstration.
(i) Puisque F est de classe $\mathscr{C}^1$ et $|F'(x_*)| < 1$, par continuité de F' en x_*, il existe $\delta > 0$ tel que $J := [x_* - \delta, x_* + \delta] \subset I$ et

$$K := \sup_{t\in J} |F'(t)| < 1.$$

D'après la proposition 1.3.8, $F_{|J}$ est contractante. De plus, pour tout $x \in J$:

$$|F(x) - x_*| = |F(x) - F(x_*)| \leq K|x - x_*| \leq K\delta \leq \delta,$$

donc $F(x) \in J$, c'est-à-dire $F(J) \subset J$. Ainsi J est un intervalle fermé stable par F, sur lequel F est contractante : le théorème de point fixe de Banach–Picard s'applique, et donne que pour tout $x_0 \in J$, la suite des itérées $(x_n)_{n\in\mathbb{N}}$ converge vers l'unique point fixe de F dans J, qui est x_*.

(ii) De même, par continuité de F' en x_*, il existe $\delta > 0$ tel que $J := [x_* - \delta, x_* + \delta] \subset I$ et

$$K := \inf_{t\in J} |F'(t)| > 1.$$

Soit $x_0 \in J \setminus \{x_*\}$. Supposons par l'absurde que $x_n \in J$ pour tout $n \in \mathbb{N}$. Pour chaque n, l'inégalité des accroissements finis appliquée entre x_n et x_* donne :

$$|x_{n+1} - x_*| = |F(x_n) - F(x_*)| \geq K|x_n - x_*|,$$

d'où, par récurrence :

$$\forall n \in \mathbb{N}, \quad |x_n - x_*| \geq K^n |x_0 - x_*|.$$

Comme $K > 1$ et $x_0 \neq x_*$, le membre de droite tend vers $+\infty$, ce qui contredit $|x_n - x_*| \leq \delta$ pour tout n. Il existe donc $n \in \mathbb{N}$ tel que $x_n \notin J$.
En particulier, si la suite $(x_n)_{n\in\mathbb{N}}$ convergeait vers x_* sans être stationnaire égale à x_*, tous ses termes appartiendraient à J à partir d'un certain rang, ce qui contredit ce qui précède. La suite ne peut donc converger vers x_* que si elle est stationnaire égale à x_*. □

Définition 1.3.13. Sous les hypothèses du théorème précédent, un point fixe x_* est dit :

- **attractif** si $|F'(x_*)| < 1$,
- **superattractif** si $F'(x_*) = 0$,
- **répulsif** si $|F'(x_*)| > 1$.

Exemple 1.3.14. Soit $F(x) = x^2$ sur $I = \mathbb{R}$. C'est une fonction de classe $\mathscr{C}^1$ sur I avec $F'(x) = 2x$. Les points fixes de F sont les solutions de $x^2 = x$, c'est-à-dire 0 et 1.
En $x_* = 0$, on a $F'(0) = 0$, donc 0 est un point fixe superattractif : pour x_0 proche de 0, la suite $(x_n)_{n\in\mathbb{N}}$ converge (très rapidement) vers 0.
En $x_* = 1$, on a $|F'(1)| = 2 > 1$, donc 1 est un point fixe répulsif. En effet, on montre par récurrence que $x_n = x_0^{2^n}$. On a donc :

$$\lim_{n\to+\infty} x_n = \begin{cases} +\infty & \text{si } x_0 > 1 \\ 1 & \text{si } x_0 = 1 \\ 0 & \text{si } 0 \leq x_0 < 1 \end{cases}$$

Ceci illustre bien la conclusion du théorème précédent : la suite des itérées (autres que la suite stationnaire en x_*) s'échappe de tout voisinage de x_*. ◂

Dans le cas où $F'(x_*) = 0$, la démonstration du théorème précédent montre que F est contractante au voisinage de x_*, de constante K aussi petite que l'on veut. La vitesse de convergence de $(x_n)_{n\in\mathbb{N}}$ est donc superlinéaire. Si on suppose que F est $\mathscr{C}^2$, on montre en fait que la convergence est au moins d'ordre 2.

Théorème 1.3.15. Supposons que F soit $\mathscr{C}^2$ sur I, que $F'(x_*) = 0$ (x_* est superattractif) et que $|F''| \leq M$ sur I. Alors il existe un voisinage $J \subset I$ de x_* tel que $F(J) \subset J$, et pour tout $x_0 \in J$, la suite des itérées $(x_n)_{n\in\mathbb{N}}$ est bien définie et converge vers x_* avec une convergence (au moins) quadratique. De plus, si $F''(x_*) \neq 0$, la convergence est exactement d'ordre 2.

Démonstration. Choisissons $\delta > 0$ tel que $J := [x_* - \delta, x_* + \delta] \subset I$ et $\delta < \frac{2}{M}$.
Montrons d'abord que $F(J) \subset J$. Soit $x \in J$. Par la formule de Taylor–Lagrange en x_*, il existe ξ compris entre x_* et x tel que :

$$F(x) = \underbrace{F(x_*)}_{=x_*} + \underbrace{F'(x_*)}_{=0}(x - x_*) + \frac{1}{2}F''(\xi)(x - x_*)^2,$$

d'où :

$$|F(x) - x_*| \leq \frac{M}{2}|x - x_*|^2 \leq \frac{M}{2}\delta^2 \leq \delta,$$

puisque $M\delta < 2$. Ainsi $F(x) \in J$, ce qui montre $F(J) \subset J$.
Soit maintenant $x_0 \in J$. Puisque J est stable par F, la suite $(x_n)_{n\in\mathbb{N}}$ est bien définie et reste dans J. Le calcul précédent, appliqué à $x = x_n$, fournit ξ_n compris entre x_* et x_n tel que :

$$x_{n+1} - x_* = F(x_n) - x_* = \frac{1}{2}F''(\xi_n)(x_n - x_*)^2,$$

donc :

$$|x_{n+1} - x_*| \leq \frac{M}{2}|x_n - x_*|^2.$$

En multipliant par $\frac{M}{2}$, on obtient :

$$\frac{M}{2}|x_{n+1} - x_*| \leq \left(\frac{M}{2}|x_n - x_*|\right)^2.$$

Comme $\frac{M\delta}{2} < 1$, une récurrence immédiate donne :

$$|x_n - x_*| \leq \frac{2}{M}\left(\frac{M}{2}|x_0 - x_*|\right)^{2^n} \leq \frac{2}{M}\left(\frac{M\delta}{2}\right)^{2^n} \xrightarrow[n\to+\infty]{} 0,$$

donc $(x_n)_{n\in\mathbb{N}}$ converge vers x_* et cette convergence est (au moins) d'ordre 2.
Enfin, si $F''(x_*) \neq 0$, comme $\xi_n \to x_*$ par encadrement et F'' est continue, on a $F''(\xi_n) \to F''(x_*)$, donc :

$$|x_{n+1} - x_*| \underset{n\to+\infty}{\sim} \frac{|F''(x_*)|}{2}|x_n - x_*|^2,$$

ce qui montre que la convergence est exactement d'ordre 2. □

Critère d'arrêt D'après la démonstration ci-dessus :

$$|x_n - x_*| \leq \frac{2}{M}\left(\frac{M}{2}|x_0 - x_*|\right)^{2^n}.$$

Si on dispose d'un encadrement $a \leq x_* \leq b$ avec $b - a < \frac{2}{M}$, alors en choisissant $x_0 \in [a, b]$, on a :

$$|x_n - x_*| \leq \frac{2}{M}\left(\frac{M}{2}(b - a)\right)^{2^n}.$$

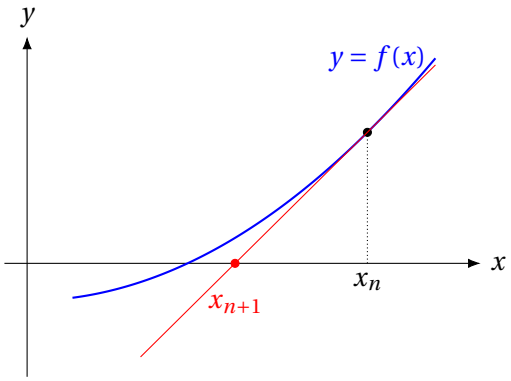


Figure 1.4 – Une itération de la méthode de Newton

Pour avoir $|x_n - x_*| \leq \varepsilon$, il suffit de choisir n assez grand pour que $\frac{2}{M}\left(\frac{M}{2}(b-a)\right)^{2^n} \leq \varepsilon$, c'est-à-dire :

$$n \geq \log_2\left(\frac{\ln\left(\frac{M\varepsilon}{2}\right)}{\ln\left(\frac{M(b-a)}{2}\right)}\right). \tag{1.5}$$

Exemple 1.3.16 (Cas où $|F'(x_*)| = 1$)**.** Dans le cas où $|F'(x_*)| = 1$, on ne peut rien conclure :

1. $F(x) = \sin(x)$, $x_* = 0$ et $x_0 \in [0, \frac{\pi}{2}]$. Puisque $\sin(x) \leq x$ sur $[0, \frac{\pi}{2}]$, la suite $(x_n)_{n\in\mathbb{N}}$ est décroissante et minorée, donc elle converge vers l'unique point fixe de sin, c'est-à-dire 0.
2. $F(x) = \mathrm{sh}(x)$, $x_* = 0$ et $x_0 > 0$. Pour tout $x \geq 0$, $\mathrm{sh}(x) \geq x$ donc la suite $(x_n)_{n\in\mathbb{N}}$ est croissante. Le seul point fixe de sh étant 0, la suite $(x_n)_{n\in\mathbb{N}}$ ne peut pas converger, donc elle diverge vers $+\infty$. ◀

Remarque 1.3.17. Dans le cas $|F'(x_*)| = 1$, même s'il y a convergence, on peut montrer que celle-ci est trop lente pour que ce soit intéressant numériquement. ◀

1.4 Méthode de Newton

Supposons qu'on dispose d'une approximation x_0 d'un zéro x_* de f. Le développement limité de f en x_0 s'écrit :

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + o(x - x_0).$$

En négligeant le reste, résoudre $f(x) = 0$ revient approximativement à résoudre :

$$f(x_0) + f'(x_0)(x - x_0) = 0,$$

dont la solution x_1 est (en supposant que $f'(x_0) \neq 0$) :

$$x_1 := x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Géométriquement, on cherche le point d'intersection de la tangente en x_0 avec l'axe des abscisses. Cette valeur x_1 est en général une meilleure approximation que x_0. On recommence ensuite à partir de la tangente en x_1, etc.

Définition 1.4.1. La **méthode de Newton** est la méthode itérative définie par :

$$\begin{cases} x_{n+1} := x_n - \dfrac{f(x_n)}{f'(x_n)} \\ x_0 \text{ donné.} \end{cases}$$

La bonne définition de la méthode de Newton n'est pas certaine *a priori* : rien ne garantit que $f'(x_n)$ ne soit jamais nul, ni que la suite $(x_n)_{n\in\mathbb{N}}$ reste dans l'intervalle de définition de f.

Remarque 1.4.2. Si on note $F(x) = x - \frac{f(x)}{f'(x)}$, sa dérivée est :

$$F'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}.$$

Si x_* est un zéro de f mais pas de f' (zéro simple), alors $F(x_*) = x_*$ et $F'(x_*) = 0$: x_* est un point fixe superattractif de F. ◀

Lemme 1.4.3. On suppose que f est $\mathscr{C}^2$ sur I et que $f(x_*) = 0$ et que f' ne s'annule pas sur I. On note $F(x) = x - \frac{f(x)}{f'(x)}$. Pour tout $x \in I$, il existe ξ compris entre x_* et x tel que :

$$F(x) - x_* = \frac{f''(\xi)}{2f'(x)}(x - x_*)^2.$$ ◀

Démonstration. Soit $x \in I$. Par Taylor–Lagrange, il existe ξ compris entre x_* et x tel que :

$$0 = f(x_*) = f(x) + f'(x)(x_* - x) + \frac{f''(\xi)}{2}(x_* - x)^2.$$

On réarrange cette égalité :

$$\begin{aligned} 0 = f(x) + f'(x)(x_* - x) + \frac{f''(\xi)}{2}(x_* - x)^2 &\iff f'(x)(x - x_*) - f(x) = \frac{f''(\xi)}{2}(x - x_*)^2 \\ &\iff x - x_* - \frac{f(x)}{f'(x)} = \frac{f''(\xi)}{2f'(x)}(x - x_*)^2 \\ &\iff F(x) - x_* = \frac{f''(\xi)}{2f'(x)}(x - x_*)^2. \end{aligned}$$ □

Théorème 1.4.4 (convergence quadratique locale)**.** Supposons que f est de classe $\mathscr{C}^2$ sur $[a,b]$, avec $f(a)f(b) < 0$ et $f' > 0$ sur $[a,b]$. On note $F(x) = x - \frac{f(x)}{f'(x)}$.

(i) Il existe un unique $x_* \in]a,b[$ tel que $f(x_*) = 0$.

(ii) Il existe un intervalle $J \subset [a,b]$ voisinage de x_* tel que $F(J) \subset J$ et pour tout $x_0 \in J$, la suite $(x_n)_{n\in\mathbb{N}}$ de la méthode de Newton est bien définie, converge vers x_* et la convergence est au moins d'ordre 2.

Démonstration. L'existence et l'unicité de x_* découlent du TVI et de la croissance stricte de f. Puisque f' et f'' sont continues sur le compact $[a,b]$, elles sont bornées et atteignent leurs bornes. Posons :

$$M := \frac{\max\limits_{t\in[a,b]} |f''(t)|}{\min\limits_{t\in[a,b]} f'(t)}.$$

D'après le lemme 1.4.3 :

$$\forall x \in [a,b], \quad |F(x) - x_*| \leq \frac{M}{2}|x - x_*|^2.$$

Le reste est identique à la démonstration de la proposition 1.3.15. □

La convergence de la méthode de Newton est d'ordre 2, donc très rapide (voir remarque 1.2.3) : au voisinage de x_*, le nombre de chiffres exacts double à chaque itération de la méthode. Le critère d'arrêt (1.5) est valable pour la méthode de Newton, avec :

$$M := \frac{\max\limits_{t\in[a,b]} |f''(t)|}{\min\limits_{t\in[a,b]} |f'(t)|}.$$

Théorème 1.4.5 (convergence globale)**.** On suppose que f est $\mathscr{C}^2$ sur $[a,b]$, que $f'>0$, $f''>0$ sur $[a,b]$, et que $f(a)f(b)<0$. Alors pour tout $x_0 \in]x_*,b]$, la méthode de Newton est bien définie, converge en décroissant vers l'unique zéro de f dans $[a,b]$, et la convergence est exactement d'ordre 2.

Remarque 1.4.6. Le théorème est valable plus généralement si f' et f'' sont de signe constant. En effet, on se ramène aux hypothèses du théorème en posant $g(x) = \pm f(\pm x)$:

- si f est concave, alors $-f$ est convexe (mais de sens de variation contraire à celui de f) ;
- si f est décroissante, alors $x \mapsto f(-x)$ est croissante.

En combinant ces deux transformations, il est toujours possible de se ramener au cas croissant convexe. L'intervalle stable sera $[a,x_*[$ ou $]x_*,b]$ selon les cas, et la suite $(x_n)_{n\in\mathbb{N}}$ sera monotone. ◂

Démonstration. Si $x \geq x_*$, alors $F(x) = x - \frac{f(x)}{f'(x)} \leq x$, avec inégalité stricte si $x > x_*$.
Pour tout $x \in [a,b]$, d'après le lemme 1.4.3, il existe ξ tel que

$$F(x) - x_* = \frac{f''(\xi)}{2f'(x)}(x-x_*)^2 \geq 0, \tag{1.6}$$

avec inégalité stricte si $x \neq x_*$. Par conséquent, pour tout $x \in]x_*,b]$, on a :

$$x_* < F(x) < x \leq b.$$

donc $F(x) \in]x_*,b]$. L'intervalle $]x_*,b]$ est stable par F, donc pour tout $x_0 \in]x_*,b]$, la suite $(x_n)_{n\in\mathbb{N}}$ de la méthode de Newton est bien définie et vérifie :

$$x_* < x_{n+1} = F(x_n) < x_n.$$

La suite $(x_n)_{n\in\mathbb{N}}$ est strictement décroissante et minorée par x_*, donc elle converge. Sa limite est l'unique point fixe de F dans $[a,b]$, c'est-à-dire x_*.
De plus, d'après le lemme 1.4.3 appliqué à x_n, il existe ξ_n compris entre x_* et x_n tel que :

$$|x_{n+1} - x_*| = |F(x_n) - x_*| = \frac{f''(\xi_n)}{2f'(x_n)}|x_n - x_*|^2.$$

Par encadrement, $\xi_n \to x_*$ donc $\frac{f''(\xi_n)}{f'(x_n)} \to \frac{f''(x_*)}{f'(x_*)} > 0$ par continuité de f' et f''. Par conséquent :

$$|x_{n+1} - x_*| \underset{n\to+\infty}{\sim} \frac{f''(x_*)}{2f'(x_*)}|x_n - x_*|^2,$$

donc la convergence est exactement d'ordre 2. □

Remarque 1.4.7. Si $x_0 \in [a,x_*[$, alors (1.6) montre que $x_1 = F(x_0) > x_*$. Par conséquent, si $x_1 \leq b$, on est ramené au cas précédent, et la suite $(x_n)_{n\in\mathbb{N}}$ converge vers x_*. ◂

Exemple 1.4.8 (Méthode de Héron)**.** On cherche à calculer $\sqrt{a}$, où $a>0$. Pour cela, appliquons la méthode de Newton pour résoudre $x^2 - a = 0$. En notant $f(x) = x^2 - a$, la fonction d'itération de la méthode de Newton est :

$$F(x) = x - \frac{f(x)}{f'(x)} = x - \frac{x^2-a}{2x} = \frac{x + \frac{a}{x}}{2}.$$

La méthode de Newton consiste donc à itérer à partir d'une approximation x_0 :

$$x_{n+1} = \frac{x_n + \frac{a}{x_n}}{2}.$$

Cet algorithme était connu des Babyloniens et des Grecs de l'antiquité, et porte le nom de **méthode de Héron** (ou méthode babylonienne). À titre d'exemple, on a appliqué cette méthode pour calculer $\sqrt{2}$ en partant de $x_0 = 1$.

n	x_n	Erreur
0	1	10^{-1}
1	1,5	10^{-2}
2	1,416 666 666 666 666 5	10^{-3}
3	1,414 215 686 274 509 7	10^{-6}
4	1,414 213 562 374 689 9	10^{-12}
5	1,414 213 562 373 095 0	10^{-16}

Il suffit de 5 itérations pour obtenir une valeur correcte à 16 chiffres significatifs (maximum permis par l'erreur machine). ◀

1.5 Méthode de la sécante

Un problème de la méthode de Newton est que f' n'est pas forcément connue ou facilement calculable. L'idée est de remplacer la dérivée en x_n par le taux d'accroissement entre x_n et x_{n-1} :

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}.$$

Géométriquement, on a remplacé la tangente en x_n par la sécante aux points x_n et x_{n-1}, d'où le nom de cette méthode. La sécante a pour équation :

$$y = f(x_n) + \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}(x - x_n),$$

donc elle intersecte l'axe des abscisses en :

$$x_{n+1} := x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n).$$

Notons que la suite $(x_n)_{n\in\mathbb{N}}$ vérifie une relation de récurrence d'ordre 2, elle nécessite deux valeurs initiales x_0 et x_1.

Définition 1.5.1. La **méthode de la sécante** est la suite définie par :

$$\begin{cases} x_{n+1} = x_n - \dfrac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n) \\ x_0 \neq x_1 \text{ donnés} \end{cases}$$

Comme pour la méthode de Newton, la bonne définition de la méthode de sécante et sa convergence ne sont pas garanties *a priori*. On peut montrer que la méthode de la sécante est convergente sous les mêmes hypothèses que la méthode de Newton, et que la convergence est superlinéaire, mais pas d'ordre 2.

Théorème 1.5.2. On suppose que f est $\mathscr{C}^2$ sur $[a,b]$, que $f' > 0$, $f'' > 0$ sur $[a,b]$, et que $f(a)f(b) < 0$. Alors pour tous $x_0, x_1 \in]x_*, b]$ avec $x_1 < x_0$, la méthode de la sécante est bien définie et converge en décroissant vers l'unique zéro de f dans $[a,b]$.

Démonstration. Les hypothèses sur f impliquent l'existence et l'unicité de x_* dans $[a,b]$.
Montrons par récurrence la propriété H_n : « $x_* < x_{n+1} < x_n$ ». L'initialisation est vraie par hypothèse. Soit $n \geq 1$, supposons que H_{n-1} est vraie. Puisque f est strictement croissante et que $x_* < x_n < x_{n-1}$, on a :

$$\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} > 0, \qquad f(x_n) > f(x_*) = 0.$$

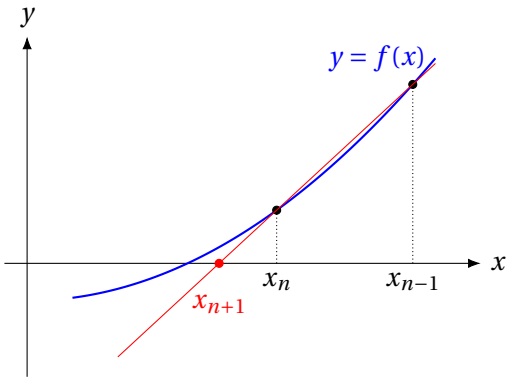


Figure 1.5 – Une itération de la méthode de la sécante

Par conséquent :

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n) < x_n.$$

Montrons maintenant que $x_{n+1} > x_*$. D'après le théorème des accroissements finis appliqué à f entre x_n et x_{n-1}, il existe $\xi_n \in]x_n, x_{n-1}[$ tel que :

$$\frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} = f'(\xi_n).$$

De même, appliqué entre x_* et x_n, il existe $\eta_n \in]x_*, x_n[$ tel que :

$$f(x_n) = f(x_n) - f(x_*) = f'(\eta_n)(x_n - x_*).$$

Puisque $x_* < \eta_n < x_n < \xi_n < x_{n-1}$ et que $f'' > 0$ sur $[a, b]$, la dérivée f' est strictement croissante, donc $0 < f'(\eta_n) < f'(\xi_n)$. Par conséquent :

$$x_{n+1} - x_* = (x_n - x_*) - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n) = (x_n - x_*)\left(1 - \frac{f'(\eta_n)}{f'(\xi_n)}\right) > 0,$$

d'où $x_{n+1} > x_*$. Ceci achève la démonstration de H_n.
Par récurrence, on a montré que $\forall n \in \mathbb{N}$, $x_* < x_{n+1} < x_n$. La suite $(x_n)_{n\in\mathbb{N}}$ est décroissante et minorée, donc elle converge et sa limite $\ell \in [x_*, b]$ vérifie :

$$\ell = \ell - \frac{f(\ell)}{f'(\ell)},$$

c'est-à-dire $f(\ell) = 0$. L'unique zéro de f dans $[a, b]$ étant x_*, on a $\ell = x_*$. □

Chapitre 2

Approximation polynomiale

Soit I un intervalle et $f : I \to \mathbb{R}$ une fonction continue. L'objectif de l'approximation polynomiale est de trouver un polynôme qui soit proche de f, en un sens à préciser. L'approximation polynomiale est à la base de nombreuses méthodes en analyse numérique. L'idée est de remplacer un objet compliqué, la fonction f, par des objets plus simples, des polynômes, qui sont facilement manipulables par un ordinateur. C'est la technique qu'on emploiera au chapitre 3 sur le calcul approché d'intégrales : on remplacera la fonction qu'on souhaite intégrer par un polynôme, dont on sait calculer explicitement l'intégrale.

Dans tout ce chapitre, la fonction f sera supposée continue sur I. Quand $I = [a,b]$ (intervalle fermé borné), on notera $\|\cdot\|_\infty$ la norme sup sur $[a,b]$:

$$\forall g \in \mathscr{C}([a,b]), \quad \|g\|_\infty := \sup_{x\in[a,b]} |g(x)|.$$

On rappelle que toute fonction continue sur un compact est bornée et atteint ses bornes, donc le supremum ci-dessus est fini et il est atteint (c'est un maximum). Enfin, on rappelle qu'une suite de fonctions $(g_n)_{n\in\mathbb{N}}$ converge uniformément vers g sur $[a,b]$ si et seulement si :

$$\|g_n - g\|_\infty \xrightarrow[n\to+\infty]{} 0.$$

2.1 Interpolation de Lagrange

Dans cette section, on fixe $x_0, x_1, \ldots, x_n$ des réels distincts de l'intervalle $[a,b]$, appelés **nœuds**. L'interpolation polynomiale consiste à déterminer un polynôme P, de plus bas degré possible, qui coïncide avec f aux points $x_0, \ldots, x_n$. L'espoir étant que si P coïncide avec f en suffisamment de points, $P(x)$ pourrait être une bonne approximation de $f(x)$ pour $x \in [a,b]$.

Si on considère les coefficients de P comme les inconnues du problème, on est amené à résoudre le système linéaire :

$$\begin{cases} P(x_0) = f(x_0) \\ P(x_1) = f(x_1) \\ \quad\vdots \\ P(x_n) = f(x_n) \end{cases} \tag{2.1}$$

Pour $P \in \mathbb{R}_d[X]$, c'est un système linéaire à $n+1$ équations et $d+1$ inconnues. Si ces équations sont linéairement indépendantes (on verra que c'est le cas), on comprend qu'il faut $d \geq n$ pour qu'une solution existe, et $d = n$ pour que celle-ci soit unique. On cherche donc P appartenant à $\mathbb{R}_n[X]$.
Bien sûr, on pourrait calculer P en résolvant le système (2.1) par la méthode du pivot de Gauss par exemple (pour n petit, c'est certainement le plus simple). La matrice de ce système est la matrice

de Vandermonde :

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix}.$$

En général, c'est une mauvaise idée, et ce pour deux raisons :

1. D'une part, la résolution d'un système $n \times n$ par pivot de Gauss nécessite $O(n^3)$ opérations.
2. D'autre part, la matrice de Vandermonde devient numériquement presque singulière dès que n est un peu grand ou que des nœuds sont proches les uns des autres : une petite erreur sur les données $f(x_i)$, ou une simple erreur d'arrondi commise en cours de calcul, peut alors se traduire par une erreur énorme sur les coefficients de P obtenus, rendant le résultat inutilisable.

On va voir des algorithmes qui résolvent le problème (2.1) en $O(n^2)$ opérations, ce qui répond au premier défaut. La question de la stabilité numérique est plus délicate : elle ne disparaît pas simplement en changeant de méthode de calcul, et dépend aussi du choix des nœuds d'interpolation eux-mêmes. Nous y reviendrons plus loin dans ce chapitre.

2.1.a Polynômes d'interpolation de Lagrange

On commence par montrer que le système (2.1) possède une unique solution dans $\mathbb{R}_n[X]$. De plus, on va exhiber une base de $\mathbb{R}_n[X]$ dans laquelle la solution s'écrit simplement.

Proposition 2.1.1. L'application $\Phi\colon \mathbb{R}_n[X] \to \mathbb{R}^{n+1}$ définie par

$$\forall P \in \mathbb{R}_n[X], \quad \Phi(P) := \big(P(x_0), P(x_1), \dots, P(x_n)\big)$$

est un isomorphisme. ◀

Démonstration. L'application Φ est linéaire. Calculons son noyau. Soit $P \in \mathrm{Ker}(\Phi)$, alors pour tout $i \in [\![0, n]\!]$, $P(x_i) = 0$. Puisque les x_i sont distincts, P possède $n+1$ racines. Or P est de degré au plus n, donc P est le polynôme nul. Par conséquent, $\mathrm{Ker}(\Phi) = \{0_{\mathbb{R}[X]}\}$ donc Φ est injective. Puisque $\dim(\mathbb{R}_n[X]) = \dim(\mathbb{R}^{n+1})$, Φ est bijective. □

Corollaire 2.1.2. Pour tout $(y_0, y_1, \dots, y_n) \in \mathbb{R}^{n+1}$, il existe un unique polynôme $P \in \mathbb{R}_n[X]$ tel que :

$$\forall i \in [\![0, n]\!], \quad P(x_i) = y_i$$ ◀

Démonstration. L'application Φ de la proposition 2.1.1 est un isomorphisme, donc elle est inversible. Le polynôme $P := \Phi^{-1}(y_0, y_1, \dots, y_n)$ est le polynôme recherché. □

Notons $(e_0, e_1, \dots, e_n)$ la base canonique de $\mathbb{R}^{n+1}$. L'application Φ^{-1} décrite précédemment étant linéaire, le polynôme d'interpolation prenant la valeur y_i en x_i s'écrit :

$$P = \Phi^{-1}\left(\sum_{i=0}^{n} y_i e_i\right) = \sum_{i=0}^{n} y_i \Phi^{-1}(e_i).$$

Ainsi, il suffit de connaitre les $\Phi^{-1}(e_i)$ pour pouvoir calculer un polynôme interpolateur prenant n'importe quelles valeurs aux nœuds x_i.
Les polynômes $L_i = \Phi^{-1}(e_i)$ sont appelés les polynômes interpolateurs de Lagrange. Ils satisfont :

$$\forall i, j \in [\![0, n]\!], \quad L_i(x_j) = \delta_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

où $\delta_{i,j}$ est le delta de Kronecker. Autrement dit, le polynôme L_i est de degré au plus n et a pour racines x_j pour tout $j \neq i$, il s'écrit donc :

$$L_i(X) = \lambda \prod_{\substack{j=0 \\ j\neq i}}^{n} (X - x_j),$$

avec $\lambda \in \mathbb{R}$. La condition $L_i(x_i) = 1$ donne :

$$\lambda = \left(\prod_{\substack{j=0 \\ j\neq i}}^{n} (x_i - x_j) \right)^{-1}.$$

Définition 2.1.3. On appelle **polynômes interpolateurs de Lagrange** aux nœuds $x_0, x_1, \ldots, x_n$ les polynômes :

$$\forall i \in [\![0, n]\!], \quad L_i(X) := \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{X - x_j}{x_i - x_j}.$$

Exemple 2.1.4. Les polynômes interpolateurs de Lagrange aux nœuds $(x_0, x_1, x_2) = (1, 2, 3)$ sont :

$$L_0(X) = \frac{(X-2)(X-3)}{(1-2)(1-3)}, \quad L_1(X) = \frac{(X-1)(X-3)}{(2-1)(2-3)}, \quad L_2(X) = \frac{(X-1)(X-2)}{(3-1)(3-2)}.$$ ◀

Proposition 2.1.5. Les polynômes $(L_i)_{0\leq i\leq n}$ forment une base de $\mathbb{R}_n[X]$. ◀

Démonstration. La famille $(L_i)_{0\leq i\leq n}$ est l'image de la base canonique de $\mathbb{R}^{n+1}$ par l'isomorphisme $\Phi^{-1} : \mathbb{R}^{n+1} \to \mathbb{R}_n[X]$, donc c'est une base de $\mathbb{R}_n[X]$. □

Finalement, on a démontré le théorème suivant.

Théorème 2.1.6. Il existe un unique polynôme $P \in \mathbb{R}_n[X]$ satisfaisant $P(x_i) = f(x_i)$ pour tout $i \in [\![0, n]\!]$. Dans la base des polynômes de Lagrange, il s'écrit :

$$P(X) = \sum_{i=0}^{n} f(x_i) L_i(X).$$

2.1.b Calcul effectif du polynôme interpolateur

Notons P_n le polynôme interpolateur de f aux nœuds $x_0, \ldots, x_n$. La base des polynômes de Lagrange se prête mal au calcul effectif de P_n. En effet, l'évaluation de P_n en un point x à partir de cette base nécessite $O(n^2)$ opérations :

- pour chaque $i \in [\![0, n]\!]$, calculer $L_i(x)$ nécessite $O(n)$ opérations ;
- calculer $P_n(x)$ nécessite de calculer une somme de $n+1$ termes, chaque terme coûtant $O(n)$ opérations, soit $O(n^2)$ opérations au total.

De plus ce calcul doit être entièrement recommencé si on souhaite évaluer P_n en un autre point. Enfin, si on ajoute un nouveau nœud x_{n+1}, tous les polynômes de Lagrange $L_0, \ldots, L_n$ sont modifiés, puisqu'ils dépendent de l'ensemble des nœuds.

On cherche donc une base mieux adaptée, dans laquelle le passage du polynôme interpolateur P_{n-1} (aux nœuds $x_0, \ldots, x_{n-1}$) au polynôme interpolateur P_n (aux nœuds $x_0, \ldots, x_n$) s'effectue par une simple mise à jour, sans tout recalculer.

Définition 2.1.7. On appelle **polynômes de Newton** associés aux nœuds $x_0, \dots, x_n$ (dans cet ordre) les polynômes :

$$\forall k \in [\![0, n]\!], \quad N_k(X) := \prod_{i=0}^{k-1} (X - x_i),$$

avec la convention $N_0(X) := 1$.

Proposition 2.1.8. Les polynômes de Newton $(N_k)_{0 \le k \le n}$ forment une base de $\mathbb{R}_n[X]$. ◀

Dans cette base, les coordonnées de P_n se calculent par récurrence à partir des valeurs $f(x_i)$.

Définition 2.1.9. On appelle **différences divisées** les quantités définies par récurrence par :

$$\begin{cases} f[x_0] := f(x_0) \\ f[x_0, \dots, x_{k+1}] := \dfrac{f[x_1, \dots, x_{k+1}] - f[x_0, \dots, x_k]}{x_{k+1} - x_0} \end{cases}$$

On peut voir les différences divisées comme une généralisation de la notion de taux d'accroissement. En effet, la différence divisée en 2 points est exactement le taux d'accroissement :

$$f[x_0, x_1] = \frac{f[x_1] - f[x_0]}{x_1 - x_0} = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

La différence divisée en trois points s'écrit quant à elle :

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = \frac{\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0}.$$

Théorème 2.1.10. Le polynôme interpolateur de f aux nœuds $x_0, x_1, \dots, x_n$ s'écrit dans la base de Newton :

$$P_n(X) = \sum_{k=0}^{n} f[x_0, \dots, x_k] N_k(X).$$

La démonstration de ce théorème s'appuie sur la proposition suivante.

Proposition 2.1.11. La différence divisée $f[x_0, \dots, x_n]$ est le coefficient du monôme X^n du polynôme interpolateur de f aux nœuds $x_0, x_1, \dots, x_n$. ◀

Démonstration. Montrons-le par récurrence sur n. Si $n = 0$, c'est trivial. Soit $n \ge 1$, supposons que la proposition est vraie au rang $n-1$, montrons-la au rang n. Notons :

- P_n le polynôme interpolateur de f aux points $x_0, \dots, x_n$;
- $Q_{0:n-1}$ le polynôme interpolateur de f aux points $x_0, \dots, x_{n-1}$;
- $Q_{1:n}$ le polynôme interpolateur de f aux points $x_1, \dots, x_n$.

On a la relation :

$$P_n(X) = \frac{(X - x_0) Q_{1:n}(X) - (X - x_n) Q_{0:n-1}}{x_n - x_0}.$$

En effet, si on note Q le membre de droite, on vérifie immédiatement que $Q(x_i) = f(x_i)$ pour tout $i \in [\![0, n]\!]$. Par unicité du polynôme interpolateur de Lagrange, on a $P_n = Q$.
Par hypothèse de récurrence, le coefficient du monôme X^{n-1} de $Q_{0:n-1}$ et de $Q_{1:n}$ sont respectivement $f[x_0, \dots, x_{n-1}]$ et $f[x_1, \dots, x_n]$. D'après l'égalité précédente, le coefficient de X^n de P_n est donc :

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

c'est-à-dire $f[x_0, \dots, x_n]$. Par récurrence, la proposition est démontrée. □

Corollaire 2.1.12. La quantité $f[x_0,\dots,x_n]$ ne dépend pas de l'ordre des nœuds $x_0,\dots,x_n$. ◀

Démonstration. Le polynôme interpolateur de f aux nœuds $x_0,\dots,x_n$ ne dépend pas de l'ordre des nœuds, donc ses coefficients non plus. En particulier, c'est le cas du coefficient de X^n, c'est-à-dire $f[x_0,\dots,x_n]$. □

On peut maintenant démontrer le théorème 2.1.10 donnant l'expression du polynôme interpolateur dans la base des polynômes de Newton.

Démonstration du théorème 2.1.10. Pour tout $k \in [\![0,n]\!]$, notons P_k le polynôme interpolateur de f aux nœuds $x_0,\dots,x_k$.
Alors $P_k - P_{k-1}$ s'annule en $x_0,\dots,x_{k-1}$, donc il est divisible par N_k. Comme $\deg(P_k - P_{k-1}) \leq k$, le quotient est un scalaire. Le coefficient de X^k dans $P_k - P_{k-1}$ étant $f[x_0,\dots,x_k]$ d'après la proposition 2.1.11, on a l'égalité :

$$P_k(X) - P_{k-1}(X) = f[x_0,\dots,x_k]N_k(X).$$

Par télescopage, on en déduit :

$$\begin{aligned} P_n(X) &= P_0(X) + \sum_{k=1}^{n} \big(P_k(X) - P_{k-1}(X)\big) \\ &= f(x_0) + \sum_{k=1}^{n} f[x_0,\dots,x_k]N_k(X) \\ &= \sum_{k=0}^{n} f[x_0,\dots,x_k]N_k(X). \end{aligned}$$

□

Algorithme des différences divisées. On organise les différences divisées en un tableau triangulaire à double entrée : pour $i \in [\![0,n]\!]$ et $k \in [\![0,n-i]\!]$, on note $D_{i,k} := f[x_i, x_{i+1},\dots,x_{i+k}]$.

1. Initialiser $D_{i,0} = f(x_i)$ pour tout $i \in [\![0,n]\!]$.
2. Pour k allant de 1 à n :

$$\forall i \in [\![0,n-k]\!], \quad D_{i,k} = \frac{D_{i+1,k-1} - D_{i,k-1}}{x_{i+k} - x_i}.$$

3. Les coefficients recherchés sont ceux de la première ligne : $D_{0,k}$, $k \in [\![0,n]\!]$.

Exemple 2.1.13. Avec 4 points :

	$D_{\bullet,0}$	$D_{\bullet,1}$	$D_{\bullet,2}$	$D_{\bullet,3}$
x_0	$f(x_0)$	$f[x_0,x_1]$	$f[x_0,x_1,x_2]$	$f[x_0,x_1,x_2,x_3]$
x_1	$f(x_1)$	$f[x_1,x_2]$	$f[x_1,x_2,x_3]$	
x_2	$f(x_2)$	$f[x_2,x_3]$		
x_3	$f(x_3)$			

Par exemple, pour $f(x) := \frac{1}{1+x^2}$ et $(x_0,x_1,x_2,x_3) := (-1,0,1,2)$:

	$D_{\bullet,0}$	$D_{\bullet,1}$	$D_{\bullet,2}$	$D_{\bullet,3}$
-1	0,5	0,5	$-0,5$	0,2
0	1	$-0,5$	0,1	
1	0,5	$-0,3$		
2	0,2			

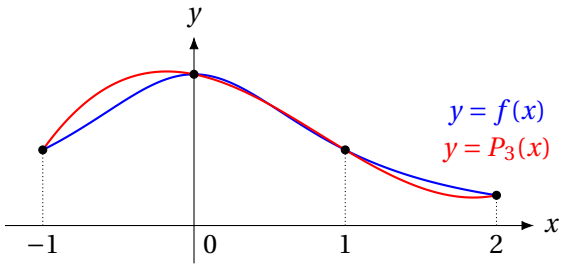


Figure 2.1 – Interpolation de $f(x) = \frac{1}{1+x^2}$ aux nœuds -1, 0, 1 et 2.

Par conséquent, $P_3(X) = 0{,}5 + 0{,}5(X+1) - 0{,}5(X+1)X + 0{,}2(X+1)X(X-1)$. La figure 2.1 représente les graphes de f et de P_3. ◀

Remarque 2.1.14.

1. Chaque case du tableau nécessite 2 soustractions et 1 division, donc le calcul du tableau nécessite $O(n^2)$ opérations.
2. Une fois le tableau calculé, on peut évaluer le polynôme interpolateur en un point x en $O(n)$ opérations en utilisant une variante de la méthode de Horner (cf. TP2). Si on veut évaluer en un autre point x', inutile de recalculer le tableau.
3. Si on veut ajouter un nœud x_{n+1}, il suffit de calculer n cases supplémentaires, soit $O(n)$ opérations. ◀

2.2 Erreur d'interpolation

Quitte à renuméroter les nœuds, on suppose que $x_0 < x_1 < \cdots < x_n$. On note P_n le polynôme interpolateur de Lagrange de f aux nœuds $x_0, \ldots, x_n$. On note également :

$$\Pi_n(X) := \prod_{i=0}^{n} (X - x_i).$$

L'erreur ponctuelle commise entre f et son polynôme interpolateur est donnée par le théorème suivant.

Théorème 2.2.1. On suppose que f est continue sur $[a,b]$ et $n+1$ fois dérivable sur $]a,b[$. Pour tout $x \in [a,b]$, il existe $\xi \in \left]\min(x_0,x), \max(x_n,x)\right[$ tel que :

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

Lemme 2.2.2 (Rolle itéré)**.** Soit $g\colon [a,b] \to \mathbb{R}$ une fonction continue sur $[a,b]$ et n fois dérivable sur $]a,b[$. On suppose qu'il existe $c_0 < c_1 < \cdots < c_n$ dans $[a,b]$ tels que $g(c_i) = 0$. Alors il existe $\xi \in]c_0, c_n[$ tel que $g^{(n)}(\xi) = 0$. ◀

Démonstration. Par récurrence sur n. Pour $n = 1$, c'est le théorème de Rolle. Soit $n \geq 1$. Supposons le théorème vrai pour n, démontrons-le pour $n+1$. Par le théorème de Rolle, il existe $\gamma_i \in]c_i, c_{i+1}[$ tels que $g'(\gamma_i) = 0$. Par hypothèse de récurrence, il existe $\xi \in]\gamma_0, \gamma_{n-1}[\subset]c_0, c_n[$ tel que $(g')^{(n)}(\xi) = 0$, c'est-à-dire $g^{(n+1)}(\xi) = 0$. □

Démonstration du théorème 2.2.1. Soit $x \in [a,b]$. Si x est l'un des x_i, l'égalité est triviale. Supposons que $x \notin \{x_0, \ldots, x_n\}$. Notons $R(x) = f(x) - P_n(x)$ et $F_x(t) = R(t)\Pi_n(x) - R(x)\Pi_n(t)$. La fonction F_x s'annule en $n+2$ points :

- $F_x(x_i) = 0$ pour tout $i \in [\![0, n]\!]$, car $R(x_i) = 0$ et $\Pi_n(x_i) = 0$;
- $F_x(x) = R(x)\Pi_n(x) - R(x)\Pi_n(x) = 0$.

Par le lemme de Rolle itéré, il existe $\xi \in]\min(x_0, x), \max(x_n, x)[$ tel que $F^{(n+1)}(\xi) = 0$, c'est-à-dire :

$$R^{(n+1)}(\xi)\Pi_n(x) = R(x)\Pi_n^{(n+1)}(\xi).$$

Or $\Pi_n^{(n+1)}$ est constant égal à $(n+1)!$ et $R^{(n+1)}(\xi) = f^{(n+1)}(\xi)$ car $P_n^{(n+1)} = 0$, d'où

$$f(x) - P_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\Pi_n(x). \qquad \square$$

Corollaire 2.2.3. Si f est de classe $\mathscr{C}^{n+1}$ sur $[a, b]$, alors :

$$\|f - P_n\|_\infty \leq \frac{\|f^{(n+1)}\|_\infty}{(n+1)!}\|\Pi_n\|_\infty. \qquad \blacktriangleleft$$

Remarque 2.2.4. L'erreur d'interpolation se décompose donc en deux facteurs de nature différente : le terme $\|f^{(n+1)}\|_\infty$, qui dépend uniquement des caractéristiques de f mais pas des nœuds, et le terme $\|\Pi_n\|_\infty$, qui ne dépend que de la répartition des nœuds $x_0, \ldots, x_n$ et non de f. Cette majoration suggère qu'on pourrait espérer faire tendre l'erreur vers 0 en augmentant indéfiniment le nombre de nœuds n, à condition de bien choisir leur répartition pour contrôler $\|\Pi_n\|_\infty$. C'est loin d'être automatique, comme on le verra dans la sous-section suivante. $\blacktriangleleft$

Pour finir, les théorèmes 2.1.10 et 2.2.1 ont pour conséquence le résultat suivant, qui généralise le théorème des accroissements finis.

Corollaire 2.2.5 (accroissements finis généralisés)**.** On suppose que f est continue sur $[a, b]$ et n fois dérivable sur $]a, b[$. Pour tous réels distincts $x_0 < \cdots < x_n$ de $[a, b]$, il existe $\xi \in]x_0, x_n[$ tel que :

$$f[x_0, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!} \qquad \blacktriangleleft$$

Démonstration. Soit P_{n-1} le polynôme interpolateur de f aux nœuds $x_0, \ldots, x_{n-1}$. D'après le théorème 2.2.1, il existe $\xi \in]x_0, x_n[$ tel que :

$$f(x_n) - P_{n-1}(x_n) = \frac{f^{(n)}(\xi)}{n!}\Pi_{n-1}(x_n).$$

Par ailleurs, le théorème 2.1.10 nous donne :

$$f(x_n) - P_{n-1}(x_n) = P_n(x_n) - P_{n-1}(x_n) = f[x_0, \ldots, x_n]N_n(x_n).$$

En remarquant que :

$$\Pi_{n-1}(x_n) = N_n(x_n) = \prod_{i=0}^{n-1}(x_n - x_i) \neq 0$$

on déduit des égalités précédentes que $\frac{f^{(n)}(\xi)}{n!} = f[x_0, \ldots, x_n]$. $\square$

En particulier, les différences divisées restent bornées par $\|f^{(n+1)}\|_\infty$ quelle que soit la position des nœuds dans $[a, b]$.

2.3 Convergence uniforme et choix des nœuds

Considérons des nœuds $x_0, \dots, x_n$ et notons P_n le polynôme interpolateur de f en ces nœuds. On se demande comment évolue l'erreur de P_n par rapport à f lorsque le nombre de nœuds augmente : converge-t-elle vers 0 ? Autrement dit, a-t-on convergence uniforme des polynômes interpolateurs vers f lorsque n tend vers $+\infty$?

Proposition 2.3.1. On suppose que f est $\mathscr{C}^\infty$ sur $[a,b]$. Si $\frac{1}{n!}\|f^{(n)}\|_\infty (b-a)^n \to 0$, alors quels que soient les nœuds, $(P_n)_{n\in\mathbb{N}}$ converge uniformément vers f sur $[a,b]$. ◀

Démonstration. D'après le corollaire 2.2.3 :

$$\|f - P_n\|_\infty \leq \frac{\|f^{(n+1)}\|_\infty}{(n+1)!}\|\Pi_n\|_\infty.$$

Or, pour tout $x \in [a,b]$, on a :

$$|\Pi_n(x)| = \prod_{i=0}^{n} |x - x_i| \leq \prod_{i=0}^{n} (b-a) = (b-a)^{n+1}.$$

Par conséquent, $\|f - P_n\|_\infty \leq \frac{1}{(n+1)!}\|f^{(n+1)}\|_\infty (b-a)^{n+1}$. Cette quantité tend vers 0 par hypothèse, donc $(P_n)_{n\in\mathbb{N}}$ converge uniformément vers f sur $[a,b]$. □

Exemple 2.3.2. Cette proposition s'applique par exemple si les dérivées de f sont uniformément bornées sur $[a,b]$, c'est-à-dire si $\exists M > 0$, $\forall n \in \mathbb{N}$, $\|f^{(n)}\|_\infty \leq M$.
Par exemple, si $f(x) = \mathrm{e}^x$, alors $\|f^{(n)}\|_\infty = \mathrm{e}^b$ pour tout n. Par conséquent, quels que soient les nœuds d'interpolation, $(P_n)_{n\in\mathbb{N}}$ converge uniformément vers f sur $[a,b]$. ◀

Cette dernière proposition donne un critère de convergence indépendant de la position des nœuds. Mais cette majoration est très pessimiste en général : pour que $\|\Pi_n\|_\infty = (b-a)^{n+1}$, il faudrait que tous les nœuds soient égaux à l'une des bornes de l'intervalle ! On peut espérer améliorer ce résultat si on répartit mieux les nœuds.

2.3.a Nœuds équidistants

On considère la subdivision régulière de $[a,b]$ en n intervalles (donc $n+1$ nœuds) :

$$\forall i \in [\![0,n]\!], \quad x_i = a + ih$$

avec $h = \frac{b-a}{n}$ le pas de la subdivision.

Lemme 2.3.3. Pour les nœuds équidistants, on a l'encadrement :

$$h^{n+1}\frac{n!}{4n} \leq \|\Pi\|_\infty \leq h^{n+1} n!$$ ◀

Remarque 2.3.4. D'après le lemme ci-dessus :

$$\frac{1}{4n}\frac{(b-a)^{n+1}n!}{n^{n+1}} \leq \|\Pi_n\|_\infty \leq \frac{(b-a)^{n+1}n!}{n^{n+1}}.$$

En utilisant la formule de Stirling $n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} \mathrm{e}^{-n}$, on a

$$\frac{(b-a)^{n+1}n!}{n^{n+1}} \sim \sqrt{2\pi}\frac{(b-a)^{n+1}n^{n+\frac{1}{2}}\mathrm{e}^{-n}}{n^{n+1}} \sim \frac{\sqrt{2\pi}\,\mathrm{e}}{\sqrt{n}}\left(\frac{b-a}{\mathrm{e}}\right)^{n+1}$$

Ainsi, il existe des constantes $C, C' > 0$ telles que pour n assez grand :

$$\frac{C}{n\sqrt{n}}\left(\frac{b-a}{\mathrm{e}}\right)^{n+1} \leq \|\Pi_n\|_\infty \leq \frac{C'}{\sqrt{n}}\left(\frac{b-a}{\mathrm{e}}\right)^{n+1}.$$

L'ordre de grandeur de $\|\Pi_n\|_\infty$ est donc $\left(\frac{b-a}{\mathrm{e}}\right)^{n+1}$, si on ne tient compte que du terme géométrique. ◀

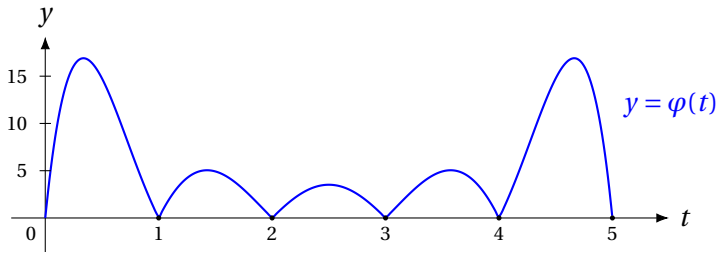


Figure 2.2 – Graphe de $\varphi_5(t) = \prod_{i=0}^{5} |t-i|$ sur $[0,5]$. Conformément à la preuve, les bosses sont les plus hautes près des bords de l'intervalle ($[0,1]$ et $[4,5]$, où φ_5 dépasse 16), et les plus basses au centre (où φ_5 ne dépasse pas 4).

Démonstration. Soit $x \in [a,b]$. On écrit $x = a + th$ avec $t \in [0,n]$. On a alors :

$$|\Pi_n(x)| = \prod_{i=0}^{n} |x - x_i| = \prod_{i=0}^{n} |(a+th) - (a+ih)| = h^{n+1} \prod_{i=0}^{n} |t-i|.$$

Notons $\varphi_n(t) \coloneqq \prod_{i=0}^{n} |t-i|$, pour $t \in [0,n]$. On remarque que $\forall t \in [0,n]$, $\varphi_n(n-t) = \varphi_n(t)$ donc la fonction φ_n est symétrique par rapport à $\frac{n}{2}$, on peut donc restreindre l'étude à l'intervalle $[0, \frac{n}{2}]$.
Pour $t \in [0,n] \setminus \mathbb{N}$, on a :

$$\frac{\varphi_n(t-1)}{\varphi_n(t)} = \frac{\prod_{i=1}^{n} |t-(i+1)|}{\prod_{i=0}^{n} |t-i|} = \frac{|t-(n+1)|}{|t|} = \frac{n+1-t}{t}.$$

donc $\frac{\varphi_n(t-1)}{\varphi_n(t)} > 1$ sur $[1, \frac{n}{2}] \setminus \mathbb{N}$. Par conséquent, pour tout $k \in \mathbb{N}$ compris entre 1 et $\frac{n}{2} - 1$:

$$\max_{t \in [k-1,k]} \varphi_n(t) > \max_{t \in [k,k+1]} \varphi_n(t).$$

En particulier, le maximum de φ_n sur $[0, \frac{n}{2}]$ est atteint dans $[0,1]$ (et aussi dans $[n-1,n]$ par symétrie) :

$$\max_{t \in [0,n]} \varphi_n(t) = \max_{t \in [0,1]} \varphi_n(t) = \max_{t \in [n-1,n]} \varphi_n(t).$$

Pour $t \in [0,1]$, on a d'une part :

$$\varphi_n(t) = \prod_{i=0}^{n} |t-i| = t \prod_{i=1}^{n} (i-t) \leq \prod_{i=1}^{n} i = n!$$

et d'autre part :

$$\varphi_n\left(\frac{1}{2}\right) = \prod_{i=0}^{n} \left|\frac{1}{2} - i\right| = \frac{1}{2} \prod_{i=1}^{n} \left(i - \frac{1}{2}\right) = \frac{1}{2^{n+1}} \prod_{i=1}^{n} (2i-1) \geq \frac{1}{2^{n+1}} \prod_{i=1}^{n-1} 2i = \frac{(n-1)!}{4} = \frac{n!}{4n}.$$

Par conséquent, on a l'encadrement :

$$\frac{n!}{4n} \leq \max_{t \in [0,n]} \varphi_n(t) \leq n!,$$

d'où l'encadrement :

$$h^{n+1} \frac{n!}{4n} \leq \max_{x \in [a,b]} |\Pi_n(x)| \leq h^{n+1} n!$$ □

Remarque 2.3.5. La preuve montre que le maximum est atteint aux bords de l'intervalle $[a,b]$. En effet, la fonction φ de la preuve atteint son maximum dans les intervalles $[0,1]$ et $[n-1,n]$, ce qui correspond aux intervalles $[a,a+h]$ et $[b-h,b]$ avec $h = \frac{b-a}{n}$. Ainsi, l'interpolation avec des nœuds équidistants produit des erreurs plus importantes aux bords de l'intervalle qu'en son centre. ◂

Proposition 2.3.6. Si f est $\mathscr{C}^{n+1}$ sur $[a,b]$, alors le polynôme P_n interpolateur de f aux nœuds équidistants sur $[a,b]$ vérifie :

$$\|f - P_n\|_\infty \leq \frac{\|f^{(n+1)}\|_\infty}{n+1} \left(\frac{b-a}{n}\right)^{n+1}.$$ ◂

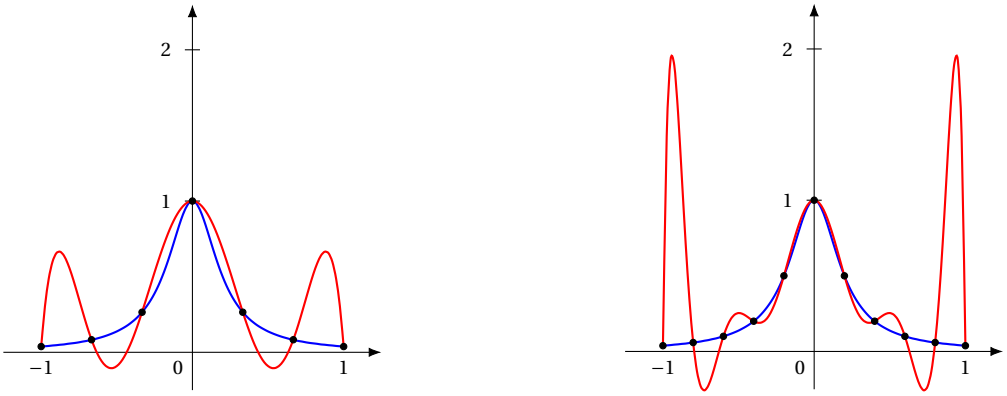


Figure 2.3 – Interpolation de la fonction de Runge $f(x) = \frac{1}{1+25x^2}$ (bleu) par un polynôme aux nœuds équidistants (rouge), pour $n = 6$ (gauche) et $n = 10$ (droite). Plus n augmente, plus les oscillations près des bords ± 1 s'aggravent, alors que f elle-même reste bornée par 1.

2.3.b Phénomène de Runge

On a vu au corollaire 2.2.3 que l'erreur d'interpolation est majorée par le produit $\|f^{(n+1)}\|_\infty \|\Pi_n\|_\infty$. On pourrait donc penser qu'il suffit d'augmenter indéfiniment le nombre de nœuds n pour rendre cette erreur aussi petite que l'on veut. L'exemple suivant, dû à Runge, montre que non : pour des nœuds équidistants, il existe des fonctions f pourtant $\mathscr{C}^\infty$ (et même analytique) pour lesquelles $\|f - P_n\|_\infty$ diverge vers $+\infty$ quand $n \to +\infty$. Plus on augmente le nombre de nœuds, plus l'approximation de la fonction se dégrade !

On considère la fonction :

$$f(x) := \frac{1}{1+25x^2}, \qquad x \in [-1,1].$$

Si $|x| < \frac{1}{5}$, un développement en série entière donne :

$$f(x) = \frac{1}{1+(5x)^2} = \sum_{k=0}^{+\infty} (-1)^k (5x)^{2k}.$$

Pour n pair, on a donc $|f^{(n)}(0)| = 5^n n!$, ce qui montre que $\|f^{(n)}\|_\infty \geq 5^n n!$. Cette minoration donne le bon ordre de grandeur de la dérivée n-ième, y compris pour n impair. En effet, on peut montrer que :

$$\|f^{(n)}\|_\infty \underset{n\to+\infty}{\sim} 5^n n!$$

Or pour des nœuds équidistants sur $[-1,1]$, on a montré que $\|\Pi_n\|_\infty$ était d'ordre $\left(\frac{2}{\mathrm{e}}\right)^{n+1}$, donc il existe $C > 0$ tel que pour n assez grand :

$$\frac{\|f^{(n+1)}\|_\infty}{(n+1)!} \|\Pi_n\|_\infty \geq C \left(\frac{10}{\mathrm{e}}\right)^{n+1} \xrightarrow[n\to+\infty]{} +\infty.$$

En soi, le fait que le majorant de $\|f - P_n\|_\infty$ diverge ne prouve rien. Mais on peut montrer $\|f - P_n\|_\infty$ que diverge effectivement (voir Demailly, 2016, chap. II, section 2.3).

On illustre cette divergence avec la figure 2.3. On peut voir que lorsque n augmente, le polynôme interpolateur oscille de plus en plus aux bords de l'intervalle. C'est ce qu'on appelle le **phénomène de Runge**. L'interpolation sur des nœuds équidistants produit généralement ce phénomène d'oscillation aux bords.

2.3.c Nœuds de Tchebychev

Les nœuds équidistants ne sont pas le meilleur choix qu'on puisse faire pour interpoler une fonction. Reprenons la majoration de l'erreur :

$$\|f - P_n\|_\infty \leq \frac{\|f^{n+1}\|_\infty}{(n+1)!}\|\Pi_n\|_\infty.$$

Si on cherche à minimiser ce majorant indépendamment de la fonction f, il faut répartir les nœuds de sorte à minimiser $\|\Pi_n\|_\infty$. Sur $[-1,1]$, on peut montrer que $\|\Pi_n\|_\infty$ est minimal lorsque $\Pi_n = \frac{1}{2^n}T_{n+1}$ où T_{n+1} est le $(n+1)$-ième polynôme de Tchebychev (voir Yger et Weil, 2009, p. 91). Les nœuds correspondants sont les racines de ce polynôme et sont appelés les nœuds de Tchebychev.

Définition 2.3.7. On appelle **polynômes de Tchebychev** les polynômes $(T_n)_{n\in\mathbb{N}}$ définis par :

$$\forall x \in \mathbb{R}, \quad T_n(x) := \cos(n \arccos x).$$

Proposition 2.3.8. Les polynômes de Tchebychev satisfont la relation de récurrence :

$$\begin{cases} T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \\ T_0(x) = 1 \\ T_1(x) = x \end{cases}$$

En particulier, ce sont bien des polynômes. De plus, T_n est de degré n et de coefficient dominant 2^{n-1} pour tout $n \geq 1$. ◀

Démonstration. Les expressions de T_0 et T_1 sont claires au vu de la définition. Notons $\theta := \arccos x$, de sorte que $T_n(x) = T_n(\cos\theta) = \cos(n\theta)$. Pour $n \geq 1$, on a :

$$\begin{aligned} T_{n+1}(x) + T_{n-1}(x) &= \cos(n\theta + \theta) + \cos(n\theta - \theta) \\ &= 2\cos(n\theta)\cos(\theta) \\ &= 2xT_n(x), \end{aligned}$$

d'où la relation $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$. Puisque $T_0(x) = 1$ et $T_1(x) = x$, on montre par récurrence que T_n est un polynôme de degré n pour tout $n \in \mathbb{N}$, et que son coefficient dominant est 2^{n-1} pour tout $n \in \mathbb{N}^*$. □

Proposition 2.3.9. Le polynôme T_n possède n racines réelles distinctes, données par :

$$x_i = \cos\left(\frac{(2k+1)\pi}{2n}\right), \quad k \in [\![0, n-1]\!].$$ ◀

Démonstration. Soit $x \in \mathbb{R}$. On a :

$$\begin{aligned} T_n(x) = 0 &\iff \cos(n \arccos x) = 0 \\ &\iff \exists k \in \mathbb{Z},\ n \arccos x = \frac{\pi}{2} + k\pi \\ &\iff \exists k \in \mathbb{Z},\ \arccos x = \frac{\pi}{2n} + \frac{k\pi}{n} \\ &\iff \exists k \in \mathbb{Z},\ \arccos x = \frac{(2k+1)\pi}{2n}. \end{aligned}$$

Or, $\arccos(x) \in [0,\pi]$, et les seules valeurs de $k \in \mathbb{Z}$ pour lesquelles $\frac{(2k+1)\pi}{2n}$ est compris entre 0 et π sont $k = 0, 1, \ldots, n-1$. Ainsi, T_n possède n racines réelles distinctes :

$$x_i = \cos\left(\frac{(2k+1)\pi}{2n}\right), \quad k \in [\![0, n-1]\!].$$

Puisque T_n est de degré n, il n'a pas d'autres racines. □

Les $n+1$ racines de T_{n+1} sont appelées les nœuds de Tchebychev d'ordre n sur $[-1,1]$. Pour obtenir les nœuds de Tchebychev sur un intervalle $[a,b]$ quelconque, il suffit d'utiliser la bijection affine :

$$\begin{array}{ccc} [-1,1] & \longrightarrow & [a,b] \\ x & \longmapsto & \frac{a+b}{2} + \frac{b-a}{2}x. \end{array}$$

Définition 2.3.10. On appelle **nœuds de Tchebychev** d'ordre n sur $[a,b]$ les nœuds :

$$\forall i \in [\![0,n]\!], \quad x_i = \frac{a+b}{2} + \frac{b-a}{2}\cos\left(\frac{(2i+1)\pi}{2n+2}\right).$$

Si on considère les nœuds de Tchebychev sur $[-1,1]$, alors :

$$\|\Pi_n\|_\infty = \frac{1}{2^n}\|T_{n+1}\|_\infty = \frac{1}{2^n},$$

car T_{n+1} est borné par 1 et la borne est atteinte en 0. Si on considère les nœuds de Tchebychev sur $[a,b]$ quelconque, la bijection affine avec $[-1,1]$ nous permet d'obtenir :

$$\|\Pi_n\|_\infty = \left(\frac{b-a}{2}\right)^{n+1}\frac{1}{2^n} = 2\left(\frac{b-a}{4}\right)^{n+1}.$$

Proposition 2.3.11. Si f est de classe $\mathscr{C}^{n+1}$ sur $[a,b]$, le polynôme P_n d'interpolation de f aux nœuds de Tchebychev d'ordre n sur $[a,b]$ vérifie :

$$\|f - P_n\|_\infty \leq \frac{\|f^{(n+1)}\|_\infty}{(n+1)!} \times 2\left(\frac{b-a}{4}\right)^{n+1}. \quad \blacktriangleleft$$

Remarque 2.3.12. Dans le cas de la fonction de Runge $f(x) = \frac{1}{1+25x^2}$ sur $[-1,1]$, si on utilise les nœuds de Tchebychev, notre majoration de l'erreur devient :

$$\|f - P_n\|_\infty = \mathrm{O}\left(\left(\frac{5}{2}\right)^{n+1}\right).$$

C'est mieux que le $\mathrm{O}\big((10/\mathrm{e})^{n+1}\big)$ des nœuds équidistants, mais ça ne suffit toujours pas pour la convergence. Néanmoins, on peut montrer que les polynômes d'interpolation sur les nœuds de Tchebychev convergent bien uniformément vers f (c'est en fait le cas pour toute fonction Lipschitzienne). À titre de comparaison, la figure 2.4 montre les polynômes interpolateurs de f aux nœuds équidistants et aux nœuds de Tchebychev, pour $n = 10$. On voit qu'avec les nœuds de Tchebychev, le phénomène de Runge (oscillations aux bords) disparait. ◀

Nœuds	$\|\Pi_n\|_\infty$
Quelconques	$\mathrm{O}\big((b-a)^{n+1}\big)$
Équidistants	$\left(\frac{b-a}{\mathrm{e}}\right)^{n+1}$
Tchebychev	$\left(\frac{b-a}{4}\right)^{n+1}$

Table 2.1 – Ordre de grandeur de $\|\Pi_n\|_\infty$ pour les différents choix de nœuds. Comme $\mathrm{e} \approx 2{,}718 < 4$, les nœuds de Tchebychev ont une constante $\|\Pi_n\|_\infty$ beaucoup plus petite que les nœuds équidistants.

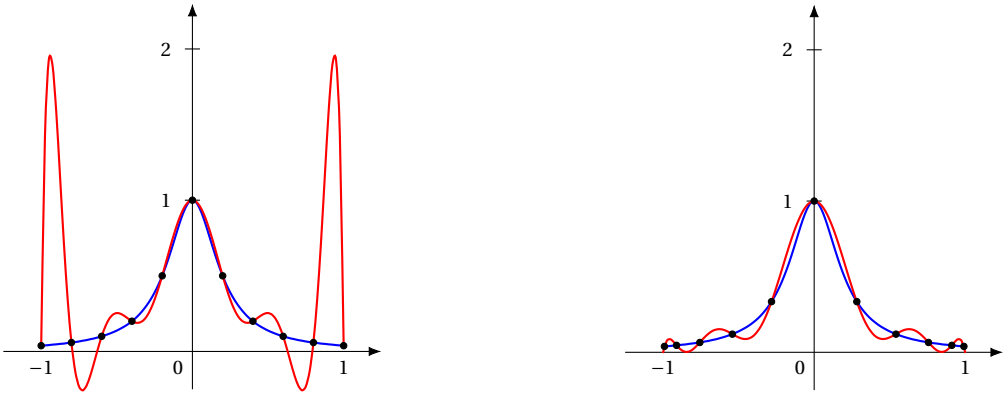


Figure 2.4 – Interpolation de la fonction de Runge $f(x) = \frac{1}{1+25x^2}$ (en bleu) par un polynôme de degré 10 (en rouge), aux nœuds équidistants (à gauche) et aux nœuds de Tchebychev (à droite). Les points noirs marquent les nœuds d'interpolation.

2.4 Stabilité de l'interpolation de Lagrange

Jusqu'ici, on s'est intéressé à la qualité de l'approximation de f par P_n, en supposant les valeurs $f(x_i)$ connues exactement. En pratique, on n'en connait qu'une approximation : erreur de mesure si f provient de données expérimentales, ou simplement erreur d'arrondi si $f(x_i)$ est calculée en machine.

Soit $\tilde{y}_i$ une valeur approchée de $f(x_i)$. Notons $\tilde{P}_n$ le polynôme interpolateur prenant la valeur $\tilde{y}_i$ en x_i. On souhaite estimer l'écart entre $\tilde{P}_n$ et P_n, le polynôme interpolateur pour les valeurs exactes. Soit $x \in [a,b]$, on a :

$$\begin{aligned} |P_n(x) - \tilde{P}_n(x)| &= \left| \sum_{i=0}^{n} (f(x_i) - \tilde{y}_i) L_i(x) \right| \\ &\leq \sum_{i=0}^{n} \left| f(x_i) - \tilde{y}_i \right| |L_i(x)| \\ &\leq \max_{0 \leq i \leq n} \left| f(x_i) - \tilde{y}_i \right| \times \sum_{i=0}^{n} |L_i(x)|. \end{aligned}$$

Si on note $\Lambda_n := \sup_{x \in [a,b]} \sum_{i=0}^{n} |L_i(x)|$, alors :

$$\| P_n - \tilde{P}_n \|_\infty \leq \Lambda_n \max_{0 \leq i \leq n} \left| f(x_i) - \tilde{y}_i \right|. \tag{2.2}$$

Cette constante Λ_n, qui ne dépend que des nœuds $(x_i)_{0 \leq i \leq n}$, s'interprète comme le facteur d'amplification des erreurs du procédé d'interpolation de Lagrange. Plus elle est grande, plus l'interpolation est instable : une petite erreur sur les valeurs $f(x_i)$ peut se traduire par une grosse erreur sur le polynôme d'interpolation.

Définition 2.4.1. On appelle **constante de Lebesgue** associée aux nœuds $x_0, \ldots, x_n$ le réel :

$$\Lambda_n := \sup_{x \in [a,b]} \sum_{i=0}^{n} |L_i(x)|,$$

où $(L_i)_{0 \leq i \leq n}$ sont les polynômes d'interpolation de Lagrange des $(x_i)_{0 \leq i \leq n}$.

La majoration (2.2) n'est pas améliorable : il existe des fonctions pour lesquelles c'est une égalité. C'est ce qu'on démontre dans le théorème suivant.

Théorème 2.4.2. Soit $\mathscr{L}_n : \mathscr{C}([a,b]) \to \mathbb{R}_n[X]$ l'application qui à toute fonction associe son polynôme interpolateur de Lagrange aux nœuds $x_0, \dots, x_n$. On munit $\mathscr{C}([a,b])$ et $\mathbb{R}_n[X]$ de la norme $\|\cdot\|_\infty$ sur $[a,b]$. Alors $\mathscr{L}_n$ est une application linéaire continue, et :

$$\sup_{f \in \mathscr{C}([a,b]) \setminus \{0\}} \frac{\|\mathscr{L}_n[f]\|_\infty}{\|f\|_\infty} = \Lambda_n,$$

où Λ_n est la constante de Lebesgue.

Remarque 2.4.3.

1. Plus généralement, si $(E, \|\cdot\|_E)$ et $(F, \|\cdot\|_F)$ sont des e.v.n. et $\Phi : E \to F$ une application linéaire, alors Φ est continue si et seulement s'il existe $C > 0$ tel que $\forall x \in E$, $\|\Phi(x)\|_F \leq C\|x\|_E$. On définit alors la norme subordonnée de Φ comme :

   $$|||\Phi||| := \sup_{x \in E \setminus \{0_E\}} \frac{\|\Phi(x)\|_F}{\|x\|_E}.$$

   On a donc :

   $$\forall x \in E, \quad \|\Phi(x)\|_F \leq |||\Phi|||\,\|x\|_E,$$

   et $|||\Phi|||$ est la plus petite constante pour laquelle l'inégalité ci-dessus est vraie pour tout $x \in E$. On peut montrer que $|||\cdot|||$ définit une norme sur $\mathscr{L}_c(E,F)$, l'espace des applications linéaires continues de E dans F.
2. Le théorème précédent affirme donc que $|||\mathscr{L}_n||| = \Lambda_n$. Ainsi, $\|\mathscr{L}_n[f]\|_\infty \leq \Lambda_n \|f\|_\infty$ pour toute fonction, et Λ_n est la plus petite constante pour laquelle cette inégalité est toujours vraie.
3. Si $\tilde{f}(x)$ est la valeur approchée de $f(x)$ qu'on calcule en pratique, avec une erreur $\|\tilde{f} - f\|_\infty \leq \varepsilon$, alors l'erreur sur le polynôme d'interpolation est majorée par :

   $$\|\mathscr{L}_n[\tilde{f}] - \mathscr{L}_n[f]\|_\infty = \|\mathscr{L}_n[\tilde{f} - f]\|_\infty \leq \Lambda_n \|\tilde{f} - f\|_\infty \leq \Lambda_n \varepsilon.$$

4. On peut montrer que $\Lambda_n \to +\infty$ quels que soient les nœuds : l'interpolation est une procédure de plus en plus instable quand on augmente le nombre de nœuds. En revanche, la vitesse à laquelle Λ_n diverge dépend du choix des nœuds. ◂

Démonstration.
• *Linéarité.* Soient $f, g \in \mathscr{C}([a,b])$ et $\lambda \in \mathbb{R}$. Par définition, $\mathscr{L}_n[f](x_i) = f(x_i)$ et $\mathscr{L}_n[g] = g(x_i)$ pour tout $i \in [\![0,n]\!]$. Posons $P_n := \lambda\mathscr{L}_n[f] + \mathscr{L}_n[g]$. Alors $P_n \in \mathbb{R}_n[X]$ car $\mathscr{L}_n[f], \mathscr{L}_n[g] \in \mathbb{R}_n[X]$. De plus :

$$\forall i \in [\![0,n]\!], \quad P_n(x_i) = \lambda\mathscr{L}_n[f](x_i) + \mathscr{L}_n[g](x_i) = \lambda f(x_i) + g(x_i).$$

Par unicité du polynôme interpolateur de Lagrange, on a $P_n = \mathscr{L}_n[\lambda f + g]$, c'est-à-dire :

$$\lambda\mathscr{L}_n[f] + \mathscr{L}_n[g] = \mathscr{L}_n[\lambda f + g].$$

• *Continuité.* Soit $f \in \mathscr{C}([a,b])$ et soit $x \in [a,b]$. On a :

$$\begin{aligned} |\mathscr{L}_n[f](x)| &= \left| \sum_{i=0}^{n} f(x_i) L_i(x) \right| \\ &\leq \sum_{i=0}^{n} |f(x_i)||L_i(x)| \\ &\leq \|f\|_\infty \sum_{i=0}^{n} |L_i(x)| \\ &\leq \|f\|_\infty \times \Lambda_n. \end{aligned}$$

Cette majoration étant vraie pour tout $x \in [a,b]$, on en déduit que :

$$\|\mathscr{L}_n[f]\|_\infty \leq \Lambda_n \|f\|_\infty.$$

Par linéarité, on a donc pour tous $f, g \in \mathscr{C}([a,b])$:

$$\|\mathscr{L}_n[f] - \mathscr{L}_n[g]\|_\infty = \|\mathscr{L}_n[f-g]\|_\infty \leq \Lambda_n \|f-g\|_\infty.$$

Ainsi, $\mathscr{L}_n$ est Lipschitzienne, donc elle est continue.
• *Norme subordonnée.* D'après le calcul précédent, on a :

$$\forall f \in \mathscr{C}([a,b]) \setminus \{0\}, \quad \frac{\|\mathscr{L}_n[f]\|_\infty}{\|f\|_\infty} \leq \Lambda_n.$$

Nous allons construire une fonction $f \in \mathscr{C}([a,b]) \setminus \{0\}$ telle que $\|\mathscr{L}_n[f]\|_\infty = \Lambda_n \|f\|_\infty$, ce qui prouvera l'égalité :

$$\sup_{f \in \mathscr{C}([a,b]) \setminus \{0\}} \frac{\|\mathscr{L}_n[f]\|_\infty}{\|f\|_\infty} = \Lambda_n.$$

La fonction $x \mapsto \sum_{i=0}^n |L_i(x)|$ est continue sur $[a,b]$, donc elle atteint son maximum :

$$\exists \xi \in [a,b], \quad \sum_{i=0}^n |L_i(\xi)| = \max_{x \in [a,b]} \sum_{i=0}^n |L_i(x)| = \Lambda_n.$$

On définit f affine par morceaux telle que $\|f\|_\infty = 1$ et $f(x_i) = \operatorname{sgn}(L_i(\xi))$.

$$\mathscr{L}_n[f](\xi) = \sum_{i=0}^n f(x_i) L_i(\xi) = \sum_{i=0}^n \operatorname{sgn}(L_i(\xi)) L_i(\xi) = \sum_{i=0}^n |L_i(\xi)| = \Lambda_n = \Lambda_n \|f\|_\infty.$$

Par conséquent, $\|\mathscr{L}_n[f]\|_\infty \geq \mathscr{L}_n[f](\xi) = \Lambda_n \|f\|_\infty$. L'inégalité contraire étant toujours vraie, on a montré que pour cette fonction f, on a $\|\mathscr{L}_n[f]\|_\infty = \Lambda_n \|f\|_\infty$, ce qui conclut la preuve. □

Exemple 2.4.4.

1. Pour des nœuds équidistants, on peut montrer que $\Lambda_n \sim \frac{2^{n+1}}{e n \ln(n)}$. La constante de Lebesgue croît exponentiellement vite vers $+\infty$. Par conséquent, l'interpolation sur des nœuds équidistants devient rapidement très instable. C'est pourquoi même si $\|f - P_n\|_\infty \to 0$, il n'est pas recommandé de faire une interpolation avec beaucoup de nœuds dans le cas équidistant.
2. Pour les nœuds de Tchebychev, on peut montrer que $\Lambda_n \sim \frac{2}{\pi} \ln(n)$. La constante Λ_n croît toujours vers $+\infty$, mais très lentement. En pratique (c'est-à-dire pour des valeurs de n modérées), on peut considérer que l'interpolation sur des nœuds de Tchebychev est stable. ◂

2.5 Théorème d'approximation de Weierstrass

On l'a vu avec l'exemple de Runge, les polynômes d'interpolation de Lagrange ne convergent pas nécessairement vers f lorsque $n \to +\infty$. Néanmoins, le théorème d'approximation de Weierstrass affirme qu'il existe toujours une suite de polynômes (mais pas forcément des polynômes d'interpolation) qui converge uniformément vers f sur $[a,b]$.

Théorème 2.5.1 (approximation de Weierstrass)**.** Pour toute fonction continue $f : [a,b] \to \mathbb{R}$, il existe une suite de polynômes $(P_n)_{n \in \mathbb{N}}$ qui converge uniformément vers f sur $[a,b]$.

Remarque 2.5.2. Si on note F l'ensemble des fonctions polynomiales définies sur $[a,b]$, le théorème d'approximation de Weierstrass est équivalent aux affirmations suivantes :

1. $\forall f \in \mathscr{C}([a,b]),\ \forall \varepsilon > 0,\ \exists P \in F,\ \|f - P\|_\infty \leq \varepsilon$.
2. $\overline{F} = \mathscr{C}([a,b])$, où $\overline{F}$ désigne l'adhérence de F

On dit que F est dense dans $\mathscr{C}([a,b])$. ◂

Il existe plusieurs démonstrations du théorème d'approximation de Weierstrass. On donne ici une démonstration constructive, basée sur les polynômes de Bernstein, donnant explicitement une suite de polynômes qui converge vers f.

Définition 2.5.3. On appelle **polynômes de Bernstein** associé à f les polynômes :

$$\forall n \in \mathbb{N}, \quad B_n[f](X) := \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k} X^k (1-X)^{n-k},$$

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

Démonstration du théorème 2.5.1. Il suffit de prouver le théorème pour $[a,b] = [0,1]$. En effet, supposons que le théorème est vrai sur $[0,1]$, et montrons qu'il est vrai sur tout intervalle $[a,b]$. Si $f \in \mathscr{C}([a,b])$, on pose $g(t) := f(a + t(b-a))$ pour tout $t \in [0,1]$. La fonction g est continue sur $[0,1]$, donc il existe une suite de polynômes $(Q_n)_{n\in\mathbb{N}}$ qui converge uniformément vers g sur $[0,1]$. On pose alors $P_n(x) := Q_n\left(\frac{x-a}{b-a}\right)$. La suite de polynômes $(P_n)_{n\in\mathbb{N}}$ converge uniformément vers f, en vertu de l'égalité :

$$\sup_{x\in[a,b]} \left|f(x) - P_n(x)\right| = \sup_{x\in[a,b]} \left|g\left(\frac{x-a}{b-a}\right) - Q_n\left(\frac{x-a}{b-a}\right)\right| = \sup_{t\in[0,1]} \left|g(t) - Q_n(t)\right| \to 0.$$

Démontrons le théorème sur $[0,1]$. Soit $f \in \mathscr{C}([0,1])$. On va montrer que la suite des polynômes de Bernstein associés à f converge uniformément vers f sur $[0,1]$.
Soient $x \in [0,1]$ et $\varepsilon > 0$. On considère S_n une variable aléatoire de loi binomiale $\mathscr{B}(n,x)$, et on note $X_n = \frac{1}{n}S_n$. Alors par le théorème de transfert, on a $B_n[f](x) = \mathbb{E}[f(X_n)]$. Par linéarité de l'espérance et par inégalité triangulaire :

$$\left|f(x) - B_n[f](x)\right| = \left|\mathbb{E}[f(x)] - \mathbb{E}[f(X_n)]\right| \leq \mathbb{E}\left[\left|f(x) - f(X_n)\right|\right].$$

Soit $\delta > 0$, on considère l'évènement $A_\delta := \{|X_n - x| \geq \delta\}$ et on note $\mathbb{1}_{A_\delta}$ la fonction indicatrice de A_δ. On décompose par linéarité :

$$\mathbb{E}\left[\left|f(x) - f(X_n)\right|\right] = \mathbb{E}\left[\left|f(x) - f(X_n)\right|\mathbb{1}_{A_\delta}\right] + \mathbb{E}\left[\left|f(x) - f(X_n)\right|\mathbb{1}_{A_\delta^c}\right].$$

On étudie ces deux termes séparément :

- D'une part, pour tout $\delta > 0$, on a :

  $$\mathbb{E}\left[\left|f(x) - f(X_n)\right|\mathbb{1}_{A_\delta}\right] \leq \mathbb{E}\left[2\|f\|_\infty \mathbb{1}_{A_\delta}\right] = 2\|f\|_\infty \mathbb{E}\left[\mathbb{1}_{A_\delta}\right] = 2\|f\|_\infty \mathbb{P}(A_\delta).$$

  On utilise alors l'inégalité de Bienaymé–Tchebychev :

  $$\mathbb{P}(A_\delta) = \mathbb{P}(|X_n - x| \geq \delta) = \mathbb{P}(|S_n - nx| \geq n\delta) \leq \frac{\mathrm{Var}(S_n)}{n^2\delta^2} = \frac{x(1-x)}{n\delta^2} \leq \frac{1}{4n\delta^2},$$

  car $x(1-x) \leq \frac{1}{4}$ sur $[0,1]$. Par conséquent :

  $$\mathbb{E}\left[\left|f(x) - f(X_n)\right|\mathbb{1}_{A_\delta}\right] \leq \frac{\|f\|_\infty}{2n\delta^2}. \tag{2.3}$$

- D'autre part, f est uniformément continue sur $[0,1]$ par le théorème de Heine, donc il existe $\delta(\varepsilon) > 0$ tel que :

  $$\forall u, v \in [0,1], \quad |u - v| < \delta(\varepsilon) \implies |f(u) - f(v)| < \frac{\varepsilon}{2}.$$

  On applique cette proposition avec $u = x$ et $v = X_n$: sur l'évènement $A_{\delta(\varepsilon)}^c$, on a $|x - X_n| < \delta(\varepsilon)$, donc $|f(x) - f(X_n)| < \frac{\varepsilon}{2}$. Par conséquent :

  $$\mathbb{E}\left[\left|f(x) - f(X_n)\right|\mathbb{1}_{A_{\delta(\varepsilon)}^c}\right] \leq \mathbb{E}\left[\frac{\varepsilon}{2}\mathbb{1}_{A_{\delta(\varepsilon)}^c}\right] \leq \frac{\varepsilon}{2}. \tag{2.4}$$

En combinant les majorations (2.3) et (2.4) avec $\delta = \delta(\varepsilon)$, on obtient :

$$\forall x \in [0,1], \quad \left|f(x) - B_n[f](x)\right| \leq \frac{\|f\|_\infty}{2n\delta(\varepsilon)^2} + \frac{\varepsilon}{2}.$$

Le majorant est uniforme en x, donc :

$$\left\|f - B_n[f]\right\|_\infty \leq \frac{\|f\|_\infty}{2n\delta(\varepsilon)^2} + \frac{\varepsilon}{2}.$$

Ainsi, pour tout $n \geq \frac{\|f\|_\infty}{\varepsilon\delta(\varepsilon)^2}$, on a $\left\|f - B_n[f]\right\|_\infty \leq \varepsilon$. □

Remarque 2.5.4. Pour une fonction Lipschitzienne, on peut montrer que :

$$\left\|f - B_n[f]\right\|_\infty = \mathrm{O}\left(\frac{1}{\sqrt{n}}\right),$$

ce qui est sous-optimal. Les polynômes d'approximation de Jackson, d'autres polynômes conçus pour approcher f mais dont la construction est plus délicate, ont une vitesse de convergence en $\mathrm{O}\left(\frac{1}{n}\right)$ (voir Demailly, 2016, chap. II, section 3). ◂

2.6 Approximation quadratique

2.6.a Norme intégrale à poids

On s'intéresse maintenant à l'approximation polynomiale pour la norme 2 sur $[a,b]$:

$$\|f\|_2 := \left(\int_a^b f(x)^2 \,\mathrm{d}x\right)^{\frac{1}{2}}.$$

Contrairement à la norme $\|\cdot\|_\infty$, cette norme est euclidienne, c'est-à-dire qu'elle provient d'un produit scalaire :

$$\|f\|_2 = \sqrt{\langle f, f\rangle}, \quad \text{avec} \quad \langle f, g\rangle := \int_a^b f(x)g(x)\,\mathrm{d}x.$$

Plus généralement, on s'intéressera à des normes euclidiennes définies par une intégrale sur un intervalle I non nécessairement fermé ni borné.

Définition 2.6.1. Soit $w\colon I \to \mathbb{R}_+^*$ une fonction continue strictement positive, appelée fonction **poids**, telle que :

$$\forall n \in \mathbb{N}, \quad \int_I x^{2n} w(x)\,\mathrm{d}x \text{ converge.}$$

On note $\mathscr{C}(I,w)$ l'ensemble des fonctions f continues sur I pour lesquelles :

$$\int_I f(x)^2 w(x)\,\mathrm{d}x \text{ converge.}$$

Sur $\mathscr{C}(I,w)$, on définit le produit scalaire $\langle\cdot,\cdot\rangle_w$:

$$\forall f,g \in \mathscr{C}(I,w), \quad \langle f,g\rangle_w := \int_I f(x)g(x)w(x)\,\mathrm{d}x.$$

On note $\|\cdot\|_w$ la norme associée :

$$\forall f \in \mathscr{C}(I,w), \quad \|f\|_w = \sqrt{\langle f,f\rangle_w} = \left(\int_I f(x)^2 w(x)\,\mathrm{d}x\right)^{\frac{1}{2}}.$$

Proposition 2.6.2. L'ensemble $\mathscr{C}(I,w)$ est un sous-espace vectoriel de $\mathscr{C}(I)$, et l'ensemble des fonctions polynomiales est un sous-espace vectoriel de $\mathscr{C}(I,w)$. De plus, l'application $\langle\cdot,\cdot\rangle_w$ définie ci-dessus est bien définie sur $\mathscr{C}(I,w)$ et est un produit scalaire. ◀

Démonstration.
• $\mathscr{C}(I,w)$ *est un s.e.v.* La fonction nulle appartient à $\mathscr{C}(I,w)$. Soient $f,g\in\mathscr{C}(I,w)$ et $\lambda\in\mathbb{R}$, montrons que $\lambda f+g\in\mathscr{C}(I,w)$. On utilise l'inégalité :

$$\forall x\in I,\quad (\lambda f(x)+g(x))^2\leq 2\lambda^2 f(x)^2+2g(x)^2.$$

Or, les intégrales :

$$\int_I f(x)^2 w(x)\,\mathrm{d}x \quad\text{et}\quad \int_I g(x)^2\,\mathrm{d}x$$

convergent car $f,g\in\mathscr{C}(I,w)$, donc par comparaison l'intégrale :

$$\int_I (\lambda f(x)+g(x))^2 w(x)\,\mathrm{d}x$$

est convergente, c'est-à-dire $\lambda f+g\in\mathscr{C}(I,w)$. Ainsi, $\mathscr{C}(I,w)$ est un sous-espace vectoriel de $\mathscr{C}(I)$.

• $\mathscr{C}(I,w)$ *contient les polynômes.* Par hypothèse, les fonctions monômes $x\mapsto x^n$ appartiennent à $\mathscr{C}(I,w)$. Par stabilité par combinaisons linéaires, les fonctions polynômes appartiennent donc aussi à $\mathscr{C}(I,w)$.

• $\langle\cdot,\cdot\rangle_w$ *est un produit scalaire.* Montrons que $\langle\cdot,\cdot\rangle_w$ est bien définie sur $\mathscr{C}(I,w)$. Soient $f,g\in\mathscr{C}(I,w)$, on utilise l'inégalité :

$$\forall x\in I,\quad |f(x)g(x)|\leq\frac{1}{2}\big(f(x)^2+g(x)^2\big).$$

Par comparaison, l'intégrale :

$$\int_I |f(x)g(x)|w(x)\,\mathrm{d}x,$$

est convergente. Ainsi, l'intégrale définissant $\langle f,g\rangle_w$ est absolument convergente.
L'application $\langle\cdot,\cdot\rangle_w$ est clairement bilinéaire, symétrique et positive, montrons qu'elle est définie positive. Soit $f\in\mathscr{C}(I,w)$ telle que $\langle f,f\rangle_w=0$. Alors :

$$\int_I f(x)^2 w(x)\,\mathrm{d}x=0,$$

donc par continuité, $f(x)^2w(x)=0$ pour tout $x\in I$ (cf. cours d'intégration). Puisque w ne s'annule pas, on a $f(x)=0$ pour tout $x\in I$. Ainsi, on a montré que $\langle\cdot,\cdot\rangle_w$ est un produit scalaire sur $\mathscr{C}(I,w)$. □

Exemple 2.6.3. Quelques exemples classiques de fonctions poids.

1. $I=[a,b]$ et $w(x)=1$. Dans ce cas, $\mathscr{C}(I,w)=\mathscr{C}([a,b])$ et $\|\cdot\|_w$ est la norme 2 sur $[a,b]$:

$$\|f\|_w=\|f\|_2=\left(\int_a^b f(x)^2\,\mathrm{d}x\right)^{\frac{1}{2}}.$$

2. $I=]-1,1[$ et $w(x)=\frac{1}{\sqrt{1-x^2}}$. Le changement de variable $x=\cos\theta$ donne :

$$\int_{-1}^1 P(x)^2\frac{1}{\sqrt{1-x^2}}\,\mathrm{d}x=\int_0^\pi P(\cos\theta)^2\,\mathrm{d}\theta.$$

Par conséquent, l'intégrale est convergente pour tout polynôme P.

3. $I = \mathbb{R}_+$ et $w(x) = \mathrm{e}^{-x}$. Par croissance comparée, l'intégrale :

$$\int_0^{+\infty} P(x)^2 \mathrm{e}^{-x} \, \mathrm{d}x,$$

est bien convergente pour tout polynôme P.

4. $I = \mathbb{R}$ et $w(x) = \mathrm{e}^{-x^2}$. De même :

$$\int_{-\infty}^{+\infty} P(x)^2 \mathrm{e}^{-x^2} \, \mathrm{d}x,$$

converge par croissance comparée, quel que soit le polynôme P. ◀

2.6.b Polynômes orthogonaux

Rappelons quelques propriétés des espaces euclidiens (voir le cours d'algèbre bilinéaire). Soit E un espace euclidien, c'est-à-dire un espace vectoriel réel de dimension finie muni d'un produit scalaire $\langle \cdot, \cdot \rangle$, dont on note $\|\cdot\|$ la norme associée. Si F est un sous-espace vectoriel de E, on définit son orthogonal par :

$$F^\perp := \{ y \in E \mid \forall x \in F, \ \langle x, y \rangle = 0 \}.$$

C'est un sous-espace vectoriel de E. De plus, F et $F^\perp$ sont supplémentaires dans E :

$$E = F \oplus F^\perp.$$

Par conséquent, tout vecteur de E se décompose de manière unique comme la somme d'un vecteur de F et d'un vecteur de $F^\perp$. On note p_F le projecteur sur F parallèlement à $F^\perp$, ce projecteur est appelé le projecteur orthogonal sur F. Le projecteur orthogonal sur $F^\perp$ est alors $\mathrm{id}_E - p_F$. Pour tout $x \in E$, $p_F(x)$ est l'unique vecteur de F tel que $x - p_F(x)$ soit orthogonal à tous les vecteurs de F.

Théorème 2.6.4. Soit E un espace euclidien et soit F un sous-espace vectoriel de dimension finie. Soit $x \in E$ et soit $p_F(x)$ le projeté orthogonal de x sur F. Alors :

$$\forall y \in F, \quad \|x - p_F(x)\| \leq \|x - y\|$$

avec égalité si et seulement si $y = p_F(x)$.

Démonstration. Soit $x \in E$ et $y \in F$. Par le théorème de Pythagore :

$$\|x - y\|^2 = \|x - p_F(x) + p_F(x) - y\|^2 = \|x - p_F(x)\|^2 + \|p_F(x) - y\|^2,$$

car $p_F(x) - y \in F$ et $x - p_F(x) \in F^\perp$. Par conséquent :

$$\|x - y\|^2 \geq \|x - p_F(x)\|^2$$

avec égalité si et seulement si $\|p_F(x) - y\|^2 = 0$, c'est-à-dire si $p_F(x) = y$. □

Autrement dit, dans un espace euclidien, la distance d'un vecteur x à un sous-espace F est atteinte en $p_F(x)$, et seulement en $p_F(x)$. Appliquons ce théorème à l'approximation polynomiale.

Théorème 2.6.5. Pour tout $f \in \mathscr{C}(I, w)$, il existe un unique $Q_n \in \mathbb{R}_n[X]$ tel que :

$$\|f - Q_n\|_w = \inf_{P \in \mathbb{R}_n[X]} \|f - P\|_w.$$

De plus, Q_n est l'unique polynôme de $\mathbb{R}_n[X]$ satisfaisant $\langle f - Q_n, P \rangle_w$ pour tout $P \in \mathbb{R}_n[X]$.

Démonstration. Notons E l'espace vectoriel engendré par f et par les polynômes de degré inférieurs ou égal à n. Alors E est de dimension finie, c'est un espace euclidien pour le produit scalaire $\langle\cdot,\cdot\rangle_w$, et $\mathbb{R}_n[X]$ est un sous-espace de E. Notons Q_n le projeté orthogonal de f sur $\mathbb{R}_n[X]$. On applique le théorème de projection orthogonale : la distance de f au sous-espace $\mathbb{R}_n[X]$ est atteinte en Q_n, et Q_n est l'unique élément de $\mathbb{R}_n[X]$ pour lequel cette distance est atteinte. □

Définition 2.6.6. On appelle **polynôme de meilleure approximation quadratique** de f à l'ordre n pour le poids w, l'unique polynôme $Q_n \in \mathbb{R}_n[X]$ tel que :

$$\|f - Q_n\|_w = \min_{P\in\mathbb{R}_n[X]} \|f - P\|_w$$

Il est facile de déterminer le projeté orthogonal d'un vecteur x sur un sous-espace F si on connait une base orthonormée $(e_1,\dots,e_n)$ de F :

$$p_F(x) = \sum_{i=1}^{n} \langle x, e_i\rangle e_i.$$

Si la base $(e_1,\dots,e_n)$ est seulement orthogonale, alors $\left(\frac{e_1}{\|e_1\|},\dots,\frac{e_n}{\|e_n\|}\right)$ est orthonormée, donc :

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

Il est toujours possible de trouver une base orthonormée de F grâce au procédé de Gram–Schmidt.

Procédé de Gram–Schmidt. Soit $(u_1,\dots,u_n)$ une base de F. On définit par récurrence :

$$\begin{cases} \tilde{e}_1 = u_1 \\ \tilde{e}_k := u_k - \displaystyle\sum_{i=1}^{k-1} \frac{\langle u_k, \tilde{e}_i\rangle}{\|\tilde{e}_i\|^2} \tilde{e}_i. \end{cases}$$

Les vecteurs ainsi obtenus vérifient :

1. $\langle \tilde{e}_k, \tilde{e}_\ell\rangle = 0$ pour tous $k \neq \ell$ (famille orthogonale) ;
2. $\forall k \in [\![1, n]\!]$, $\mathrm{Vect}(u_1,\dots,u_k) = \mathrm{Vect}(\tilde{e}_1,\dots,\tilde{e}_k)$.

En particulier, $(\tilde{e}_k)_{1\le k\le n}$ est une base orthogonale de F. En posant $e_k := \frac{\tilde{e}_k}{\|\tilde{e}_k\|}$, on obtient une base orthonormée de F.

Plus généralement, on peut appliquer ce procédé à n'importe quelle famille libre de vecteurs, même infinie, pour obtenir une famille orthogonale. En appliquant le procédé de Gram–Schmidt à la famille $(X^k)_{k\in\mathbb{N}}$ de $\mathbb{R}[X]$ muni du produit scalaire $\langle\cdot,\cdot\rangle_w$, on obtient une famille de polynômes orthogonaux $(P_n)_{n\in\mathbb{N}}$ satisfaisant :

$$\forall n \in \mathbb{N}, \quad \mathrm{Vect}(P_0, P_1,\dots,P_n) = \mathrm{Vect}(1, X,\dots,X^n) = \mathbb{R}_n[X].$$

Ainsi, la famille $(P_0,\dots,P_n)$ est une base orthogonale de $\mathbb{R}_n[X]$.

Théorème 2.6.7. Il existe une unique suite de polynômes $(P_n)_{n\in\mathbb{N}}$ satisfaisant :

(i) $\forall n \in \mathbb{N}$, $\deg(P_n) = n$.

(ii) $\forall n \in \mathbb{N}$, P_n est unitaire.

(iii) $\forall k \neq \ell$, $\langle P_k, P_\ell\rangle_w = 0$.

La suite $(P_n)_{n\in\mathbb{N}}$ est appelée la suite des **polynômes orthogonaux unitaires** pour le poids w.

Démonstration. L'existence provient du procédé de Gram–Schmidt, comme décrit précédemment. Montrons l'unicité. Soient $(P_n)_{n\in\mathbb{N}}$ et $(\tilde{P}_n)_{n\in\mathbb{N}}$ deux suites de polynômes orthogonaux unitaires. Pour $n=0$, on a nécessairement $P_0=\tilde{P}_0=1$. Pour $n\geq 1$, on a $P_n-\tilde{P}_n\in\mathbb{R}_{n-1}[X]$ car P_n et $\tilde{P}_n$ sont unitaires, et aussi $P_n-\tilde{P}_n\in\mathbb{R}_{n-1}[X]^\perp$ par définition des polynômes orthogonaux. Par conséquent :

$$P_n-\tilde{P}_n\in\mathbb{R}_{n-1}[X]\cap\mathbb{R}_{n-1}[X]^\perp=\{0_{\mathbb{R}[X]}\},$$

donc $P_n=\tilde{P}_n$. □

Remarque 2.6.8. Plus généralement, on appelle polynômes orthogonaux toute suite $(\lambda_n P_n)_{n\in\mathbb{N}}$ avec $\lambda_n\neq 0$ et $(P_n)_{n\in\mathbb{N}}$ les polynômes orthogonaux unitaires. Le fait de demander que les P_n soient unitaires n'est qu'une façon d'assurer l'unicité. D'autres conditions peuvent servir à assurer l'unicité : on peut par exemple demander à ce que les polynômes soient normalisés, ou bien prescrire leur valeur en un point. ◀

Avec le procédé de Gram–Schmidt, le calcul de P_n fait intervenir tous les P_k avec $0\leq k\leq n-1$. En fait, on peut montrer que les polynômes orthogonaux peuvent se calculer par une récurrence d'ordre 2.

Proposition 2.6.9. La suite des polynômes orthogonaux $(P_n)_{n\in\mathbb{N}}$ satisfait la relation de récurrence :

$$\forall n\geq 2,\quad P_n(X)=(X-\lambda_n)P_{n-1}(X)-\mu_n P_{n-2}(X),$$

avec :

$$\lambda_n:=\frac{\langle XP_{n-1},P_{n-1}\rangle_w}{\|P_{n-1}\|_w^2},\qquad \mu_n:=\frac{\|P_{n-1}\|_w^2}{\|P_{n-2}\|_w^2}.$$ ◀

Démonstration. Le polynôme XP_{n-1} est de degré n, donc il s'écrit :

$$XP_{n-1}(X)=\sum_{k=0}^{n}\alpha_k P_k,$$

avec $(\alpha_k)_{0\leq k\leq n}$ ses coordonnées dans la base $(P_0,\dots,P_n)$. Puisque XP_{n-1} est unitaire, on a nécessairement $\alpha_n=1$. Les autres coordonnées sont données par :

$$\alpha_k=\frac{\langle XP_{n-1},P_k\rangle_w}{\|P_k\|_w^2},$$

car les P_k sont orthogonaux. Remarquons que :

$$\langle XP_{n-1},P_k\rangle_w=\int_I xP_{n-1}(x)P_k(x)\,\mathrm{d}x=\langle P_{n-1},XP_k\rangle_w.$$

Or, $XP_k\in\mathbb{R}_{k+1}[X]$ et $P_{n-1}\in\mathbb{R}_{n-2}[X]^\perp$, donc si $k\leq n-3$, $\langle P_{n-1},XP_k\rangle_w=0$. Ainsi, $\alpha_k=0$ pour tout $k\leq n-3$ et il ne reste que α_{n-1} et α_{n-2} :

$$\alpha_{n-1}=\frac{\langle XP_{n-1},P_{n-1}\rangle_w}{\|P_{n-1}\|_w^2}=:\lambda_n,\qquad \alpha_{n-2}=\frac{\langle P_{n-1},XP_{n-2}\rangle_w}{\|P_{n-2}\|_w^2}.$$

Enfin, on écrit $XP_{n-2}(X)=P_{n-1}(X)+Q(X)$ avec $Q\in\mathbb{R}_{n-2}[X]$ car XP_{n-2} est de degré $n-1$ et unitaire, donc :

$$\langle P_{n-1},XP_{n-2}\rangle_w=\langle P_{n-1},P_{n-1}\rangle_w+\langle P_{n-1},Q\rangle=\|P_{n-1}\|_w^2+0,$$

d'où :

$$\alpha_{n-2}=\frac{\|P_{n-1}\|_w^2}{\|P_{n-2}\|_w^2}=:\mu_n.$$

Finalement, on a montré que :

$$XP_{n-1}(X)=P_n(X)+\lambda_n P_{n-1}(X)+\mu_n P_{n-2}(X),$$

ce qu'on réécrit :

$$P_n(X)=(X-\lambda_n)P_{n-1}(X)-\mu_n P_{n-2}(X).$$ □

Polynômes	Poids $w(x)$	Intervalle I
Legendre	1	$[-1,1]$
Tchebychev	$\frac{1}{\sqrt{1-x^2}}$	$]-1,1[$
Laguerre	e^{-x}	$\mathbb{R}_+$
Hermite	$\mathrm{e}^{-\frac{x^2}{2}}$	$\mathbb{R}$

Table 2.2 – Quelques polynômes orthogonaux classiques

Exemple 2.6.10 (Polynômes de Legendre)**.** On appelle **polynômes de Legendre** les polynômes orthogonaux pour le poids $w(x) = 1$ sur $[-1,1]$, et satisfaisant la condition $P_n(1) = 1$. Vous étudierez ces polynômes en TD. Ils satisfont la relation de récurrence :

$$\forall n \geq 1, \quad (n+1)P_{n+1}(X) = (2n+1)XP_n(X) - nP_{n-1}(X). \qquad \blacktriangleleft$$

Exemple 2.6.11 (Polynômes de Tchebychev)**.** Les **polynômes de Tchebychev** $(T_n)_{n\in\mathbb{N}}$ ont été définis dans la définition 2.3.7. Ces polynômes sont également les polynômes orthogonaux pour le poids $w(x) = \frac{1}{\sqrt{1-x^2}}$ sur $]-1,1[$ satisfaisant la condition $T_n(1) = 1$. En effet, par le changement de variable $x = \cos(\theta)$:

$$\langle T_k, T_\ell \rangle_w = \int_{-1}^{1} T_k(x)T_\ell(x)\frac{1}{\sqrt{1-x^2}}\,\mathrm{d}x = \int_0^\pi T_k(\cos\theta)T_\ell(\cos\theta)\,\mathrm{d}\theta = \int_0^\pi \cos(k\theta)\cos(\ell\theta)\,\mathrm{d}\theta.$$

En utilisant la formule $\cos(a)\cos(b) = \frac{\cos(a+b)+\cos(a-b)}{2}$:

$$\int_0^\pi \cos(k\theta)\cos(\ell\theta)\,\mathrm{d}\theta = \frac{1}{2}\int_0^\pi \cos((k+\ell)\theta) + \cos((k-\ell)\theta)\,\mathrm{d}\theta = 0 \text{ si } k \neq \ell.$$

Enfin, $T_n(1) = \cos(n\arccos 1) = \cos(0) = 1$. Attention, les polynômes $(T_n)_{n\in\mathbb{N}}$ ne sont pas unitaires (sauf T_0 et T_1), le coefficient dominant de T_n est 2^{n-1} si $n \geq 1$. Ces polynômes satisfont également une relation de récurrence d'ordre 2, voir proposition 2.3.8. $\blacktriangleleft$

Exemple 2.6.12. D'autres familles de polynômes orthogonaux classiques :

1. Les **polynômes de Laguerre** sont orthogonaux pour le poids $w(x) = \mathrm{e}^{-x}$ sur $I = \mathbb{R}_+$.
2. Les **polynômes d'Hermite** sont orthogonaux pour le poids $w(x) = \mathrm{e}^{-\frac{x^2}{2}}$ sur $I = \mathbb{R}$. $\blacktriangleleft$

On obtient finalement l'expression du polynôme de meilleure approximation quadratique de f.

Théorème 2.6.13. Soit $f \in \mathscr{C}(I,w)$. Le polynôme de meilleure approximation quadratique de f à l'ordre n s'écrit dans la base des polynômes orthogonaux :

$$Q_n(X) = \sum_{k=0}^{n} \frac{\langle f, P_k\rangle_w}{\|P_k\|_w^2} P_k(X).$$

Enfin, on peut se demander si les polynômes de meilleure approximation quadratique convergent vers f lorsque $n \to +\infty$. C'est faux en général, mais c'est vrai pour les familles classiques de polynômes orthogonaux qu'on a citées jusqu'à présent (voir table 2.2). On le démontre dans le cas où l'intervalle I est borné (ce qui couvre le cas des polynômes de Legendre et des polynômes de Tchebychev).

Proposition 2.6.14. Si I est borné, alors pour tout $f \in \mathscr{C}(I,w)$, on a $\|f - Q_n\|_w \to 0$ avec Q_n le polynôme de meilleure approximation de f à l'ordre n. $\blacktriangleleft$

Démonstration. Notons Q_n le polynôme de meilleure approximation quadratique de f à l'ordre n. On note $a = \inf I$ et $b = \sup I$ (a et b n'appartiennent pas nécessairement à I).

• Supposons d'abord que f est continue sur $\overline{I} = [a,b]$ (l'adhérence de I). Par le théorème d'approximation de Weierstrass, soit $(B_n)_{n\in\mathbb{N}}$ une suite de polynômes qui converge uniformément vers f, et telle que $\deg(B_n) \leq n$ pour tout $n \in \mathbb{N}$ (par exemple les polynômes de Bernstein associés à f). Par définition de Q_n, on a :

$$\|f - Q_n\|_w^2 \leq \|f - B_n\|_w^2.$$

Or sur $[a,b]$, on peut majorer la norme $\|\cdot\|_w$ par la norme $\|\cdot\|_\infty$:

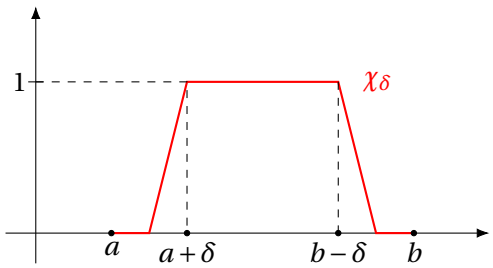
$$\|f - B_n\|_w^2 = \int_a^b |f(x) - B_n(x)|^2 w(x)\,\mathrm{d}x \leq \|f - B_n\|_\infty^2 \int_a^b w(x)\,\mathrm{d}x,$$

donc $\|f - B_n\|_w \leq C\|f - B_n\|_\infty$ avec $C := \left(\int_a^b w(x)\,\mathrm{d}x\right)^{\frac{1}{2}}$ une constante finie (l'intégrale converge par définition d'une fonction poids). Par conséquent :

$$\|f - Q_n\|_w^2 \leq C_w \|f - B_n\|_\infty \to 0,$$

car $(B_n)_{n\in\mathbb{N}}$ converge uniformément vers f. Donc $(Q_n)_{n\in\mathbb{N}}$ converge vers f pour la norme $\|\cdot\|_w$.

• On traite à présent le cas général. Soit $\varepsilon > 0$. Pour tout $\delta > 0$, on note χ_δ la fonction affine par morceaux qui vaut 1 sur $[a+\delta, b-\delta]$ et qui vaut 0 en dehors de $[a+\frac{\delta}{2}, b-\frac{\delta}{2}]$:



Puisque f est continue sur I, la fonction $f_\delta := f \times \chi_\delta$ est une fonction continue sur $[a,b]$, avec la convention $f_\delta(a) = f_\delta(b) := 0$. Soit $Q_{n,\delta}$ le polynôme de meilleure approximation quadratique de f_δ. Par définition de Q_n et par inégalité triangulaire, on a :

$$\|f - Q_n\|_w \leq \|f - Q_{n,\delta}\|_w \leq \|f - f_\delta\|_w + \|f_\delta - Q_{n,\delta}\|_w.$$

Étudions ces deux termes.

- On a $f - f_\delta = f \times (1 - \chi_\delta)$ avec $1 - \chi_\delta$ compris entre 0 et 1 et qui s'annule sur $[a+\delta, b-\delta]$. Par conséquent :

$$\|f - f_\delta\|_w^2 = \|f(1-\chi_\delta)\|_w^2 \leq \int_a^{a+\delta} f(x)^2 w(x)\,\mathrm{d}x + \int_{b-\delta}^b f(x)^2 w(x)\,\mathrm{d}x \xrightarrow[\delta\to 0^+]{} 0,$$

  car $\int_a^b f(x)^2 w(x)\,\mathrm{d}x$ est une intégrale convergente. Il existe donc $\delta(\varepsilon)$ tel que $\|f - f_{\delta(\varepsilon)}\|_w \leq \frac{\varepsilon}{2}$.
- Pour $\delta(\varepsilon)$ fixé, on sait d'après le premier point de la démonstration que $(Q_{n,\delta(\varepsilon)})_{n\in\mathbb{N}}$ converge vers $f_{\delta(\varepsilon)}$ pour la norme $\|\cdot\|_w$, car $f_{\delta(\varepsilon)}$ est continue sur $[a,b]$. Donc il existe $n(\varepsilon) \in \mathbb{N}$ tel que pour tout $n \geq n(\varepsilon)$, on a $\|f_{\delta(\varepsilon)} - Q_{n,\delta(\varepsilon)}\|_w \leq \frac{\varepsilon}{2}$.

Ainsi, on a montré que :

$$\forall \varepsilon > 0, \quad \exists n(\varepsilon) \in \mathbb{N}, \quad \forall n \geq n(\varepsilon), \quad \|f - Q_n\|_w \leq \varepsilon,$$

c'est-à-dire $\|f - Q_n\|_w \to 0$. □

Remarque 2.6.15. On a vu que le polynôme de meilleure approximation quadratique de f s'écrivait $Q_n = \sum_{k=0}^{n} \alpha_k P_k$ avec $\alpha_k := \frac{\langle f, P_k \rangle_w}{\|P_k\|_w^2}$. Si $(Q_n)_{n\in\mathbb{N}}$ converge vers f, on pourra écrire :

$$f = \sum_{k=0}^{+\infty} \alpha_k P_k.$$

Attention, cette série de fonctions converge au sens de la norme $\|\cdot\|_w$, c'est-à-dire que la suite des sommes partielles converge vers f pour cette norme :

$$\left\| f - \sum_{k=0}^{n} \alpha_k P_k \right\|_w \xrightarrow[n\to+\infty]{} 0.$$

A priori, cette série ne converge pas simplement (ni uniformément) : c'est un autre mode de convergence pour les séries de fonctions. ◀

Chapitre 3

Intégration numérique

Dans ce chapitre, on s'intéresse à l'évaluation numérique de l'intégrale d'une fonction continue f sur un intervalle $[a,b]$:

$$\int_a^b f(x)\,\mathrm{d}x.$$

Si on connait une primitive F de f, le théorème fondamental de l'analyse permet de calculer la valeur exacte de cette intégrale :

$$\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a).$$

Cependant, déterminer une primitive d'une fonction n'est pas toujours aisé. Pire, certaines fonctions, par exemple $x \mapsto \mathrm{e}^{-x^2}$, ne possèdent pas de primitives élémentaires [1] (théorème de Liouville–Rosenlicht en algèbre différentielle). Les méthodes numériques sont alors la seule façon de calculer l'intégrale. Enfin, les méthodes d'intégration numériques seront utilisées au chapitre 4 pour la résolution approchée d'équations différentielles.

On se limitera aux méthodes de quadrature [2] consistant à approcher l'intégrale de f par une somme pondérée de valeurs de f en des points donnés. On ne parlera pas des méthodes pour les intégrales multiples, ni des méthodes probabilistes comme la méthode de Monte-Carlo.

On note $\|\cdot\|_\infty$ la norme sup sur $[a,b]$.

3.1 Sommes de Riemann et méthodes des rectangles

Une première façon d'approcher numériquement une intégrale est d'utiliser des sommes de Riemann. Soit σ une subdivision pointée de l'intervalle $[a,b]$, c'est-à-dire $\sigma = \big((a_k)_{0\leq k\leq N}, (\xi_k)_{0\leq k\leq N-1}\big)$ où $(a_k)_{0\leq k\leq n}$ est une subdivision de $[a,b]$:

$$a = a_0 < a_1 < \cdots < a_{N-1} < a_N = b,$$

et $(\xi_k)_{0\leq k\leq N-1}$ est un choix de point dans chaque sous-intervalle de la subdivision :

$$\forall k \in [\![0, N-1]\!], \quad \xi_k \in [a_k, a_{k+1}].$$

La **somme de Riemann** associée à f et à la subdivision pointée σ est définie par :

$$S(f,\sigma) \coloneqq \sum_{k=0}^{N-1} f(\xi_k)(a_{k+1} - a_k).$$

1. la notion de « fonction élémentaire » est un peu subtile à définir. Dans les grandes lignes, il s'agit d'une fonction pouvant s'exprimer comme une « combinaison » de fractions rationnelles, d'exponentielles et de logarithmes.

2. le terme « quadrature » vient historiquement du calcul d'aires en géométrie. En géométrie euclidienne classique, calculer une aire revenait à construire à la règle (non graduée) et au compas un carré de même aire qu'une figure donnée, d'où le nom. C'est ce qui a donné ensuite les techniques d'approximation d'aires par des polygones, et finalement l'approximation d'intégrales, une intégrale étant l'aire sous la courbe représentative d'une fonction.

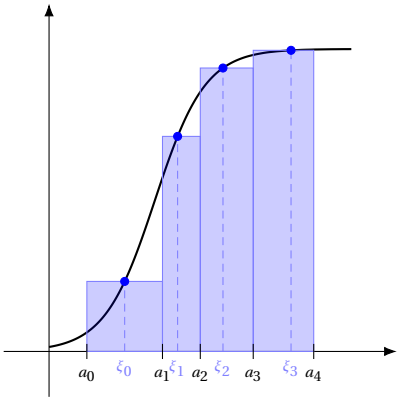


Figure 3.1 – Une somme de Riemann ($N = 4$). L'aire des rectangles (en bleu) approche l'aire sous la courbe.

Géométriquement, $S(f,\sigma)$ est la somme des aires (signées) des rectangles de hauteur $f(\xi_k)$ et de largeur $a_{k+1} - a_k$ (voir figure 3.1). Si on note :

$$|\sigma| := \max_{0\le k\le N-1} a_{k+1} - a_k,$$

le pas de la subdivision, un théorème central de la théorie de l'intégration de Riemann est que les sommes de Riemann convergent vers l'intégrale de f sur $[a,b]$ lorsque le pas de la subdivision tend vers 0, voir le cours d'intégration de L2.

Théorème 3.1.1. Si $f \in \mathscr{C}([a,b])$ alors pour toute suite de subdivisions pointées $(\sigma_n)_{n\in\mathbb{N}}$, on a :

$$|\sigma_n| \to 0 \implies S(f,\sigma_n) \to \int_a^b f(x)\,\mathrm{d}x.$$

Exemple 3.1.2. Pour la subdivision régulière $a_k = a + kh$ avec $h = \frac{b-a}{N}$ le pas (constant), les sommes de Riemann se réécrivent :

$$S(f,\sigma) = \frac{b-a}{N}\sum_{k=0}^{N-1} f(\xi_k).$$

Les choix classiques pour les points ξ_k sont les suivants :

- **méthode des rectangles gauches** : $\xi_k = a_k$.
- **méthode des rectangles droits** : $\xi_k = a_{k+1}$.
- **méthode du point milieu** : $\xi_k = \frac{a_k + a_{k+1}}{2}$.

Ces trois méthodes sont illustrées par la figure 3.2. ◀

La question est de savoir à quelle vitesse les sommes de Riemann convergent vers l'intégrale de f. Dans le cas d'une subdivision régulière, on a le résultat suivant.

Proposition 3.1.3. Si $f \in \mathscr{C}^1([a,b])$, alors pour tout pointage $(\xi_k)_{0\le k\le N-1}$ d'une subdivision régulière, on a :

$$\left|\int_a^b f(x)\,\mathrm{d}x - \frac{b-a}{N}\sum_{k=0}^{N-1} f(\xi_k)\right| \le \|f'\|_\infty \frac{(b-a)^2}{N}.$$ ◀

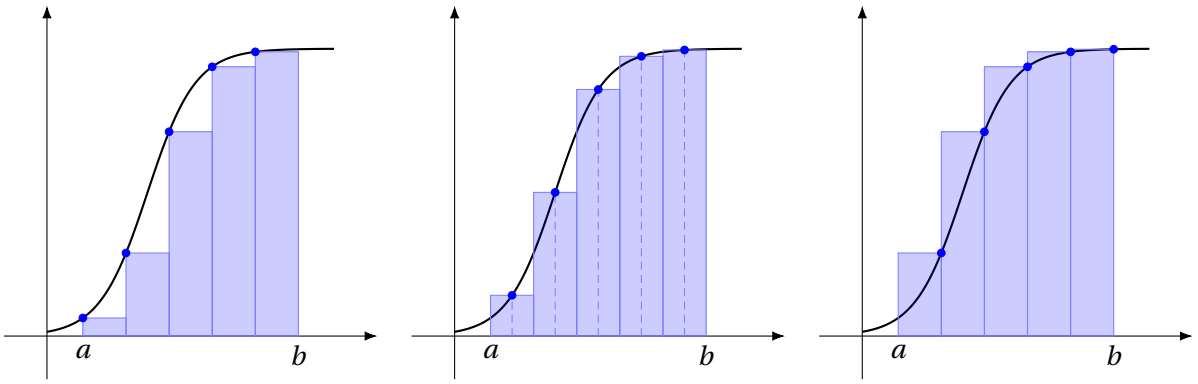


Figure 3.2 – **À gauche :** rectangles gauches. **Au centre :** point milieu. **À droite :** rectangles droits. La fonction étant croissante, la méthode des rectangles gauches sous-estime la valeur de l'intégrale, tandis que la méthode des rectangles droits la surestime.

Démonstration. Par la relation de Chasles et par inégalité triangulaire, on a :

$$\begin{aligned}\left|\int_a^b f(x)\,\mathrm{d}x - \frac{b-a}{N}\sum_{k=0}^{N-1} f(\xi_k)\right| &= \left|\sum_{k=0}^{N-1}\int_{a_k}^{a_{k+1}} f(x)\,\mathrm{d}x - \frac{b-a}{N}\sum_{k=0}^{N-1} f(\xi_k)\right| \\ &= \left|\sum_{k=0}^{N-1}\int_{a_k}^{a_{k+1}} \big(f(x)-f(\xi_k)\big)\,\mathrm{d}x\right| \\ &\leq \sum_{k=0}^{N-1}\int_{a_k}^{a_{k+1}} \big|f(x)-f(\xi_k)\big|\,\mathrm{d}x.\end{aligned}$$

On utilise l'inégalité des accroissements finis :

$$\forall x \in [a_k, a_{k+1}], \quad |f(x)-f(\xi_k)| \leq \|f'\|_\infty |x-\xi_k| \leq \|f'\|_\infty (a_{k+1}-a_k).$$

On obtient alors :

$$\begin{aligned}\left|\int_a^b f(x)\,\mathrm{d}x - \frac{b-a}{N}\sum_{k=0}^{N-1} f(a_k)\right| &\leq \sum_{k=0}^{N-1}\int_{a_k}^{a_{k+1}} \|f'\|_\infty (a_{k+1}-a_k)\,\mathrm{d}x \\ &= \|f'\|_\infty \sum_{k=0}^{N-1} (a_{k+1}-a_k)^2 \\ &= \|f'\|_\infty \sum_{k=0}^{N-1} h^2 \\ &= \|f'\|_\infty N h^2 \\ &= \|f'\|_\infty \frac{(b-a)^2}{N}.\end{aligned}$$

□

Ainsi, les méthodes des rectangles génériques ont une vitesse de convergence en $\mathrm{O}\left(\frac{1}{N}\right)$. Pour les méthodes des rectangles gauches/droits, on peut améliorer le majorant d'un facteur $\frac{1}{2}$ (voir en TD) :

$$\left|\int_a^b f(x)\,\mathrm{d}x - \frac{b-a}{N}\sum_{k=0}^{N-1} f(a_k)\right| \leq \|f'\|_\infty \frac{(b-a)^2}{2N}.$$

La vitesse de convergence reste la même.

Pour la méthode du point milieu, la majoration précédente peut être améliorée d'un facteur $\frac{1}{4}$. Mais on a en fait beaucoup mieux : on va montrer que la méthode du point milieu a une vitesse de convergence en $\mathrm{O}\left(\frac{1}{N^2}\right)$. Qualitativement, on peut le comprendre ainsi. La méthode des rectangles

gauches sous-estime l'intégrale sur les intervalles où f est croissante, et la surestime sur les intervalles où f est décroissante (et inversement pour les rectangles droits). En utilisant le point milieu, sur un intervalle $[a_k, a_{k+1}]$ où f est croissante, on surestime l'intégrale entre a_k et $\frac{a_k+a_{k+1}}{2}$ mais on la sous-estime entre $\frac{a_k+a_{k+1}}{2}$ et a_{k+1} (voir figure 3.2), de même quand f est décroissante. Il y a donc une compensation partielle des erreurs dans chaque intervalle où f est monotone. Ce phénomène de compensation des erreurs entraine une meilleure vitesse de convergence de la méthode du point milieu, comme le montre la proposition suivante.

Proposition 3.1.4. Si $f \in \mathscr{C}^2([a,b])$, alors :

$$\left| \int_a^b f(x)\,\mathrm{d}x - \frac{b-a}{N} \sum_{k=0}^{N-1} f\Big(\frac{a_k+a_{k+1}}{2}\Big) \right| \leq \|f''\|_\infty \frac{(b-a)^3}{24N^2}.$$ ◀

Démonstration. Notons $\xi_k = \frac{a_k+a_{k+1}}{2}$. Tout repose sur le fait que la fonction $x \mapsto (x-\xi_k)$ est d'intégrale nulle sur $[a_k, a_{k+1}]$. En effet, le changement de variable $x = \xi_k + t$ avec $t \in [-\frac{h}{2}, \frac{h}{2}]$ donne :

$$\int_{a_k}^{a_{k+1}} (x-\xi_k)\,\mathrm{d}x = \int_{-\frac{h}{2}}^{\frac{h}{2}} t\,\mathrm{d}t = 0. \tag{3.1}$$

On écrit comme dans la preuve précédente :

$$\begin{aligned}
\left| \int_a^b f(x)\,\mathrm{d}x - \frac{b-a}{N} \sum_{k=0}^{N-1} f(\xi_k) \right| &= \left| \sum_{k=0}^{N-1} \int_{a_k}^{a_{k+1}} \big(f(x) - f(\xi_k)\big)\,\mathrm{d}x \right| \\
&= \left| \sum_{k=0}^{N-1} \int_{a_k}^{a_{k+1}} \big(f(x) - f(\xi_k) - f'(\xi_k)(x-\xi_k)\big)\,\mathrm{d}x \right| \qquad \text{d'après (3.1)} \\
&\leq \sum_{k=0}^{N-1} \int_{a_k}^{a_{k+1}} \big|f(x) - f(\xi_k) - f'(\xi_k)(x-\xi_k)\big|\,\mathrm{d}x.
\end{aligned}$$

On utilise alors une généralisation de l'inégalité des accroissements finis. Par Taylor–Lagrange à l'ordre 2, pour tout $x \in [a_k, a_{k+1}]$, il existe θ compris entre ξ_k et x tel que

$$f(x) - f(\xi_k) = f'(\xi_k)(x-\xi_k) + \frac{f''(\theta)}{2}(x-\xi_k)^2.$$

Par conséquent :

$$\forall x \in [a_k, a_{k+1}], \quad \big|f(x) - f(\xi_k) - f'(\xi_k)(x-\xi_k)\big| \leq \frac{\|f''\|_\infty}{2}(x-\xi_k)^2.$$

Finalement :

$$\begin{aligned}
\left| \int_a^b f(x)\,\mathrm{d}x - \frac{b-a}{N} \sum_{k=0}^{N-1} f(\xi_k) \right| &\leq \sum_{k=0}^{N-1} \int_{a_k}^{a_{k+1}} \frac{\|f''\|_\infty}{2}(x-\xi_k)^2\,\mathrm{d}x \\
&= \frac{\|f''\|_\infty}{2} \sum_{k=0}^{N-1} \frac{(a_{k+1}-\xi_k)^3 - (a_k-\xi_k)^3}{3} \\
&= \frac{\|f''\|_\infty}{2} \sum_{k=0}^{N-1} \frac{\left(\frac{h}{2}\right)^3 - \left(-\frac{h}{2}\right)^3}{3} \\
&= \|f''\|_\infty \frac{Nh^3}{24} \\
&= \|f''\|_\infty \frac{(b-a)^3}{24N^2}.
\end{aligned}$$ □

3.2 Méthodes de quadrature simples

3.2.a Définition, erreur et ordre d'une méthode de quadrature

Que ce soit les méthodes des rectangles gauches/droits, la méthode du point milieu, ou une somme de Riemann quelconque, les méthodes d'intégration approchée consistent à évaluer f en des points de $[a,b]$, et à calculer une somme pondéré de ces valeurs. C'est ce type de méthode que nous allons étudier dans la suite.

Définition 3.2.1. On appelle **méthodes de quadrature** sur $[a,b]$ une application $J\colon \mathscr{C}([a,b]) \to \mathbb{R}$ de la forme :

$$\forall f \in \mathscr{C}([a,b]), \quad J(f) \coloneqq \sum_{i=0}^{n} \omega_i f(x_i),$$

avec :

- $(\omega_i)_{0\le i\le n}$ des scalaires appelés les **poids** ;
- $(x_i)_{0\le i\le n}$ des points distincts de $[a,b]$ appelés les **nœuds**.

On appelle **erreur** de la méthode de quadrature l'application $E\colon \mathscr{C}([a,b]) \to \mathbb{R}$ définie par :

$$\forall f \in \mathscr{C}([a,b]), \quad E(f) \coloneqq \int_a^b f(x)\,\mathrm{d}x - J(f).$$

Remarque 3.2.2. Il existe des méthodes de quadratures plus générales, faisant également intervenir les dérivées de f aux nœuds :

$$J(f) \coloneqq \sum_{i=0}^{n} \sum_{j=0}^{m_i} \omega_{i,j} f^{(j)}(x_i).$$

Ces méthodes ne seront pas étudiées dans ce cours. ◂

Remarque 3.2.3. On remarquera que J et E sont des applications linéaires. De plus, on prendra garde au fait que l'erreur $E(f)$ est signée. ◂

Pour une méthode de quadrature J, on cherchera à évaluer l'ordre de grandeur de l'erreur $|E(f)|$. Un autre critère pour évaluer la qualité de la méthode, plus algébrique, est le suivant.

Définition 3.2.4. On dit qu'une méthode de quadrature est **d'ordre** p si elle est exacte pour les polynômes de degré inférieur ou égal à p :

$$\forall P \in \mathbb{R}_p[X], \quad J(P) = \int_a^b P(x)\,\mathrm{d}x.$$

On dit que la méthode est exactement d'ordre p si elle est d'ordre p mais pas d'ordre $p+1$.

Exemple 3.2.5. La méthode des rectangles gauches/droits est d'ordre 0. La méthode du point milieu est d'ordre 1. ◂

Si on s'inspire des sommes de Riemann, une façon de construire une méthode de quadrature est de subdiviser l'intervalle $[a,b]$ en N morceaux, et d'approcher l'intégrale sur chacun d'eux avec une méthode simple (l'aire d'un rectangle en ce qui concerne les sommes de Riemann).
Soit $a = a_0 < a_1 < \cdots < a_N = b$ une subdivision de $[a,b]$. Par la relation de Chasles, on écrit :

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{k=0}^{N-1} \int_{a_k}^{a_{k+1}} f(x)\,\mathrm{d}x.$$

On approche l'intégrale de f sur chaque intervalle $[a_k, a_{k+1}]$ avec une méthode de quadrature J_k, dite **simple** (ou élémentaire) :

$$\int_{a_k}^{a_{k+1}} f(x)\,\mathrm{d}x \approx J_k(f).$$

On obtient ainsi une méthode de quadrature dite **composite** :

$$J(f) \coloneqq \sum_{k=0}^{N-1} J_k(f).$$

En notant E_k l'erreur de la méthode de quadrature J_k sur $[a_k, a_{k+1}]$, l'erreur de la méthode composite s'écrit simplement grâce à la relation de Chasles :

$$\begin{aligned} E(f) &= \int_a^b f(x)\,\mathrm{d}x - \sum_{k=0}^{N-1} J_k(f) \\ &= \sum_{k=0}^{N-1} \left(\int_{a_k}^{a_{k+1}} f(x)\,\mathrm{d}x - J_k(f) \right) \\ &= \sum_{k=0}^{N-1} E_k(f). \end{aligned}$$

Exemple 3.2.6. En prenant $J_k(f) \coloneqq (a_{k+1} - a_k) f(\xi_k)$ avec $\xi_k \in [a_k, a_{k+1}]$, on retrouve les sommes de Riemann. ◀

Proposition 3.2.7. L'erreur d'une méthode composite est la somme des erreurs des méthodes de quadrature simples la définissant. ◀

3.2.b Méthodes de quadrature interpolatoires

Une façon de construire une méthode de quadrature d'ordre n est d'approcher l'intégrale de f par celle d'un polynôme interpolateur P_n en $n+1$ nœuds $(x_i)_{0\le i\le n}$:

$$\int_a^b f(x)\,\mathrm{d}x \approx \int_a^b P_n(x)\,\mathrm{d}x. \tag{3.2}$$

Ceci définit bien une méthode de quadrature au sens de la définition 3.2.1. En effet, si on note $(L_i)_{0\le i\le n}$ les polynômes interpolateurs de Lagrange aux nœuds $(x_i)_{0\le i\le n}$, on a par linéarité de l'intégrale :

$$\int_a^b P_n(x)\,\mathrm{d}x = \int_a^b \sum_{i=0}^{n} f(x_i) L_i(x)\,\mathrm{d}x = \sum_{i=0}^{n} f(x_i) \int_a^b L_i(x)\,\mathrm{d}x = \sum_{i=0}^{n} \omega_i f(x_i),$$

les poids étant donnés par $\omega_i \coloneqq \int_a^b L_i(x)\,\mathrm{d}x$.

Définition 3.2.8. On appelle **méthode de quadrature interpolatoire** associée aux nœuds $(x_i)_{0\le i\le n}$ la méthode :

$$J(f) \coloneqq \sum_{i=0}^{n} \omega_i f(x_i), \qquad \omega_i \coloneqq \int_a^b L_i(x)\,\mathrm{d}x,$$

où $(L_i)_{0\le i\le n}$ sont les polynômes interpolateurs de Lagrange aux nœuds $(x_i)_{0\le i\le n}$.

Proposition 3.2.9. Toute méthode de quadrature interpolatoire à $n+1$ nœuds est d'ordre n. ◀

Démonstration. Si $f \in \mathbb{R}_n[X]$, alors f est égale à son polynôme interpolateur, donc (3.2) est une égalité. □

Exemple 3.2.10 (Méthodes à un nœud)**.** Avec un seul nœud x_0, le poids correspondant est $\omega_0 = b-a$. La méthode s'écrit donc :

$$J(f, [a,b]) = (b-a) f(x_0).$$

- Pour $x_0 = a$, c'est la méthode simple des rectangles gauches.
- Pour $x_0 = b$, c'est la méthode simple des rectangles droits.

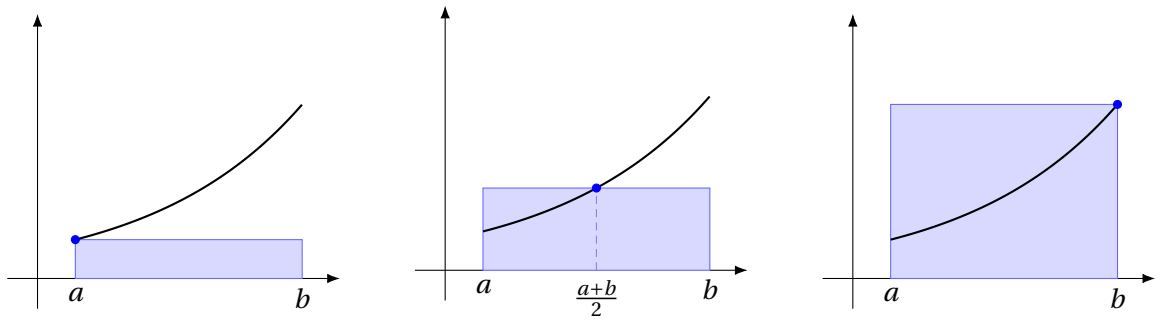


Figure 3.3 – **À gauche :** rectangles gauches. **Au centre :** point milieu. **À droite :** rectangles droits.

- Pour $x_0 = \frac{a+b}{2}$ c'est la méthode simple du point milieu. ◀

Exemple 3.2.11 (Méthode des trapèzes)**.** Avec deux nœuds $x_0 = a$ et $x_1 = b$, le polynôme interpolateur de f en a et b est la sécante :

$$P_1(x) = f(a) + \frac{f(b)-f(a)}{b-a}(x-a).$$

La **méthode simple des trapèzes** s'écrit donc :

$$J_{\text{trapèze}}(f,[a,b]) = \int_a^b P_1(x)\,\mathrm{d}x = f(a)(b-a) + \frac{f(b)-f(a)}{b-a}\frac{(b-a)^2}{2} = (b-a)\frac{f(a)+f(b)}{2}.$$

Les poids sont donc $\omega_0 = \omega_1 = \frac{b-a}{2}$. Géométriquement, on a approché l'intégrale de f sur $[a,b]$ par l'aire du trapèze de sommets $(a,0)$, $(a,f(a))$, $(b,f(b))$ et $(b,0)$. ◀

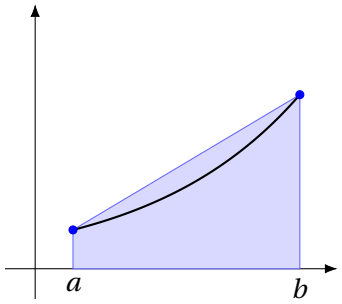


Figure 3.4 – Méthode simple des trapèzes. L'aire sous la courbe est approchée par l'aire du trapèze en bleu.

Exemple 3.2.12 (Méthode de Simpson)**.** La **méthode simple de Simpson** sur $[a,b]$ est la méthode de quadrature interpolatoire à 3 nœuds $x_0 = a$, $x_1 = \frac{a+b}{2}$ et $x_2 = b$. Pour calculer les poids, utilisons le changement de variable affine $x = \frac{a+b}{2} + \frac{b-a}{2}t$ pour se ramener à l'intervalle $[-1,1]$:

$$\omega_i = \int_a^b L_i(x)\,\mathrm{d}x = \frac{b-a}{2}\int_{-1}^1 \tilde{L}_i(t)\,\mathrm{d}t,$$

où les $\tilde{L}_i$ sont les polynômes interpolateurs de Lagrange en $t_0 = -1$, $t_1 = 0$ et $t_2 = 1$:

$$\tilde{L}_0(t) = \frac{(t-0)(t-1)}{(-1-0)(-1-1)} = \frac{t^2-t}{2},$$
$$\tilde{L}_1(t) = \frac{(t+1)(t-1)}{(0+1)(0-1)} = -t^2+1,$$
$$\tilde{L}_2(t) = \frac{(t+1)(t-0)}{(1+1)(1-0)} = \frac{t^2+t}{2}.$$

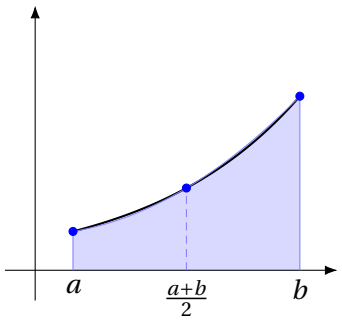


Figure 3.5 – Méthode simple de Simpson. L'aire sous la courbe est approchée par l'aire sous la parabole qui interpole la fonction aux nœuds a, $\frac{a+b}{2}$ et b (zoomer sur la figure pour voir l'écart en la fonction et la parabole).

Ainsi, les poids sur $[-1,1]$ sont :

$$\tilde{\omega}_0 = \int_{-1}^{1} \tilde{L}_0(t)\,\mathrm{d}t = \frac{1}{3}, \qquad \tilde{\omega}_1 = \int_{-1}^{1} \tilde{L}_1(t)\,\mathrm{d}t = \frac{4}{3}, \qquad \tilde{\omega}_2 = \int_{-1}^{1} \tilde{L}_2(t)\,\mathrm{d}t = \frac{1}{3}.$$

Les poids sur $[a,b]$ sont alors $\omega_i = \frac{b-a}{2}\tilde{\omega}_i$ et la méthode de Simpson est donnée par :

$$J_{\text{Simpson}}(f,[a,b]) := \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right).$$

Cette méthode est en fait d'ordre 3 (et pas juste d'ordre 2). En effet, elle est exacte sur $\mathbb{R}_2[X]$ par la proposition 3.2.9, et pour $f(x) = x^3$ sur $[-1,1]$, on a :

$$J_{\text{Simpson}}(f,[-1,1]) = \frac{1}{3}f(-1) + \frac{4}{3}f(0) + \frac{1}{3}f(1) = 0 = \int_{-1}^{1} x^3\,\mathrm{d}x,$$

donc par linéarité, la méthode sur $[-1,1]$ est exacte pour tout $f \in \mathbb{R}_3[X]$. Par changement de variable affine, on en déduit que la méthode simple de Simpson est d'ordre 3 sur tout intervalle. ◀

Les méthodes des trapèzes et de Simpson sont les méthodes interpolatoires sur 2 ou 3 nœuds équidistants de $[a,b]$. On peut généraliser en considérant des méthodes interpolatoires sur $n+1$ nœuds équidistants.

Définition 3.2.13. On appelle **méthode simple de Newton–Cotes** de degré n la méthode de quadrature interpolatoire sur les nœuds équidistants

$$\forall i \in [\![0,n]\!], \quad x_i := a + ih,$$

avec $h = \frac{b-a}{n}$ le pas.

Proposition 3.2.14. Les poids de la méthode simple de Newton–Cotes d'ordre n sur $[a,b]$ sont donnés par :

$$\forall i \in [\![0,n]\!], \quad \omega_i = h\int_0^n \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{t-j}{i-j}\,\mathrm{d}t.$$

En particulier, les poids sont symétriques : $\omega_i = \omega_{n-i}$ pour tout $i \in [\![0,n]\!]$. ◀

Démonstration. On utilise le changement de variable affine $x = a + th$, avec $t \in [0,n]$:

$$\frac{x-x_j}{x_i-x_j} = \frac{a+th-a-jh}{a+ih-a-jh} = \frac{(t-j)h}{(i-j)h} = \frac{t-j}{i-j}.$$

n	Nom	Coefficients $(c_0,\dots,c_n)$	Ordre
1	Trapèzes	$\left(\frac{1}{2},\frac{1}{2}\right)$	1
2	Simpson	$\left(\frac{1}{3},\frac{4}{3},\frac{1}{3}\right)$	3
4	Boole–Villarceau	$\frac{2}{45}(7,32,12,32,7)$	5
6	Weddle–Hardy	$\frac{1}{140}(41,216,27,272,27,216,41)$	7

Table 3.1 – Coefficients de la méthode de Newton–Cotes. En Python, la fonction `newton_cotes` du module `scipy.integrate` permet de calculer les coefficients pour n'importe quel n.

Avec ce changement de variable, les poids s'écrivent :

$$\omega_i = \int_a^b L_i(x)\,\mathrm{d}x = \int_a^b \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j}\,\mathrm{d}x = h\int_0^n \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-j}{i-j}\,\mathrm{d}t. \qquad \square$$

Les poids de la méthode de Newton–Cotes ne dépendent que de h et de :

$$\forall i \in [\![0,n]\!],\quad c_i := \int_0^n \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-j}{i-j}\,\mathrm{d}t = \frac{(-1)^{n-i}}{i!(n-i)!}\int_0^n \prod_{\substack{j=0\\ j\neq i}}^{n} (t-j)\,\mathrm{d}t. \tag{3.3}$$

Les coefficients $(c_i)_{0\le i\le n}$ ne dépendent que de n et pas de l'intervalle $[a,b]$, ils peuvent donc être calculés à l'avance une bonne fois pour toutes, voir table 3.1. La méthode de Newton–Cotes s'écrit donc :

$$\mathrm{NC}_n(f,[a,b]) = h\sum_{i=0}^{n} c_i f(a+ih),\quad \text{avec } h=\frac{b-a}{n}.$$

Proposition 3.2.15. L'ordre de la méthode de Newton–Cotes de degré n est :

$$\begin{cases} n & \text{si } n \text{ est impair,}\\ n+1 & \text{si } n \text{ est pair.}\end{cases}$$

◀

Démonstration. On sait que déjà que NC_n est d'ordre n pour tout $n\in\mathbb{N}$ (proposition 3.2.9). La seule chose à montrer est que si n est pair, alors NC_n est aussi d'ordre $n+1$. Montrons que NC_n est exacte pour $f(x)=x^{n+1}$ sur $[-1,1]$. Puisque $n+1$ est impair :

$$\int_{-1}^{1} x^{n+1}\,\mathrm{d}x = 0.$$

Montrons qu'il en est de même pour NC_n. On utilise pour ça les propriétés de symétrie des poids et des nœuds : $\forall i\in[\![0,n]\!]$, $\omega_{n-i}=\omega_i$ et $x_{n-i}=-x_i$. On a alors :

$$\begin{aligned}
\mathrm{NC}_n(f,[-1,1]) &= \sum_{i=0}^{n}\omega_i x_i^{n+1} && \\
&= \sum_{i=0}^{n}\omega_{n-i}x_{n-i}^{n+1} && \text{changement d'indice } i\leftarrow n-i\\
&= \sum_{i=0}^{n}\omega_i(-x_i)^{n+1} && \text{symétrie}\\
&= \sum_{i=0}^{n}\omega_i(-1)x_i^{n+1} && n+1 \text{ est impair}\\
&= -\mathrm{NC}_n(f,[-1,1]). &&
\end{aligned}$$

Par conséquent $\mathrm{NC}(f,[-1,1])=0$. Par linéarité, on en déduit que NC_n est exacte sur $[-1,1]$ pour tout $f\in\mathbb{R}_{n+1}[X]$. Par changement de variables affine, ça reste vrai sur tout intervalle $[a,b]$. $\square$

En dehors de $n = 1$ (méthode des trapèzes), on n'utilise que les méthodes de Newton–Cotes de degré pair, puisque les méthodes de degré impair ne font pas gagner d'ordre.

3.3 Stabilité des méthodes de quadrature

On considère une méthode de quadrature :

$$J(f) = \sum_{i=0}^{n} \omega_i f(x_i).$$

On s'interroge sur la **stabilité** d'une telle méthode : si les valeurs $f(x_i)$ sont calculées avec une erreur ε, quelle erreur commet-on sur $J(f)$?

Proposition 3.3.1. Pour tout $f \in \mathscr{C}([a,b])$, on a :

$$|J(f)| \leq \|f\|_\infty \sum_{i=0}^{n} |\omega_i|.$$

De plus, si les poids sont positifs et que la méthode est d'ordre 0, alors :

$$|J(f)| \leq (b-a)\|f\|_\infty. \quad \blacktriangleleft$$

Démonstration. Soit f continue sur $[a,b]$. Par inégalité triangulaire :

$$|J(f)| = \left| \sum_{i=0}^{n} \omega_i f(x_i) \right| \leq \sum_{i=0}^{n} |\omega_i||f(x_i)| \leq \|f\|_\infty \sum_{i=0}^{n} |\omega_i|.$$

Si les ω_i sont positifs et que la méthode est d'ordre 0 (c'est-à-dire exacte pour les constantes), alors :

$$\sum_{i=0}^{n} |\omega_i| = \sum_{i=0}^{n} \omega_i = \int_a^b 1 \,\mathrm{d}x = b-a. \quad \square$$

Remarque 3.3.2. Si $\tilde{f}(x)$ est la valeur approchée de $f(x)$ qu'on calcule en pratique, avec une erreur $\|\tilde{f} - f\|_\infty \leq \varepsilon$, alors :

$$|J(\tilde{f}) - J(f)| = |J(\tilde{f} - f)| \leq \|\tilde{f} - f\|_\infty \sum_{i=0}^{n} |\omega_i| \leq \varepsilon \sum_{i=0}^{n} |\omega_i|.$$

Quand les poids sont positifs, on a donc $|J(\tilde{f}) - J(f)| \leq (b-a)\varepsilon$, et la méthode est stable. En revanche, pour des poids non tous positifs, l'erreur peut être amplifiée d'un facteur :

$$\sum_{i=0}^{n} |\omega_i|.$$

Cette quantité peut ne pas être bornée quand $n \to +\infty$, et la méthode est alors instable. D'ailleurs, pour une méthode interpolatoire, on peut majorer :

$$\sum_{i=0}^{n} |\omega_i| = \sum_{i=0}^{n} \left| \int_a^b L_i(x) \,\mathrm{d}x \right| \leq \int_a^b \sum_{i=0}^{n} |L_i(x)| \,\mathrm{d}x \leq (b-a)\Lambda_n,$$

où Λ_n est la constante de Lebesgue (voir définition 2.4.1). Pour des nœuds équidistants, on sait que cette constante diverge vers $+\infty$ rapidement. On évitera donc d'utiliser des méthodes de quadrature dont les poids ne sont pas tous positifs. ◀

Exemple 3.3.3. Pour les méthodes de Newton–Cotes, les poids sont positifs pour tout $n \leq 7$ et $n = 9$. En revanche, pour $n = 8$ et $n \geq 10$, la méthode comporte des poids négatifs, et doit être considérée comme instable. On voit ici la traduction du phénomène de Runge pour l'intégration numérique : quand les nœuds sont équidistants, l'interpolation est instable et peut ne pas converger vers la fonction, même quand celle-ci est régulière. En pratique, on se limitera aux méthodes de Newton–Cotes de la table 3.1. ◀

3.4 Expression de l'erreur

On considère une méthode de quadrature :

$$J(f) = \sum_{i=0}^{n} \omega_i f(x_i).$$

On rappelle que l'erreur est définie par :

$$E(f) = \int_a^b f(x)\,\mathrm{d}x - J(f).$$

On suppose que la méthode est d'ordre p, avec $p \geq n$.

Définition 3.4.1. On appelle **noyau de Peano** d'ordre p de la méthode, la fonction :

$$\forall t \in [a,b], \quad K_p(t) := E(x \mapsto (x-t)_+^p) = \int_a^b (x-t)_+^p\,\mathrm{d}x - \sum_{i=0}^{n} \omega_i (x_i - t)_+^p,$$

où $(x)_+ := \max(x, 0)$ est la partie positive de x.

Théorème 3.4.2. On suppose que J est une méthode de quadrature d'ordre p, avec $p \geq n$. Si f est de classe $\mathscr{C}^{p+1}$, alors :

$$E(f) = \frac{1}{p!} \int_a^b f^{(p+1)}(t)\, K_p(t)\,\mathrm{d}t.$$

Démonstration. Écrivons la formule de Taylor avec reste intégral :

$$f(x) = \sum_{k=0}^{p} \frac{f^{(k)}(a)}{k!}(x-a)^k + R_p(x),$$

avec :

$$R_p(x) = \int_a^x \frac{f^{(p+1)}(t)}{p!}(x-t)^p\,\mathrm{d}t = \int_a^b \frac{f^{(p+1)}(t)}{p!}(x-t)_+^p\,\mathrm{d}t.$$

Si on note P le polynôme de Taylor de f en a à l'ordre p, alors par linéarité $E(f) = E(P) + E(R_p)$. Or $E(P) = 0$ car la méthode est d'ordre p. Par conséquent :

$$E(f) = E(R_p) = \int_a^b \left(\int_a^b \frac{f^{(p+1)}(t)}{p!}(x-t)_+^p\,\mathrm{d}t \right) \mathrm{d}x - \sum_{i=0}^{n} \omega_i \int_a^b \frac{f^{(p+1)}(t)}{p!}(x_i - t)_+^p\,\mathrm{d}t.$$

La fonction $(x,t) \mapsto \frac{f^{(p+1)}(t)}{p!}(x-t)_+^p$ est continue sur $[a,b] \times [a,b]$, donc on peut échanger l'ordre des intégrales par le théorème de Fubini. On échange aussi la somme et l'intégrale du 2e terme par linéarité, et on obtient :

$$\begin{aligned} E(f) &= \int_a^b \left(\int_a^b \frac{f^{(p+1)}(t)}{p!}(x-t)_+^p\,\mathrm{d}x \right) \mathrm{d}t - \int_a^b \sum_{i=0}^{n} \omega_i \frac{f^{(p+1)}(t)}{p!}(x_i - t)_+^p\,\mathrm{d}t \\ &= \int_a^b \frac{f^{(p+1)}(t)}{p!} \left(\int_a^b (x-t)_+^p\,\mathrm{d}x - \sum_{i=0}^{n} \omega_i (x_i - t)_+^p \right) \mathrm{d}t \\ &= \int_a^b \frac{f^{(p+1)}(t)}{p!} K_p(t)\,\mathrm{d}t. \end{aligned}$$

□

En majorant $|f^{(p+1)}(t)|$ par $\|f^{(p+1)}\|_\infty$ sur $[a,b]$, on obtient la majoration suivante de l'erreur.

Corollaire 3.4.3. Sous les hypothèses du théorème 3.4.2 :

$$\forall f \in \mathscr{C}^{p+1}([a,b]), \quad |E(f)| \leq \frac{1}{p!} \|f^{(p+1)}\|_\infty \int_a^b |K_p(t)|\,\mathrm{d}t.$$ ◀

Corollaire 3.4.4. Sous les hypothèses du théorème 3.4.2, si de plus K_p est de signe constant, alors il existe $\xi \in]a,b[$ tel que :

$$E(f) = \frac{f^{(p+1)}(\xi)}{p!} \int_a^b K_p(t)\,\mathrm{d}t.$$ ◀

Démonstration. Puisque K_p est de signe constant, on peut utiliser la formule de la moyenne généralisée :

$$\exists \xi \in]a,b[, \quad \int_a^b \frac{f^{(p+1)}(t)}{p!} K_p(t)\,\mathrm{d}t = \frac{f^{(p+1)}(\xi)}{p!} \int_a^b K_p(t)\,\mathrm{d}t.$$ □

Exemple 3.4.5 (Méthode des trapèzes)**.** La méthode des trapèzes est d'ordre 1, donc :

$$\forall f \in \mathscr{C}^2([a,b]), \quad E(f) = \int_a^b f''(t) K_1(t)\,\mathrm{d}t.$$

Calculons le noyau de Peano :

$$\begin{aligned} K_1(t) &= \int_a^b (x-t)_+\,\mathrm{d}x - \frac{b-a}{2}\big((a-t)_+ + (b-t)_+\big) \\ &= \frac{(b-t)^2}{2} - \frac{b-a}{2}(b-t) && \text{car } a \le t \le b \\ &= \frac{1}{2}(t-a)(t-b). \end{aligned}$$

Cette fonction est négative pour tout $t \in [a,b]$, donc le corollaire 3.4.4 s'applique. L'intégrale de K_1 vaut $-\frac{(b-a)^3}{12}$, d'où

$$\exists \xi \in]a,b[, \quad E(f) = -f''(\xi)\frac{(b-a)^3}{12}.$$

Par conséquent :

$$\forall f \in \mathscr{C}^2([a,b]), \quad |E(f)| \le \|f''\|_\infty \frac{(b-a)^3}{12}, \tag{3.4}$$

et l'erreur est négative si f est convexe, et positive si f est concave. ◀

Exemple 3.4.6 (Méthode de Simpson)**.** La méthode de Simpson est d'ordre 3, donc :

$$\forall f \in \mathscr{C}^4([a,b]), \quad E(f) = \int_a^b \frac{f^{(4)}(t)}{6} K_3(t)\,\mathrm{d}t.$$

On peut montrer par un calcul fastidieux que :

$$K_3(t) = \begin{cases} \dfrac{1}{12}(b-t)^3(2b+a-3t) & \text{si } t \le \frac{a+b}{2}, \\ \dfrac{1}{12}(a-t)^3(2a+b-3t) & \text{si } t \ge \frac{a+b}{2}. \end{cases}$$

Cette fonction est négative et son intégrale vaut $-\frac{(b-a)^5}{480}$, donc :

$$\exists \xi \in]a,b[, \quad E(f) = -f^{(4)}(\xi)\frac{(b-a)^5}{2880}.$$

Par conséquent :

$$\forall f \in \mathscr{C}^4([a,b]), \quad |E(f)| \le \|f^{(4)}\|_\infty \frac{(b-a)^5}{2880}.$$ ◀

On mentionne le théorème de Steffensen, qu'on admet. Une conséquence de ce théorème est que le corollaire 3.4.4 est toujours applicable pour les méthodes de Newton–Cotes.

Théorème 3.4.7 (Steffensen, admis)**.** Le noyau de Peano des méthodes de Newton–Cotes est toujours de signe constant.

3.5 Méthode de quadrature composite

Le problème des méthodes de quadrature interpolatoires précédentes est que l'approximation peut être assez mauvaise si l'intervalle $[a,b]$ est trop grand. Pour y remédier, on utilise une méthode composite : on considère $a = a_0 < a_1 < \cdots < a_N = b$ une subdivision de $[a,b]$ et sur chaque intervalle $[a_k, a_{k+1}]$ on utilise une méthode de quadrature interpolatoire simple. En toute généralité, on peut utiliser des méthodes différentes pour chaque intervalle $[a_k, a_{k+1}]$. Le plus souvent, on utilise la même méthode pour chaque intervalle.

Considérons le cas où la subdivision $(a_k)_{0\le k\le N}$ est régulière, c'est-à-dire :

$$\forall k \in [\![0,n]\!], \quad a_k = a + kh,$$

où $h = \frac{b-a}{N}$ est le pas de la subdivision. Sur chaque intervalle $[a_k, a_{k+1}]$, on utilise la méthode de Newton–Cotes de degré n :

$$\int_{a_k}^{a_{k+1}} f(x)\,\mathrm{d}x \approx \frac{a_{k+1}-a_k}{n} \sum_{i=0}^{n} c_i f(x_{i,k}),$$

où $x_{i,k} := a_k + i\frac{a_{k+1}-a_k}{n}$ et $(c_i)_{0\le i\le n}$ sont les coefficients définis par (3.3). En réécrivant $a_k = a+kh$ et $a_{k+1} - a_k = h$, on obtient :

$$\frac{a_{k+1}-a_k}{n} \sum_{i=0}^{n} c_i f(x_{i,k}) = \frac{h}{n} \sum_{i=0}^{n} c_i f\left(a_{k+\frac{i}{n}}\right),$$

où $a_{k+\frac{i}{n}} := a + \left(k + \frac{i}{n}\right)h$.

Définition 3.5.1. La **méthode composite de Newton–Cotes** de degré n avec N intervalles est la méthode de quadrature :

$$\mathrm{NC}_{n,N}(f,[a,b]) = \frac{h}{n} \sum_{k=0}^{N-1} \sum_{i=0}^{n} c_i f\left(a_{k+\frac{i}{n}}\right), \qquad a_{k+\frac{i}{n}} := a + \left(k + \frac{i}{n}\right)h,$$

où les coefficients $(c_i)_{0\le i\le n}$ sont définis par (3.3).

Exemple 3.5.2 (Méthode des trapèzes)**.** La **méthode composite des trapèzes** est définie par :

$$T_N(f,[a,b]) = h \sum_{k=0}^{N-1} \frac{f(a_k)+f(a_{k+1})}{2} = h \times \left(\frac{f(a)+f(b)}{2} + \sum_{k=1}^{N-1} f(a_k) \right).$$

D'après (3.4), l'erreur de la méthode simple des trapèzes est majorée par :

$$|E(f,[a_k,a_{k+1}])| \le \sup_{x\in[a_k,a_{k+1}]} |f''(x)| \frac{(a_{k+1}-a_k)^3}{12} \le \sup_{x\in[a,b]} |f''(x)| \frac{h^3}{12}.$$

L'erreur de la méthode composite est donc majorée par :

$$\begin{aligned} |E_N(f,[a,b])| &\le \sum_{k=0}^{N-1} |E(f,[a_k,a_{k+1}])| \\ &\le N \times \|f''\|_\infty \frac{h^3}{12} \\ &\le \|f''\|_\infty \frac{(b-a)^3}{12N^2}. \end{aligned}$$

◀

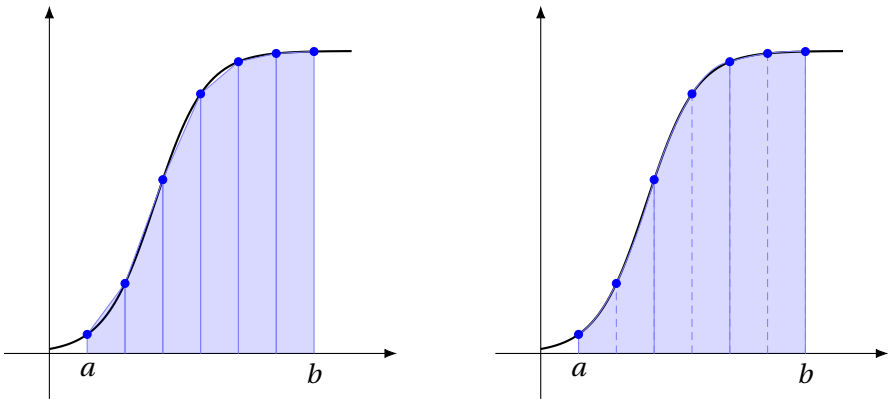


Figure 3.6 – **À gauche :** méthode des trapèzes ($N = 6$). **À droite :** méthode de Simpson ($N = 3$). Les deux méthodes évaluent la fonction 7 fois, aux mêmes points.

Exemple 3.5.3 (Méthode de Simpson)**.** La **méthode composite de Simpson** est la méthode définie par :

$$\begin{aligned} S_N(f,[a,b]) &= \frac{h}{6}\sum_{k=0}^{N-1}\left(f(a_k)+4f\left(a_{k+\frac{1}{2}}\right)+f(a_{k+1})\right) \\ &= \frac{h}{3}\left(\frac{f(a)+f(b)}{2}+\sum_{k=1}^{N-1}f(a_k)+2\sum_{k=0}^{N-1}f\left(a_{k+\frac{1}{2}}\right)\right), \end{aligned}$$

où $a_{k+\frac{1}{2}} = \frac{a_k+a_{k+1}}{2} = a+\left(k+\frac{1}{2}\right)h$. Le même raisonnement que pour la méthode des trapèzes montre que l'erreur de la méthode de Simpson composite est majorée par :

$$|E_N(f,[a,b])| \leq \|f^{(4)}\|_\infty \frac{(b-a)^5}{2880N^4}$$ ◂.

Remarque 3.5.4. Si on veut comparer les erreurs de la méthode des trapèzes et de Simpson, il faut le faire pour un nombre de nœuds équivalents. En effet, pour un même nombre N d'intervalles, la méthode des trapèzes évalue f en $N+1$ points, alors que la méthode de Simpson évalue f en $2N+1$ points (puisqu'on évalue aussi au milieu de chaque intervalle). Pour comparer les méthodes sur une base commune, on définit $h_{\text{eff}} = \frac{b-a}{nN}$ le pas effectif, et on exprime l'erreur en fonction de celui-ci. On a $h_{\text{eff}} = h$ pour les trapèzes et $h_{\text{eff}} = \frac{h}{2}$ pour Simpson.

Méthode	Erreur	Erreur (pas effecitf)
Trapèzes	$\|f''\|_\infty \frac{(b-a)^3}{12N^2}$	$h_{\text{eff}}^2 \|f''\|_\infty \frac{b-a}{12}$
Simpson	$\|f^{(4)}\|_\infty \frac{(b-a)^5}{2880N^4}$	$h_{\text{eff}}^4 \|f^{(4)}\|_\infty \frac{b-a}{180}$

À pas effectif équivalent, l'erreur de la méthode des trapèzes est en $O(h_{\text{eff}}^2)$ alors que celle de la méthode de Simpson est en $O(h_{\text{eff}}^4)$. Quand $h_{\text{eff}} \to 0$, l'erreur de la méthode de Simpson est négligeable devant l'erreur de la méthode des trapèzes. ◂

Remarque 3.5.5. Pour améliorer la précision d'une méthode composite, on a deux possibilités :

- augmenter l'ordre de la méthode ;
- augmenter le nombre N d'intervalles.

Dans le cas de Newton–Cotes, la première option n'est pas recommandée : l'interpolation devient instable et peut diverger. Il est donc préférable d'utiliser une méthode d'ordre faible (mais stable), et d'augmenter N si on veut gagner en précision. En pratique, la méthode de Simpson composite donne en général un bon compromis entre le nombre d'évaluations de la fonction et l'erreur sur l'intégrale. ◂

3.6 Méthodes de quadrature de Gauss

La méthode de Newton–Cotes étudiée précédemment se base sur l'interpolation de f en des nœuds équidistants. Or, on a vu au chapitre 2 que l'interpolation sur des nœuds équidistants n'était pas le meilleur choix pour approcher f, en plus d'être instable. On va donc chercher des nœuds pour lesquels la méthode de quadrature interpolatoire est d'ordre le plus élevé.

Généralisons un peu le problème du calcul d'intégrale. On considère I un intervalle (non nécessairement fermé ou borné) et $w\colon I \to \mathbb{R}_+$ une fonction poids au sens de la définition 2.6.1. Pour $f \in \mathscr{C}(I,w)$, on cherche à approcher numériquement l'intégrale :

$$\int_I f(x)\,w(x)\,\mathrm{d}x,$$

à l'aide d'une méthode de quadrature :

$$J(f) = \sum_{i=0}^{n} \omega_i f(x_i).$$

Définition 3.6.1. On dit que J est **d'ordre** p, pour le poids w, si J est exacte pour tout polynôme de $\mathbb{R}_p[X]$:

$$\forall P \in \mathbb{R}_p[X], \quad \int_I P(x)\,w(x)\,\mathrm{d}x = \sum_{i=0}^{n} \omega_i P(x_i).$$

Remarque 3.6.2. Attention, la notion d'ordre d'une méthode dépend du poids w. Par exemple, la méthode du point milieu est d'ordre 1 pour $w(x) = 1$ sur $[a,b]$, mais elle n'est en général que d'ordre 0 pour un poids w quelconque. ◀

Notons $(P_n)_{n\in\mathbb{N}}$ la suite des polynômes orthogonaux unitaires pour le poids w. On va voir que le meilleur choix pour les nœuds est d'utiliser les racines des polynômes orthogonaux. Mais avant, prouvons que P_n possède n racines réelles distinctes et que celles-ci appartiennent à I.

Proposition 3.6.3. Pour tout $n \in \mathbb{N}^*$, P_n possède n racines simples dans $\mathring{I}$ (l'intérieur de I). ◀

Démonstration. Soient $x_1 \ldots, x_p$ les racines de P_n appartenant à $\mathring{I}$ et notons $m_1, \ldots, m_p$ leurs multiplicités. Alors P_n s'écrit[3] :

$$P_n(X) = \prod_{i=1}^{p} (X - x_i)^{m_i} \times Q(X)$$

avec $Q \in \mathbb{R}[X]$ un polynôme qui ne s'annule pas dans $\mathring{I}$. Par contraposée du théorème des valeurs intermédiaires, Q est de signe constant sur $\mathring{I}$. Considérons le polynôme :

$$R(X) = \prod_{i=1}^{p} (X - x_i)^{\varepsilon_i}, \quad \text{avec } \varepsilon_i = \begin{cases} 1 & \text{si } m_i \text{ est impair,} \\ 0 & \text{si } m_i \text{ est pair.} \end{cases}$$

Alors $P_n(X)R(X) = \prod_{i=1}^{p} (X - x_i)^{m_i+\varepsilon_i} \times Q(X)$ avec $m_i + \varepsilon_i$ pair pour tout $i \in [\![1,p]\!]$, donc $P_n \times R$ est de signe constant sur $\mathring{I} \setminus \{x_1 \ldots, x_p\}$. Par conséquent :

$$\langle P_n, R \rangle_w = \int_I P_n(x)R(x)w(x)\,\mathrm{d}x \neq 0.$$

Or, P_n est orthogonal à tout polynôme de $\mathbb{R}_{n-1}[X]$, donc on doit avoir $\deg(R) \geq n$. Ceci n'est possible que si $p = n$ et $m_i = 1$ pour tout $i \in [\![1,n]\!]$, c'est-à-dire si P_n est à racines simples et toutes ses racines appartiennent à $\mathring{I}$. □

3. dans le cas où $p = 0$, $\prod_{i=1}^{p} (X - x_i)^{m_i}$ est un produit vide qui vaut 1 par convention.

Théorème 3.6.4. Il existe une unique méthode de quadrature interpolatoire à $n+1$ nœuds d'ordre $2n+1$. Ses nœuds sont les racines du $(n+1)$-ième polynôme orthogonal pour la fonction poids w, et ses poids sont donnés par :

$$\forall i \in [\![0,n]\!], \quad \omega_i = \int_I L_i(x)\,w(x)\,\mathrm{d}x,$$

où $(L_i)_{0\leq i\leq n}$ sont les polynômes interpolateurs de Lagrange aux nœuds $(x_i)_{0\leq i\leq n}$. Cette méthode est appelée **méthode de quadrature de Gauss**.

Démonstration. Par analyse-synthèse.
• *Analyse.* Supposons que J soit d'ordre $2n+1$. Notons $\Pi_n(X) = \prod_{i=0}^n (X - x_i)$. Si $f \in \mathbb{R}_n[X]$, alors $f \times \Pi_n \in \mathbb{R}_{2n+1}[X]$ donc :

$$\int_I f(x)\Pi_n(x)\,w(x)\,\mathrm{d}x = \sum_{i=0}^n \omega_i f(x_i)\Pi_n(x_i) = 0$$

c'est-à-dire $\langle f, \Pi_n\rangle_w = 0$. Ainsi, Π_n est unitaire, de degré $n+1$ et orthogonal à tous les polynômes de $\mathbb{R}_n[X]$, c'est donc le $(n+1)$-ième polynôme orthogonal unitaire, et les $(x_i)_{0\leq i\leq n}$ sont ses racines. Enfin, si on calcule l'intégrale des polynômes interpolateurs de Lagrange, on obtient :

$$\forall i \in [\![0,n]\!], \quad \int_I L_i(x)\,w(x)\,\mathrm{d}x = \sum_{j=0}^n \omega_j L_i(x_j) = \omega_i.$$

• *Synthèse.* Considérons la méthode de quadrature :

$$J(f) = \sum_{i=0}^n \omega_i f(x_i), \quad \omega_i = \int_I L_i(x)\,w(x)\,\mathrm{d}x,$$

avec $(x_i)_{0\leq i\leq n}$ les racines de P_{n+1} et $(L_i)_{0\leq i\leq n}$ les polynômes interpolateurs de Lagrange. D'après la proposition 3.2.9, cette méthode est d'ordre n. Montrons qu'elle est d'ordre $2n+1$.
Soit $f \in \mathbb{R}_{2n+1}[X]$. On effectue la division euclidienne de f par P_{n+1} : il existe $Q, R \in \mathbb{R}[X]$ tels que $f = P_{n+1}Q + R$ avec $\deg(R) \leq n$. De plus, $\deg(f) \leq 2n+1$ impose que $\deg(Q) \leq n$. Par conséquent $P_{n+1} \perp Q$, donc :

$$\int_I f(x)\,w(x)\,\mathrm{d}x = \underbrace{\int_I P_{n+1}(x)Q(x)\,w(x)\,\mathrm{d}x}_{=0} + \int_I R(x)\,w(x)\,\mathrm{d}x = \sum_{i=0}^n \omega_i R(x_i),$$

car la méthode est exacte pour les polynômes de $\mathbb{R}_n[X]$. Enfin, on a $f(x_i) = P_{n+1}(x_i)Q(x_i) + R(x_i) = R(x_i)$ pour tout $i \in [\![0,n]\!]$, ce qui montre que :

$$\int_I f(x)\,w(x)\,\mathrm{d}x = \sum_{i=0}^n \omega_i f(x_i).$$

Ainsi, la méthode est exacte pour les polynômes de $\mathbb{R}_{2n+1}[X]$. □

Remarque 3.6.5. La méthode est en fait exactement d'ordre $2n+1$. En effet, si on considère le polynôme $f(X) = \prod_{i=0}^n (X - x_i)^2 \in \mathbb{R}_{2n+2}[X]$, alors on a

$$J(f) = \sum_{i=0}^n \omega_i f(x_i) = 0,$$

mais :

$$\int_I f(x)\,w(x)\,\mathrm{d}x > 0,$$

car c'est l'intégrale d'une fonction continue positive non nulle. Ainsi, une méthode à $n+1$ nœuds est d'ordre au plus $2n+1$, et la seule méthode de cet ordre est la méthode de quadrature de Gauss de degré n. ◂

Exemple 3.6.6 (Méthode de Gauss–Legendre)**.** On appelle **méthode de Gauss–Legendre** la méthode de quadrature de Gauss sur $[-1,1]$ pour le poids $w(x)=1$. Les polynômes orthogonaux sont alors les polynômes de Legendre. Le tableau ci-dessous donne les valeurs des nœuds et des poids pour les premières valeurs de n.

n	P_{n+1} (unitaire)	Racines x_i	Poids ω_i	Ordre
0	X	0	2	1
1	$X^2-\frac{1}{3}$	$-\frac{1}{\sqrt{3}},\frac{1}{\sqrt{3}}$	$1,1$	3
2	$X^3-\frac{3}{5}X$	$-\sqrt{\frac{3}{5}},0,\sqrt{\frac{3}{5}}$	$\frac{5}{9},\frac{8}{9},\frac{5}{9}$	5
3	$X^4-\frac{6}{7}X^2+\frac{3}{35}$	$\pm\sqrt{\frac{3}{7}-\frac{2}{7}\sqrt{\frac{6}{5}}},\ \pm\sqrt{\frac{3}{7}+\frac{2}{7}\sqrt{\frac{6}{5}}}$	$\frac{1}{2}-\frac{1}{6}\sqrt{\frac{5}{6}},\ \frac{1}{2}+\frac{1}{6}\sqrt{\frac{5}{6}}$	7

On dispose encore d'expressions exactes pour $n=4$. En revanche pour n quelconque, les x_i doivent être calculés numériquement, par exemple avec des méthodes itératives de type Newton. En Python, la fonction `roots_legendre` du module `scipy.special` calcule numériquement les nœuds et les poids de la méthode de Gauss–Legendre. ◂

Un autre avantage des méthodes de quadrature de Gauss est que les poids ω_i sont toujours positifs. Par conséquent, elles sont stables même pour n grand contrairement à Newton–Cotes.

Proposition 3.6.7. Les poids ω_i de la méthode de quadrature de Gauss sont strictement positifs. ◂

Démonstration. Considérons les polynômes interpolateurs de Lagrange L_i. Les polynômes L_i^2 sont de degré $2n$, donc la formule de quadrature est exacte :

$$\int_I L_i(x)^2 w(x)\,\mathrm{d}x = \sum_{j=0}^{n} \omega_j L_j(x_i)^2 = \omega_i.$$

Par conséquent, $\omega_i > 0$ pour tout $i \in [\![0,n]\!]$. □

On termine avec le résultat suivant, qu'on admettra (voir Demailly, 2016, chapitre III, section 3).

Théorème 3.6.8. On suppose que I est borné et on note $a=\inf I$ et $b=\sup I$. On considère la méthode de Gauss de degré n sur I pour une fonction de poids w. Pour tout $f\in\mathscr{C}^{2n+2}([a,b])$, il existe $\xi\in\,]a,b[$ tel que :

$$E(f)=\frac{f^{(2n+2)}(\xi)}{(2n+2)!}\int_a^b P_{n+1}(x)^2 w(x)\,\mathrm{d}x.$$

Exemple 3.6.9 (Méthode de Gauss–Legendre)**.** Pour $n=1$, la méthode de Gauss–Legendre évalue la fonction en 2 points. D'après le théorème précédent, pour $f\in\mathscr{C}^4([-1,1])$, il existe $\xi\in\,]-1,1[$ tel que :

$$E(f)=\frac{f^{(4)}(\xi)}{4!}\int_{-1}^{1} P_2(x)^2\,\mathrm{d}x.$$

On calcule :

$$\int_{-1}^{1} P_2(x)^2\,\mathrm{d}x=\int_{-1}^{1}\left(x^2-\frac{1}{3}\right)^2\mathrm{d}x=\frac{8}{45},$$

donc l'erreur est majorée par :

$$\left|E(f)\right|\le\frac{\|f^{(4)}\|_\infty}{4!}\times\frac{8}{45}=\frac{\|f^{(4)}\|_\infty}{135}.$$

À titre de comparaison, la méthode de Simpson sur $[-1,1]$ nécessite d'évaluer 3 fois la fonction, et l'erreur est majorée par :

$$\left|E(f)\right| \leq \frac{\|f^{(4)}\|_\infty}{90}.$$

Ce majorant est moins bon que celui de Gauss–Legendre, pour un nombre de points d'évaluation plus grand. ◀

Chapitre 4

Résolution numérique des équations différentielles ordinaires

Dans ce chapitre, on s'intéresse à la résolution approchée d'un problème de Cauchy d'ordre 1 :

$$\begin{cases} y'(t) = f(t, y(t)) \\ y(t_0) = x_0 \end{cases}$$

Ce genre de problème décrit l'évolution d'une quantité (ou d'un vecteur) $y(t)$ en fonction du temps, la loi d'évolution étant donnée par la fonction f. En dehors de quelques cas particuliers, on ne sait pas donner l'expression explicite de la solution (l'existence et l'unicité d'une solution n'étant pas toujours garanties). En pratique, on se tourne vers les méthodes numériques pour approcher la solution d'un tel problème.

Dans tout ce chapitre, on considère l'équation différentielle ordinaire [1] :

$$y'(t) = f(t, y(t)), \tag{$\mathscr{E}$}$$

où $f\colon I \times \mathbb{R} \to \mathbb{R}$ est une fonction continue et I est un intervalle ouvert non vide de $\mathbb{R}$.

Les résultats de ce cours se généralisent aux systèmes différentiels :

$$\begin{cases} y_1'(t) = f_1(t, y_1(t), \dots, y_d(t)) \\ \quad \vdots \\ y_d'(t) = f_d(t, y_1(t), \dots, y_d(t)) \end{cases}$$

qui s'écrivent aussi sous la forme ($\mathscr{E}$) avec $y = (y_1, \dots, y_d)$ et $f\colon I \times \mathbb{R}^d \to \mathbb{R}^d$. On se limite au cas $d = 1$ pour simplifier les démonstrations.

4.1 Résultats théoriques

Avant de chercher à approcher numériquement les solutions de ($\mathscr{E}$), donnons quelques résultats théoriques d'existence et d'unicité de celles-ci.

Définition 4.1.1. On appelle solution de l'équation différentielle ($\mathscr{E}$) un couple (y, J) où $J \subset I$ est un intervalle d'intérieur non vide, et $y\colon J \to \mathbb{R}$ est une fonction dérivable telle que :

$$\forall t \in J, \quad y'(t) = f(t, y(t)).$$

Soit $(t_0, x_0) \in I \times \mathbb{R}$. On appelle **problème de Cauchy** la recherche d'une solution (y, J) de ($\mathscr{E}$) satisfaisant la condition initiale $y(t_0) = x_0$, avec t_0 dans l'intérieur de J.

1. on parle d'équation différentielle *ordinaire*, en opposition aux équations aux dérivées partielles, c'est-à-dire les équations différentielles avec des fonctions de plusieurs variables.

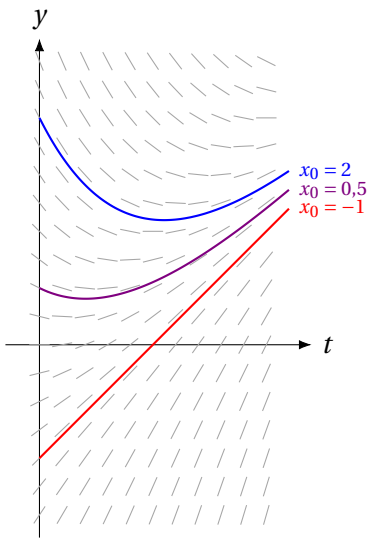


Figure 4.1 – Champ de pentes associé à l'équation $y'(t) = f(t, y(t))$ avec $f(t, y) = t - y$. En chaque point (t, y), un petit segment de pente $f(t, y)$ indique la direction que doit suivre une solution passant par ce point. On a tracé trois solutions issues d'une condition initiale $y(0) = x_0$ différente.

L'existence et l'unicité d'une solution au problème de Cauchy ne sont pas automatiques pour une fonction f continue quelconque[2]. Et quand elle existe, elle n'est pas toujours globale, c'est-à-dire définie sur I tout entier. Le théorème de Cauchy–Lipschitz affirme que c'est le cas si f satisfait une condition de Lipschitz par rapport à y.

Définition 4.1.2. Soit $J \subset I$. On dit que f est **globalement lipschitzienne en espace uniformément en temps** sur J s'il existe $L \geq 0$ tel que :

$$\forall t \in J, \quad \forall y_1, y_2 \in \mathbb{R}, \quad |f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2|.$$

Théorème 4.1.3 (Cauchy–Lipschitz, version globale)**.** On suppose que f est continue sur $I \times \mathbb{R}$ et globalement lipschitzienne en espace uniformément en temps sur tout segment $[a, b] \subset I$. Alors, pour tout $(t_0, x_0) \in I \times \mathbb{R}$, le problème de Cauchy de condition initiale (t_0, x_0) possède une unique solution $y : I \to \mathbb{R}$, définie sur I tout entier.

La preuve de ce théorème dépasse le cadre de ce cours. Vous l'étudierez plus en détail dans le cours *espaces vectoriels normés et calcul différentiel* du semestre 6.

Remarque 4.1.4. Ce théorème garantit que le problème de Cauchy est bien posé : il possède une unique solution, définie sur I tout entier. C'est cette hypothèse que l'on retiendra pour la suite du chapitre : elle assure que la quantité que l'on cherche à approcher numériquement, $y(t)$, est bien définie partout et unique. ◀

4.2 Exemples de méthodes numériques

On suppose dans toute la suite du chapitre que $f : I \times \mathbb{R} \to \mathbb{R}$ est continue et globalement lipschitzienne en espace uniformément en temps sur tout compact de I. Les hypothèses du théorème de Cauchy–Lipschitz sont satisfaites, donc pour tout $(t_0, x_0) \in I \times \mathbb{R}$, il existe une unique solution y de l'équation différentielle $(\mathscr{E})$ telle que $y(t_0) = x_0$.

2. la continuité de f garantit l'existence d'une solution locale (théorème de Cauchy–Peano–Arzelà), mais pas son unicité.

Soit $T > 0$ tel que $[t_0, t_0+T] \subset I$. On cherche à approcher y sur l'intervalle $[t_0, t_0+T]$ au sens suivant : si $(t_n)_{0\leq n\leq N}$ est une subdivision de $[t_0, t_0+T]$, construire une suite $(y_n)_{0\leq n\leq N}$ telle que $y_n \approx y(t_n)$. On note $h_n = t_{n+1} - t_n$ le pas entre les temps t_n et t_{n+1}, et :

$$h_{\max} = \max_{0\leq n\leq N-1} h_n,$$

le pas de la subdivision. L'objectif est de construire $(y_n)_{0\leq n\leq N}$ de sorte que **l'erreur globale** (aussi appelée erreur de discrétisation) :

$$\max_{0\leq n\leq N} \big|y_n - y(t_n)\big|,$$

tende vers 0 lorsque le pas de la subdivision tend vers 0.

4.2.a Schémas d'Euler

Soit y une solution de $(\mathscr{E})$. Considérons le développement limité de y en t :

$$\begin{aligned} y(t+h) &= y(t) + hy'(t) + \mathrm{o}(h) \\ &= y(t) + hf(t, y(t)) + \mathrm{o}(h). \end{aligned}$$

Pour $t = t_n$ et $h = h_n$ et en négligeant le reste, on a l'approximation :

$$y(t_{n+1}) \approx y(t_n) + h_n f(t_n, y(t_n)).$$

Connaissant la valeur au temps t_n, on peut approcher la valeur au temps t_{n+1} avec cette formule. On appelle **schéma d'Euler explicite** (schéma d'Euler en progression) la suite définie par récurrence :

$$y_{n+1} = y_n + h_n f(t_n, y_n),$$

avec y_0 une valeur initiale (typiquement $y_0 = x_0$, la valeur initiale du problème de Cauchy).

On peut aussi faire le développement limité à rebours :

$$y(t-h) = y(t) - hf(t, y(t)) + \mathrm{o}(h).$$

Pour $h = h_n$ et $t = t_{n+1}$, ceci conduit à l'approximation :

$$y(t_n) \approx y(t_{n+1}) - h_n f(t_{n+1}, y(t_{n+1})),$$

ce qu'on réécrit :

$$y(t_{n+1}) \approx y(t_n) + h_n f(t_{n+1}, y(t_{n+1})).$$

On appelle **schéma d'Euler implicite** (schéma d'Euler en régression) la suite définie par récurrence :

$$y_{n+1} = y_n + h_n f(t_{n+1}, y_{n+1}).$$

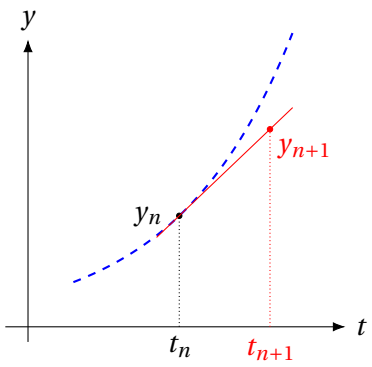


Figure 4.2 – Une itération du schéma d'Euler explicite. Partant du point (t_n, y_n), on avance le long de la tangente de pente $f(t_n, y_n)$ jusqu'à $t_{n+1} = t_n + h_n$, ce qui donne $y_{n+1} = y_n + h_n f(t_n, y_n)$.

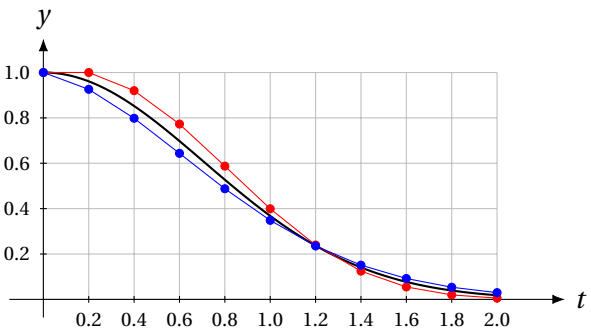


Figure 4.3 – Approximation de la solution de $y'(t) = -2ty(t)$, $y(0) = 1$, sur $[0,2]$ avec un pas $h = 0{,}2$. En noir : vraie solution. En rouge : schéma d'Euler explicite. En bleu : schéma d'Euler implicite.

Remarque 4.2.1. La méthode est dite implicite, car la relation de récurrence est en fait une équation à résoudre pour calculer y_{n+1}. En effet, y_{n+1} est un point fixe de l'application :

$$F_n(x) = y_n + h_n f(t_{n+1}, x).$$

Si f est globalement lipschitzienne en espace uniformément en temps de constante L, alors F_n est lipschitzienne de constante $h_n L$. Donc si $h_{\max} L < 1$, F_n est contractante pour tout n et le schéma d'Euler implicite est bien défini. De plus, on peut calculer y_{n+1} avec une méthode itérative. ◀

Le schéma d'Euler implicite est donc bien défini, à condition que le pas soit assez petit, mais il nécessite des étapes de calcul supplémentaires pour résoudre l'équation de récurrence numériquement. Selon la forme de la fonction f, il est parfois possible de résoudre l'équation algébriquement et de rendre le schéma explicite.

Exemple 4.2.2. Considérons l'équation $y' = -2ty$. C'est une équation différentielle linéaire admettant pour solutions les fonctions $y(t) = \lambda \exp(-t^2)$, $\lambda \in \mathbb{R}$.

- Le schéma d'Euler explicite s'écrit :

$$\begin{aligned} y_{n+1} &= y_n - 2h_n t_n y_n \\ &= (1 - 2h_n t_n) y_n. \end{aligned}$$

- Le schéma d'Euler implicite s'écrit :

$$y_{n+1} = y_n - 2h_n t_{n+1} y_{n+1}.$$

  Pour des temps $(t_n)_{0 \le n \le N}$ positifs, on peut isoler y_{n+1} et rendre le schéma explicite :

$$y_{n+1} = \frac{1}{1 + 2h_n t_{n+1}} y_n.$$

La figure 4.3 illustre ces deux méthodes. ◀

4.2.b Méthode de Taylor

Une généralisation naturelle de la méthode d'Euler est d'utiliser un développement limité à un ordre plus élevé. Considérons par exemple le développement à l'ordre 2. On suppose que f est de classe $\mathscr{C}^1$ sur $I \times \mathbb{R}$. Alors y' est de classe $\mathscr{C}^1$ sur I (car $y'(t) = f(t, y(t))$ pour tout $t \in I$), donc y est bien de classe $\mathscr{C}^2$ sur I. On peut alors écrire :

$$y(t+h) = y(t) + hy'(t) + \frac{1}{2}h^2 y''(t) + o(h^2).$$

On a $y'(t) = f(t, y(t))$, cherchons l'expression de $y''(t)$. En dérivant $y'(t) = f(t, y(t))$, on obtient :

$$\begin{aligned} y''(t) &= \frac{\mathrm{d}}{\mathrm{d}t} f(t, y(t)) \\ &= \frac{\partial f}{\partial t}(t, y(t)) \times 1 + \frac{\partial f}{\partial y}(t, y(t)) \times y'(t) \\ &= \frac{\partial f}{\partial t}(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t)) \times f(t, y(t)). \end{aligned}$$

Ainsi, si on note $f^{[1]} \in \mathscr{C}^0(I \times \mathbb{R})$ la fonction définie par :

$$\forall (t, y) \in I \times \mathbb{R}, \quad f^{[1]}(t, y) = \frac{\partial f}{\partial t}(t, y) + \frac{\partial f}{\partial y}(t, y) \times f(t, y),$$

alors on a $y''(t) = f^{[1]}(t, y(t))$ pour tout $t \in I$. Le développement de Taylor–Young à l'ordre 2 de y s'écrit :

$$y(t + h) = y(t) + h f(t, y(t)) + \frac{1}{2} h^2 f^{[1]}(t, y(t)) + \mathrm{o}(h^2),$$

ce qui conduit au schéma suivant, appelé **méthode de Taylor d'ordre 2** :

$$y_{n+1} \coloneqq y_n + h_n f(t_n, y_n) + \frac{1}{2} h_n^2 f^{[1]}(t_n, y_n).$$

Cette construction se généralise au développement à l'ordre p. Si f est de classe $\mathscr{C}^p(I \times \mathbb{R})$, alors la fonction y sera de classe $\mathscr{C}^{p+1}$ sur I, et le développement de Taylor–Young à l'ordre p s'écrit :

$$y(t + h) = y(t) + \sum_{k=1}^{p} \frac{h^k}{k!} y^{(k)}(t) + \mathrm{o}(h^p).$$

Il nous faut l'expression des dérivées $y^{(k)}$ en fonction de f.

Définition 4.2.3. Supposons que $f \in \mathscr{C}^p(I \times \mathbb{R})$. On définit par récurrence les fonctions $f^{[k]}$:

$$\forall k \in [\![1, p]\!], \quad \forall (t, y) \in I \times \mathbb{R}, \quad f^{[k]}(t, y) \coloneqq \frac{\partial f^{[k-1]}}{\partial t}(t, y) + \frac{\partial f^{[k-1]}}{\partial y}(t, y) \times f(t, y),$$

avec l'initialisation $f^{[0]} \coloneqq f$.

Le même calcul qui a montré que $y''(t) = f^{[1]}(t, y(t))$ permet de montrer par récurrence le résultat suivant.

Proposition 4.2.4. Si $f \in \mathscr{C}^p(I \times \mathbb{R})$, alors toute solution y de $(\mathscr{E})$ est de classe $\mathscr{C}^{p+1}$ sur I, et on a :

$$\forall k \in [\![0, p]\!], \quad \forall t \in I, \quad y^{(k+1)}(t) = f^{[k]}(t, y(t)). \qquad \blacktriangleleft$$

On définit la **méthode de Taylor d'ordre** p par :

$$y_{n+1} \coloneqq y_n + \sum_{k=1}^{p} \frac{1}{k!} h_n^k f^{[k-1]}(t_n, y_n).$$

Remarque 4.2.5. La méthode de Taylor n'est pas vraiment utilisée en pratique au-delà de $p = 2$, du fait de la complexité du calcul des $f^{[k]}$. $\blacktriangleleft$

4.2.c Méthode du point milieu

Un autre point de vue sur les méthodes d'Euler est le suivant. Si y est solution de l'équation différentielle $(\mathscr{E})$, le théorème fondamental de l'analyse permet d'écrire :

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + \int_{t_n}^{t_{n+1}} y'(s)\,\mathrm{d}s \\ &= y(t_n) + \int_{t_n}^{t_{n+1}} f(s, y(s))\,\mathrm{d}s. \end{aligned}$$

En appliquant la méthode des rectangles gauches, on obtient l'approximation :

$$y(t_{n+1}) \approx y(t_n) + h_n f(t_n, y(t_n)).$$

Cette approximation est celle qu'on a utilisée pour définir le schéma d'Euler explicite. Si on utilise la méthode des rectangles droits pour approcher l'intégrale, on retrouve le schéma d'Euler implicite. En utilisant d'autres méthodes de quadrature pour approcher l'intégrale, on trouve de nouvelles méthodes pour définir y_n. Considérons la méthode du point milieu :

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + \int_{t_n}^{t_{n+1}} f(s, y(s))\,\mathrm{d}s \\ &\approx y(t_n) + h_n f\left(t_n + \frac{h_n}{2}, y\left(t_n + \frac{h_n}{2}\right)\right). \end{aligned}$$

La valeur de $y(t_n + \frac{h_n}{2})$ est inconnue, mais si on dispose d'une approximation $\tilde{y}_{n+\frac{1}{2}}$, on peut définir :

$$y_{n+1} = y_n + h_n f\left(t_n + \frac{h_n}{2}, \tilde{y}_{n+\frac{1}{2}}\right).$$

Pour approcher $y(t_{n+\frac{1}{2}})$, on va utiliser le schéma d'Euler :

$$\tilde{y}_{n+\frac{1}{2}} = y_n + \frac{h_n}{2} f(t_n, y_n).$$

On obtient finalement le schéma suivant, appelé **méthode du point milieu** :

$$\begin{cases} \tilde{y}_{n+\frac{1}{2}} := y_n + \dfrac{h_n}{2} f(t_n, y_n) \\ y_{n+1} := y_n + h_n f\left(t_n + \dfrac{h_n}{2}, \tilde{y}_{n+\frac{1}{2}}\right) \end{cases}$$

Plus généralement, toute méthode de quadrature va nous permettre de définir un schéma numérique pour approcher la solution d'une équation différentielle, comme on le verra à la section 4.4.

4.3 Étude des méthodes à un pas

Les méthodes qu'on a vues précédemment peuvent toutes se mettre sous une même forme.

Définition 4.3.1. On appelle **méthode à un pas** un schéma numérique de la forme :

$$y_{n+1} = y_n + h_n \Phi(t_n, y_n, h_n),$$

où $\Phi\colon [t_0, t_0 + T] \times \mathbb{R} \times [0, H] \to \mathbb{R}$ est une application continue.

Remarque 4.3.2. Il existe des méthodes à pas multiples, dans lesquelles y_{n+1} est calculé à partir de plusieurs valeurs antérieures $y_n, y_{n-1}, \ldots, y_{n-q+1}$. Nous ne les étudierons pas dans ce cours. ◀

Exemple 4.3.3. Voici l'expression de la fonction Φ pour les méthodes précédentes.

1. Euler explicite : $\Phi(t,y,h) = f(t,y)$.
2. Euler implicite : $\Phi(t,y,h) = f(t+h, x(t,y,h))$ où $x(t,y,h)$ est l'unique point fixe de la fonction :
$$F_{y,h}(x) = y + hf(t+h,x).$$
On peut montrer que si $h \in [0,H]$ avec $HL < 1$, alors $x(t,y,h)$ est bien défini et dépend continument de (t,y,h).
3. Taylor d'ordre p : $\Phi(t,y,h) = \sum_{k=0}^{p-1} \frac{1}{(k+1)!} h^k f^{[k]}(t,y)$.
4. Point milieu : c'est bien une méthode à un pas, même si elle utilise deux points y_n et $\tilde{y}_{n+\frac{1}{2}}$ pour calculer y_{n+1}. En effet, la valeur de $\tilde{y}_{n+\frac{1}{2}}$ est calculée à partir de y_n, donc y_{n+1} ne dépend finalement que de y_n. La fonction Φ est définie par :
$$\Phi(t,y,h) = f\left(t + \frac{h}{2}, y + \frac{h}{2} f(t,y)\right).$$ ◂

Dans la suite, on fixe une méthode à un pas définie par une fonction Φ. L'objectif est de déterminer à quelle condition l'erreur globale :
$$\max_{0 \le n \le N} |y_n - y(t_n)|,$$
tend vers 0 quand $h_{\max} \to 0$, et de déterminer la vitesse de convergence.

La difficulté est la suivante. Partant de y_0, la méthode calcule une valeur approchée y_1 de $y(t_1)$. Cette valeur approchée y_1 est ensuite utilisée pour calculer une valeur approchée y_2 de $y(t_2)$, et ainsi de suite. À chaque pas de la méthode, on commet une erreur d'approximation qui se propage sur les calculs suivants. Il y a donc deux sources d'approximation :

- $\Phi(t_n, y_n, h_n)$ calcule une valeur approchée de l'accroissement $\frac{y(t_{n+1})-y(t_n)}{h_n}$ pour calculer y_{n+1} ;
- la valeur y_n utilisée dans le calcul de Φ est elle-même une approximation de $y(t_n)$.

Tout l'enjeu de l'étude est de contrôler la propagation de ces erreurs d'approximation.

4.3.a Consistance

Définition 4.3.4. Si y est une solution de l'équation différentielle $(\mathscr{E})$, on appelle **erreur de consistance** relativement à y les quantités :
$$\forall n \in [\![0, N-1]\!], \quad \varepsilon_n := y(t_{n+1}) - y(t_n) - h_n \Phi(t_n, y(t_n), h_n).$$
On dit que la méthode est **consistante** si pour toute solution y de l'équation différentielle $(\mathscr{E})$ et toute subdivision $(t_n)_{0 \le n \le N}$, on a :
$$\lim_{h_{\max} \to 0} \sum_{n=0}^{N-1} |\varepsilon_n| = 0.$$

L'erreur de consistance s'écrit $\varepsilon_n = y(t_{n+1}) - \hat{y}_{n+1}$ où :
$$\hat{y}_{n+1} = y(t_n) + h_n \Phi(t_n, y(t_n), h_n).$$
L'erreur de consistance est l'erreur entre la vraie valeur $y(t_{n+1})$ au temps t_{n+1}, et la valeur approchée $\hat{y}_{n+1}$ qu'on aurait si la valeur au temps t_n était correcte, c'est-à-dire si on avait $y_n = y(t_n)$.

On verra que si la méthode a de bonnes propriétés de stabilité, l'erreur globale est contrôlée par la somme des erreurs de consistance. Si cette somme tend vers 0 avec le pas, c'est-à-dire si la méthode est consistante, alors l'erreur globale tend vers 0.

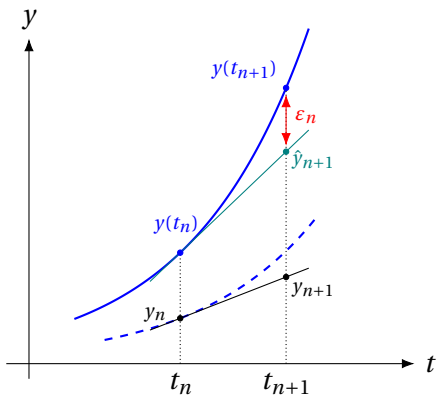


Figure 4.4 – Erreur de consistance, en rouge. Essentiellement, l'erreur globale $y(t_{n+1}) - y_{n+1}$ cumule l'erreur de consistance et l'erreur $y_n - y(t_n)$ déjà présente au départ.

Lemme 4.3.5. Pour toute solution y de $(\mathscr{E})$ et toute subdivision $(t_n)_{0\leq n\leq N}$ de $[t_0, t_0+T]$, on a :

$$\lim_{h_{\max}\to 0} \sum_{n=0}^{N-1} |\varepsilon_n| = \int_{t_0}^{t_0+T} \big|f(s,y(s)) - \Phi(s,y(s),0)\big|\,\mathrm{d}s. \qquad \blacktriangleleft$$

Démonstration. Soit y une solution de $(\mathscr{E})$. Par le théorème des accroissements finis, il existe $\tau_n \in]t_n, t_{n+1}[$ tel que :

$$y(t_{n+1}) - y(t_n) = h_n y'(\tau_n) = h_n f(\tau_n, y(\tau_n)).$$

On décompose :

$$\begin{aligned}
\varepsilon_n &= y(t_{n+1}) - y(t_n) - h_n\Phi(t_n, y(t_n), h_n) \\
&= h_n\big(f(\tau_n, y(\tau_n) - \Phi(t_n, y(t_n), h_n)\big) \\
&= h_n \underbrace{\big(f(\tau_n, y(\tau_n) - \Phi(\tau_n, y(\tau_n), 0)\big)}_{=:A_n} + h_n \underbrace{\big(\Phi(\tau_n, y(\tau_n), 0) - \Phi(t_n, y(t_n), h_n)\big)}_{=:B_n}.
\end{aligned}$$

• *Limite de* $\sum h_n|A_n|$. On écrit :

$$\sum_{n=0}^{N-1} h_n|A_n| = \sum_{n=0}^{N-1} \big|f(\tau_n, y(\tau_n)) - \Phi(\tau_n, y(\tau_n), 0)\big|(t_{n+1} - t_n).$$

On reconnait la somme de Riemann associée à la fonction continue $t \mapsto |f(t,y(t)) - \Phi(t,y(t),0)|$, pour la subdivision $(t_n)_{0\leq n\leq N}$ et le pointage $(\tau_n)_{0\leq n\leq N-1}$. Par conséquent :

$$\lim_{h_{\max}\to 0} \sum_{n=0}^{N-1} h_n|A_n| = \int_{t_0}^{t_0+T} \big|f(s,y(s)) - \Phi(s,y(s),0)\big|\,\mathrm{d}s.$$

• *Limite de* $\sum h_n|B_n|$. La fonction $(t,h) \mapsto \Phi(t,y(t),h)$ est continue sur le compact $[t_0, t_0+T] \times [0,H]$, donc elle est uniformément continue :

$$\begin{aligned}
\forall \varepsilon > 0, \quad \exists \delta > 0, \quad \forall (t,h),(t',h') \in [t_0, t_0+T] \times [0,H], \\
|t-t'| \leq \delta \text{ et } |h-h'| \leq \delta \implies |\Phi(t,y(t),h) - \Phi(t',y(t'),h')| \leq \varepsilon.
\end{aligned}$$

Introduisons le module de continuité :

$$\omega(\delta) := \sup\big\{|\Phi(t,y(t),h) - \Phi(t',y(t'),h')| : |t-t'| \leq \delta \text{ et } |h-h'| \leq \delta\big\}.$$

La continuité uniforme équivaut à $\lim_{\delta\to 0} \omega(\delta) = 0$. Dans le cas de B_n, on a :

$$\forall n \in [\![0, N-1]\!], \quad \big|\Phi(\tau_n, y(\tau_n), 0) - \Phi(t_n, y(t_n), h_n)\big| \leq \omega(h_{\max}),$$

car $|\tau_n - t_n| \leq h_n \leq h_{\max}$ et $|0 - h_n| \leq h_{\max}$. Par conséquent :

$$\sum_{n=0}^{N-1} h_n |B_n| \leq \sum_{n=0}^{N-1} h_n \omega(h_{\max}) = T\,\omega(h_{\max}) \xrightarrow[h_{\max}\to 0]{} 0.$$

• *Limite de* $\sum|\varepsilon_n|$. Par inégalité triangulaire inversée :

$$\left| \sum_{n=0}^{N-1} |\varepsilon_n| - \sum_{n=0}^{N-1} h_n |A_n| \right| \leq \sum_{n=0}^{N-1} |\varepsilon_n - h_n A_n| = \sum_{n=0}^{N-1} h_n |B_n| \xrightarrow[h_{\max}\to 0]{} 0.$$

Donc :

$$\lim_{h_{\max}\to 0} \sum_{n=0}^{N-1} |\varepsilon_n| = \lim_{h_{\max}\to 0} \sum_{n=0}^{N-1} h_n |A_n| = \int_{t_0}^{t_0+T} \big| f(s, y(s)) - \Phi(s, y(s), 0) \big| \, \mathrm{d}s. \qquad \square$$

On en déduit la caractérisation suivante de la consistance d'une méthode à un pas.

Théorème 4.3.6. La méthode est consistante si et seulement si :

$$\forall (t, y) \in [t_0, t_0 + T] \times \mathbb{R}, \quad \Phi(t, y, 0) = f(t, y).$$

Démonstration. Supposons que $\Phi(t, y, 0) = f(t, y)$ pour tout $(t, y) \in [t_0, t_0 + T] \times \mathbb{R}$. Alors pour $y = y(t)$ une solution de $(\mathscr{E})$, d'après le lemme 4.3.5, pour toute subdivision on a :

$$\lim_{h_{\max}\to 0} \sum_{n=0}^{N-1} |\varepsilon_n| = \int_{t_0}^{t_0+T} \big| f(s, y(s)) - \Phi(s, y(s), 0) \big| \, \mathrm{d}s = 0,$$

donc la méthode est consistante.
Réciproquement, supposons que la méthode est consistante. Alors pour toute solution y de $(\mathscr{E})$, toujours d'après le lemme 4.3.5, on doit avoir :

$$\int_{t_0}^{t_0+T} \big| f(s, y(s)) - \Phi(s, y(s), 0) \big| \, \mathrm{d}s = 0,$$

c'est-à-dire :

$$\forall s \in [t_0, t_0 + T], \quad f(s, y(s)) = \Phi(s, y(s), 0).$$

Soit $(t, x) \in [t_0, t_0 + T] \times \mathbb{R}$. Par le théorème de Cauchy–Lipschitz, il existe une solution de $(\mathscr{E})$ telle que $y(t) = x$. Pour une telle solution, on a $f(t, x) = \Phi(t, x, 0)$. Ceci étant vrai pour tout (t, x), le théorème est démontré. $\square$

Exemple 4.3.7. Avec ce théorème, on vérifie que toutes les méthodes évoquées précédemment (Euler, Taylor, point milieu) sont consistantes. ◀

4.3.b Stabilité

Le calcul numérique de y_{n+1} à partir de y_n se fait en pratique avec une erreur d'arrondis. Pour que la méthode soit utilisable, il faut que la propagation de ces erreurs soit contrôlée.

Définition 4.3.8. On dit que la méthode est **stable** s'il existe une constante $S \geq 0$ appelée **constante de stabilité** telle que pour tout h et pour toutes suites $(y_n)_{0\leq n\leq N}$, $(z_n)_{0\leq n\leq N}$, $(\delta_n)_{0\leq n\leq N}$ satisfaisant :

$$\forall n \in [\![0, N-1]\!], \quad y_{n+1} = y_n + h_n \Phi(t_n, y_n, h_n),$$
$$\forall n \in [\![0, N-1]\!], \quad z_{n+1} = z_n + h_n \Phi(t_n, z_n, h_n) + \delta_n,$$

alors on a :

$$\max_{0\leq n\leq N} |y_n - z_n| \leq S \left(|y_0 - z_0| + \sum_{n=0}^{N-1} |\delta_n| \right).$$

Notons que la stabilité est une propriété de la méthode et n'a rien à voir avec l'équation différentielle. Si la fonction Φ vérifie une condition de Lipschitz en espace, on montre que la méthode est stable.

Théorème 4.3.9. On suppose que Φ est globalement lipschitzienne en sa 2e variable uniformément en ses autres variables, c'est-à-dire qu'il existe $\Lambda \geq 0$ telle que :

$$\forall t \in [t_0, t_0+T], \quad \forall y_1, y_2 \in \mathbb{R}, \quad \forall h \in [0, H], \quad \big|\Phi(t, y_1, h) - \Phi(t, y_2, h)\big| \leq \Lambda |y_1 - y_2|.$$

Alors la méthode est stable, de constante de stabilité $S = \mathrm{e}^{\Lambda T}$.

La démonstration s'appuie sur le lemme suivant, dont la démonstration par récurrence est immédiate.

Lemme 4.3.10 (Grönwall discret)**.** Soient $(e_n)_{n\in\mathbb{N}}$, $(a_n)_{n\in\mathbb{N}}$ et $(b_n)_{n\in\mathbb{N}}$ des suites positives telles que :

$$\forall n \in \mathbb{N}, \quad e_{n+1} \leq a_n e_n + b_n.$$

Alors on a[3] :

$$\forall n \in \mathbb{N}, \quad e_n \leq \left(\prod_{k=0}^{n-1} a_k\right) e_0 + \sum_{j=0}^{n-1} \left(\prod_{k=j+1}^{n-1} a_k\right) b_j. \qquad \blacktriangleleft$$

Démonstration du théorème 4.3.9. En faisant la différence de y_{n+1} et z_{n+1}, on a :

$$\begin{aligned} |y_{n+1} - z_{n+1}| &\leq |y_n - z_n| + h_n \big|\Phi(t_n, y_n, h) - \Phi(t_n, z_n, h)\big| + |\delta_n| \\ &\leq (1 + \Lambda h_n)|y_n - z_n| + |\delta_n|. \end{aligned}$$

On applique le lemme de Grönwall discret aux suites $e_n = |y_n - z_n|$, $a_n = 1 + \Lambda h_n$ et $b_n = |\delta_n|$:

$$\forall n \in [\![0, N]\!], \quad |y_n - z_n| \leq \left(\prod_{k=0}^{n-1} (1 + \Lambda h_k)\right) |y_0 - z_0| + \sum_{j=0}^{n-1} \left(\prod_{k=j+1}^{n-1} (1 + \Lambda h_k)\right) |\delta_j|.$$

On utilise la majoration $1 + x \leq \mathrm{e}^x$:

$$\prod_{k=j+1}^{n-1} (1 + \Lambda h_k) \leq \prod_{k=0}^{n-1} (1 + \Lambda h_k) \leq \prod_{k=0}^{n-1} \mathrm{e}^{\Lambda h_k} = \mathrm{e}^{\Lambda(t_n - t_0)}.$$

D'où :

$$\forall n \in [\![0, N]\!], \quad |y_n - z_n| \leq \mathrm{e}^{\Lambda(t_n - t_0)} \left(|y_0 - z_0| + \sum_{j=0}^{n-1} |\delta_j|\right).$$

Finalement :

$$\max_{0 \leq n \leq N} |y_n - z_n| \leq \mathrm{e}^{\Lambda T} \left(|y_0 - z_0| + \sum_{j=0}^{N-1} |\delta_j|\right). \qquad \square$$

Exemple 4.3.11. Supposons que la fonction f est globalement lipschitzienne en espace uniformément en temps sur $[t_0, t_0 + T]$ de constante L. Le théorème précédent s'applique aux cas suivants :

1. Euler explicite, avec $\Lambda = L$. La constante de stabilité est donc $S = \mathrm{e}^{LT}$.
2. Euler implicite, avec $\Lambda = \frac{L}{1-LH}$ (si $LH < 1$). La constante de stabilité est $S = \mathrm{e}^{\frac{LT}{1-LH}}$.
3. Point milieu, avec $\Lambda = L\big(1 + \frac{LH}{2}\big)$. La constante de stabilité est $S = \mathrm{e}^{LT(1+\frac{LH}{2})}$.

Pour la méthode de Taylor, c'est plus délicat car cette méthode fait intervenir les dérivées partielles de f dont on ne sait rien. Si les $f^{[k]}$ sont aussi globalement lipschitziennes en espace uniformément en temps, alors la méthode de Taylor est stable. $\blacktriangleleft$

Remarque 4.3.12. La constante de stabilité devient vite très grande si T ou L est grand. Dans ce cas, on parle de problèmes raides (*stiff equations* en anglais). Des méthodes spécifiques existent pour étudier et approcher ce genre de problèmes. En ce qui nous concerne, on évitera d'utiliser les méthodes vues dans ce cours pour approcher une solution en temps long, ou pour des équations différentielles avec une constante de Lipschitz L trop grande. $\blacktriangleleft$

3. avec la convention qu'une somme vide est nulle et qu'un produit vide vaut 1.

4.3.c Convergence et ordre d'une méthode

Définition 4.3.13. On dit que la méthode est **convergente** si pour toute solution y de $(\mathscr{E})$ et toute subdivision $(t_n)_{0\le n\le N}$, on a :

$$\max_{0\le n\le N} |y_n - y(t_n)| \to 0,$$

lorsque $(h_{\max}, y_0) \to (0, y(t_0))$.

Théorème 4.3.14. Si la méthode est stable et consistante, alors elle est convergente.

Démonstration. Par définition de l'erreur de consistance, on a :

$$y(t_{n+1}) = y(t_n) + h_n \Phi(t_n, y(t_n), h) + \varepsilon_n.$$

Puisque la méthode est stable, on a :

$$|y_n - y(t_n)| \le S\left(|y_0 - y(t_0)| + \sum_{k=0}^{n-1} |\varepsilon_k|\right).$$

avec S la constante de stabilité. Par conséquent :

$$\max_{0\le n\le N} |y_n - y(t_n)| \le S\left(|y_0 - y(t_0)| + \sum_{k=0}^{N-1} |\varepsilon_k|\right).$$

Puisque la méthode est consistante, le majorant tend vers 0 quand $h_{\max} \to 0$ et $y_0 \to y(t_0)$. □

On s'intéresse maintenant à la vitesse de convergence de l'erreur globale.

Définition 4.3.15. On dit que la méthode est **consistante à l'ordre** p si pour toute solution y de l'équation différentielle, il existe $C > 0$ tel que pour tout h, les erreurs de consistance relatives à y satisfont :

$$\forall n \in [\![0, N-1]\!], \quad |\varepsilon_n| \le C h_n^{p+1}.$$

Attention au $p+1$ en puissance de h_n, alors que l'ordre vaut p. La raison de cette définition est que si $|\varepsilon_n| = \mathrm{O}(h_n^{p+1})$, alors la somme des erreurs de consistance est en $\mathrm{O}(h_{\max}^p)$ comme on le verra.

Le théorème suivant donne une condition nécessaire et suffisante sur Φ pour que la méthode soit consistante à l'ordre p.

Théorème 4.3.16. Supposons que $f \in \mathscr{C}^p([t_0, t_0+T]\times\mathbb{R})$ et que pour tout $k \le p$, $\frac{\partial^k \Phi}{\partial h^k}$ existe et est continue sur $[t_0, t_0+T]\times\mathbb{R}\times[0, H]$. Alors la méthode est consistante à l'ordre p si et seulement si :

$$\forall (t,y) \in [t_0, t_0+T]\times\mathbb{R}, \quad \forall k \in [\![0, p-1]\!], \quad \frac{\partial^k \Phi}{\partial h^k}(t, y, 0) = \frac{1}{k+1} f^{[k]}(t, y), \tag{4.1}$$

où les $f^{[k]}$ ont été définis par la définition 4.2.3.

Remarque 4.3.17. Pour une fonction f de classe $\mathscr{C}^1$, les théorèmes 4.3.6 et 4.3.16 entrainent :

$$\text{la méthode est d'ordre } 1 \iff \forall (t,y),\ \Phi(t,y,0) = f(t,y) \iff \text{la méthode est consistante.}$$ ◀

Démonstration. Soit y une solution de $(\mathscr{E})$. La formule de Taylor–Lagrange appliquée à y et Φ donne l'existence de $\tau_n \in]t_n, t_{n+1}[$ et $\theta_n \in]0, h_n[$ tels que :

$$y(t_{n+1}) - y(t_n) = \sum_{k=1}^{p} \frac{y^{(k)}(t_n)}{k!} h_n^k + \frac{y^{(p+1)}(\tau_n)}{(p+1)!} h_n^{p+1} = \sum_{k=1}^{p} \frac{f^{[k-1]}(t_n, y(t_n))}{k!} h_n^k + \frac{f^{[p]}(\tau_n, y(\tau_n))}{(p+1)!} h_n^{p+1},$$

$$\Phi(t_n, y(t_n), h_n) = \sum_{k=0}^{p-1} \frac{1}{k!} \frac{\partial^k \Phi}{\partial h^k}(t_n, y(t_n), 0) h_n^k + \frac{1}{p!} \frac{\partial^p \Phi}{\partial h^p}(t_n, y(t_n), \theta_n) h_n^p.$$

Les erreurs de consistance relativement à y sont donc :

$$\begin{aligned}\varepsilon_n &= y(t_{n+1}) - y(t_n) - h_n\Phi(t_n, y(t_n), h_n)\\ &= \sum_{k=0}^{p-1}\left(\frac{f^{[k]}(t_n, y(t_n))}{k+1} - \frac{\partial^{(k)}\Phi}{\partial h^k}(t_n, y(t_n), 0)\right)\frac{h_n^{k+1}}{k!} + h_n^{p+1}\left(\frac{f^{[p]}(\tau_n, y(\tau_n))}{(p+1)!} - \frac{1}{p!}\frac{\partial^p\Phi}{\partial h^p}(t_n, y(t_n), \theta_n)\right).\end{aligned}$$

Les fonctions $t \mapsto f^{[p]}(t, y(t))$ et $(t,h) \mapsto \frac{\partial^p\Phi}{\partial h^p}(t, y(t), h)$ sont continues sur les compacts $[t_0, t_0+T]$ et $[t_0, t_0+T] \times [0, H]$, donc elles sont bornées. Il existe une constante $C > 0$ qui dépend de y et de Φ telle que :

$$\forall n \in [\![0, N]\!] \quad \left|\frac{f^{[p]}(\tau_n, y(\tau_n))}{(p+1)!} - \frac{1}{p!}\frac{\partial^p\Phi}{\partial h^p}(t_n, y(t_n), \theta_n)\right| \leq C.$$

Par conséquent, les erreurs de consistance relativement à y sont en $\mathrm{O}(h_n^{p+1})$ si et seulement si :

$$\forall n \in [\![0, N]\!], \quad \forall k \in [\![0, p-1]\!], \quad \frac{\partial^k\Phi}{\partial h^k}(t_n, y(t_n), 0) = \frac{1}{k+1}f^{[k]}(t_n, y(t_n)). \tag{4.2}$$

On peut maintenant démontrer l'équivalence du théorème.

($\Longrightarrow$) Si (4.1) est vérifiée, alors (4.2) aussi donc la méthode est d'ordre p.

($\Longleftarrow$) Réciproquement, si la méthode est d'ordre p, on doit avoir (4.2) pour toute solution y et toute subdivision. Soit $t \in [t_0, t_0+T]$ et soit $x \in \mathbb{R}$. Par le théorème de Cauchy–Lipschitz, il existe une solution y de $(\mathscr{E})$ telle que $y(t) = x$. Soit n_0 tel que $t \in [t_{n_0}, t_{n_0+1}]$. Alors $t_{n_0} \to t$ lorsque $h_{\max} \to 0$, et par continuité $y(t_{n_0}) \to y(t) = x$, donc en prenant $n = n_0$ et $h_{\max} \to 0$ dans (4.2), on obtient que :

$$\forall k \in [\![0, p-1]\!], \quad \frac{\partial^k\Phi}{\partial h^k}(t, x, 0) = \frac{1}{k+1}f^{[k]}(t, x).$$

Ceci étant vrai pour tout $t \in [t_0, t_0+T]$ et tout $x \in \mathbb{R}$, on a prouvé (4.1). □

Exemple 4.3.18. Passons en revue l'ordre des méthodes étudiées jusqu'à présent.

1. La méthode d'Euler explicite est d'ordre 1 mais pas d'ordre 2. En effet, on a déjà vu que $\Phi(t, y, 0) = f(t, y)$ (consistance), mais :
$$\frac{\partial\Phi}{\partial h}(t, y, 0) = 0 \neq \frac{1}{2}f^{[1]}(t, y).$$
2. On peut vérifier que la méthode d'Euler implicite est d'ordre 1 mais pas d'ordre 2.
3. La méthode de Taylor d'ordre p est, par construction, consistante à l'ordre p si $f \in \mathscr{C}^p(I \times \mathbb{R})$. En effet :
$$\Phi(t, y, h) = \sum_{k=0}^{p-1}\frac{1}{(k+1)!}h^k f^{[k]}(t, y),$$
donc pour tout $k \in [\![0, p-1]\!]$:
$$\frac{\partial^k\Phi}{\partial h^k}(t, y, 0) = \frac{k!}{(k+1)!}f^{[k]}(t, y) = \frac{1}{k+1}f^{[k]}(t, y).$$
4. La méthode du point milieu est d'ordre 2 si $f \in \mathscr{C}^2(I \times \mathbb{R})$. En effet, elle est au moins d'ordre 1 car elle est consistante. De plus :
$$\Phi(t, y, h) = f\left(t + \frac{h}{2}, y + \frac{h}{2}f(t, y)\right),$$
donc :
$$\frac{\partial\Phi}{\partial h}(t, y, h) = \frac{1}{2}\frac{\partial f}{\partial t}\left(t + \frac{h}{2}, y + \frac{h}{2}f(t, y)\right) + \frac{1}{2}f(t, y)\frac{\partial f}{\partial y}\left(t + \frac{h}{2}, y + \frac{h}{2}f(t, y)\right).$$
Pour $h = 0$, on a bien $\frac{\partial\Phi}{\partial h}(t, y, h) = \frac{1}{2}f^{[1]}(t, y)$, donc la méthode est d'ordre 2. On peut vérifier qu'elle n'est pas d'ordre 3. ◀

Théorème 4.3.19. Si la méthode est stable et consistante à l'ordre p, alors pour toute solution y de l'équation différentielle $(\mathscr{E})$, il existe $C>0$ tel que pour toute subdivision $(t_n)_{0\leq n\leq N}$:

$$\max_{0\leq n\leq N}|y_n-y(t_n)|\leq S\big(|y_0-y(t_0)|+CTh_{\max}^p\big).$$

Démonstration. En reprenant le raisonnement du théorème 4.3.14, on a :

$$\max_{0\leq n\leq N}|y_n-y(t_n)|\leq S\left(|y_0-y(t_0)|+\sum_{n=0}^{N-1}|\varepsilon_n|\right).$$

Puisque la méthode est consistante à l'ordre p, il existe $C>0$ tel que :

$$\forall n\in[\![0,N-1]\!],\quad |\varepsilon_n|\leq Ch_n^{p+1}.$$

Donc :

$$\begin{aligned}\sum_{n=0}^{N-1}|\varepsilon_n|&\leq C\sum_{n=0}^{N-1}h_n^{p+1}\\&\leq C\sum_{n=0}^{N-1}h_n\times h_{\max}^p\\&\leq CTh_{\max}^p.\end{aligned}$$

□

Remarque 4.3.20. Si on choisit $y_0=y(t_0)$ (condition initiale du problème de Cauchy), le théorème précédent affirme qu'une méthode consistante à l'ordre p a une erreur globale majorée par :

$$\max_{0\leq n\leq N}|y_n-y(t_n)|\leq SCT\,h_{\max}^p.$$

Quand $h_{\max}\to 0$, la vitesse de convergence de l'erreur globale est donc en $\mathrm{O}(h_{\max}^p)$. ◀

4.4 Méthodes de Runge–Kutta

4.4.a Construction et exemples

Considérons l'expression intégrale de la solution de l'équation différentielle :

$$y(t_{n+1})=y(t_n)+\int_{t_n}^{t_{n+1}}f(s,y(s))\,\mathrm{d}s.$$

On fixe $c_1,\dots,c_q\in[0,1]$ (pas nécessairement distincts) et on considère les temps intermédiaires :

$$t_{n,i}=t_n+c_ih_n.$$

L'idée des méthodes de Runge–Kutta est d'approcher l'intégrale par une méthode de quadrature sur les nœuds $(t_{n,i})_{0\leq i\leq q}$:

$$\int_{t_n}^{t_{n+1}}f(s,y(s))\,\mathrm{d}s\approx h_n\sum_{i=1}^{q}b_if(t_{n,i},y(t_{n,i})),$$

avec $(b_i)_{0\leq i\leq q}$ des poids. Si on dispose d'approximations $\tilde{y}_{n,i}\approx y(t_{n,i})$, on peut définir le schéma :

$$y_{n+1}=y_n+h_n\sum_{i=1}^{q}b_if(t_{n,i},\tilde{y}_{n,i}).$$

Pour approcher les $y(t_{n,i})$, on va de nouveau utiliser des méthodes de quadrature (potentiellement différente pour chacun). L'équation intégrale s'écrit :

$$y(t_{n,i})=y(t_n)+\int_{t_n}^{t_{n,i}}f(s,y(s))\,\mathrm{d}s,$$

on approche l'intégrale par une méthode de quadrature sur les mêmes nœuds $(t_{n,j})_{1\leq j\leq q}$:

$$\int_{t_n}^{t_{n,i}} f(s,y(s))\,\mathrm{d}s \approx h_n \sum_{j=1}^{q} a_{i,j} f(t_{n,j},\tilde{y}_{n,j}),$$

avec $a_{i,j}$ des poids.

On définit ainsi la **méthode de Runge–Kutta** à q étages :

$$\left\{\begin{array}{l} \left\{\begin{array}{l} t_{n,i} = t_n + c_i h_n \\ \tilde{y}_{n,i} = y_n + h_n \displaystyle\sum_{j=1}^{q} a_{i,j} f(t_{n,j},\tilde{y}_{n,j}) \end{array}\right\} \quad \forall i \in [\![0,q]\!] \\ y_{n+1} = y_n + h_n \displaystyle\sum_{i=1}^{q} b_i f(t_{n,i},\tilde{y}_{n,i}) \end{array}\right. \tag{RK$_q$}$$

La méthode est entièrement déterminée par les coefficients $(a_{i,j})_{1\leq i,j\leq q}$, $(b_i)_{1\leq i\leq q}$ et $(c_i)_{1\leq i\leq q}$. On résume la méthode par le tableau ci-dessous, appelé **tableau de Butcher** :

$$\begin{array}{c|cccc} c_1 & a_{1,1} & a_{1,2} & \cdots & a_{1,q} \\ c_2 & a_{2,1} & a_{2,2} & \cdots & a_{2,q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_q & a_{q,1} & a_{q,2} & \cdots & a_{q,q} \\ \hline & b_1 & b_2 & \cdots & b_q \end{array}$$

En toute généralité, les $(\tilde{y}_{n,i})_{1\leq i\leq q}$ sont définis implicitement par un système non linéaire de q équations. On peut montrer que si $h_{\max}$ est assez petit, ce système admet bien une unique solution qui peut être déterminée par des méthodes de point fixe. Le plus souvent on préfèrera utiliser les méthodes de Runge–Kutta pour lesquelles :

$$\forall i,j \in [\![1,q]\!], \quad i \leq j \implies a_{i,j} = 0,$$

de sorte que chaque $\tilde{y}_{n,i}$ s'exprime directement en fonction des $(\tilde{y}_{n,j})_{1\leq j\leq i-1}$ précédents. Cette condition revient à dire que la matrice des $(a_{i,j})_{1\leq i,j\leq q}$ est triangulaire inférieure stricte. Dans ce cas, la méthode de Runge–Kutta est explicite et les $(\tilde{y}_{n,i})_{1\leq i\leq q}$ se calculent par récurrence.

Remarque 4.4.1. Les méthodes de Runge–Kutta sont parfois présentées en version « standardisée ». Si on pose $g(u) = f(t_n + uh_n, y(t_n + uh_n))$ avec $u \in [0,1]$, un changement de variable permet d'écrire :

$$\int_{t_n}^{t_{n+1}} f(s,y(s))\,\mathrm{d}s = h_n \int_0^1 g(u)\,\mathrm{d}u, \quad \text{et} \quad \int_{t_n}^{t_{n,i}} f(s,y(s))\,\mathrm{d}s = h_n \int_0^{c_i} g(u)\,\mathrm{d}u.$$

Une méthode de Runge–Kutta est définie en se donnant une méthode de quadrature pour chacune de ces intégrales de g :

- une méthode de quadrature principale :

$$\int_0^1 g(u)\,\mathrm{d}u \approx \sum_{i=1}^{q} b_i g(c_i),$$

- des méthodes de quadratures intermédiaires :

$$\forall i \in [\![1,q]\!], \quad \int_0^{c_i} g(u)\,\mathrm{d}u \approx \sum_{j=1}^{q} a_{i,j} g(c_j). \qquad \blacktriangleleft$$

Remarque 4.4.2. Si on note $p_{n,i} = f(t_{n,i}, \tilde{y}_{n,i})$ la pente en $(t_{n,i}, \tilde{y}_{n,i})$, le schéma (RK_q) s'écrit aussi :

$$\left\{ \begin{array}{l} \left\{ \begin{array}{l} t_{n,i} = t_n + c_i h_n \\ p_{n,i} = f\left(t_{n,i},\, y_n + h_n \displaystyle\sum_{j=1}^{q} a_{i,j} p_{n,j} \right) \end{array} \right\} \quad \forall i \in [\![0, q]\!] \\ y_{n+1} = y_n + h_n \displaystyle\sum_{i=1}^{q} b_i p_{n,i} \end{array} \right.$$

On préférera cette forme lors de l'implémentation, car les quantités dont on a vraiment besoin pour calculer y_{n+1} sont les $(p_{n,i})_{1 \leq i \leq q}$. On évite ainsi d'évaluer f inutilement. ◂

Dans la suite, on supposera que toutes les méthodes de quadrature utilisées sont d'ordre 0 (exactes pour les constantes). Par conséquent, on a les relations :

$$\sum_{i=1}^{q} b_i = 1, \quad \text{et} \quad \forall i \in [\![1, q]\!], \quad c_i = \sum_{j=1}^{q} a_{i,j}. \tag{4.3}$$

Remarque 4.4.3. Pour une méthode de Runge–Kutta explicite, on a nécessairement $a_{1,j} = 0$ pour tout $j \in [\![1, q]\!]$ et $c_1 = 0$. Autrement dit, la première ligne du tableau de Butcher est entièrement nulle. On a donc $t_{n,1} = t_n$ et $y_{n,1} = y_n$. ◂

Exemple 4.4.4 (Méthodes RK_1)**.** La seule méthode RK_1 explicite a pour tableau de Butcher :

0	0
	1

C'est la méthode d'Euler explicite. La méthode d'Euler implicite est aussi une méthode RK_1, de tableau de Butcher :

1	1
	1

◂

Exemple 4.4.5 (Méthodes RK_2)**.** Les méthodes RK_2 les plus connues sont :

- **Méthode du point milieu.** C'est une méthode RK_2 explicite. La méthode de quadrature principale est la méthode du point milieu, et la méthode de quadrature intermédiaire est la méthode d'Euler explicite. Son tableau de Butcher est :

0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0
	0	1

- **Méthode de Heun.** Cette méthode consiste à utiliser la méthode des trapèzes comme méthode de quadrature principale :

$$\int_{t_n}^{t_{n+1}} f(s, y(s))\,\mathrm{d}s \approx h_n \frac{f(t_n, y(t_n)) + f(t_{n+1}, y(t_{n+1}))}{2}.$$

  Comme pour la méthode du point milieu, la valeur inconnue $y(t_{n+1})$ est approchée avec la méthode d'Euler explicite, ce qui conduit au schéma :

$$\left\{ \begin{array}{l} \tilde{y}_{n+1} = y_n + h_n f(t_n, y_n) \\ y_{n+1} = y_n + h_n \left(\dfrac{1}{2} f(t_n, y_n) + \dfrac{1}{2} f(t_{n+1}, \tilde{y}_{n+1}) \right). \end{array} \right.$$

C'est une méthode RK_2 explicite de tableau de Butcher :

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

- **Méthode de Crank–Nicolson.** Cette méthode utilise aussi la méthode des trapèzes comme méthode de quadrature principale :

$$\int_{t_n}^{t_{n+1}} f(s, y(s))\,\mathrm{d}s \approx h_n \frac{f(t_n, y(t_n)) + f(t_{n+1}, y(t_{n+1}))}{2},$$

mais au lieu d'approcher $y(t_{n+1})$, on laisse le schéma sous forme implicite :

$$y_{n+1} = y_n + h_n\left(\frac{1}{2}f(t_n, y_n) + \frac{1}{2}f(t_{n+1}, y_{n+1})\right).$$

On vérifie que le schéma est bien défini si $h_{\max}L < 2$, où L est la constante de Lipschitz de f. C'est une méthode RK_2 implicite, de tableau de Butcher :

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Toutes ces méthodes sont stables et d'ordre 2, comme on peut le vérifier (ce sera une conséquence des résultats qui suivent). ◀

Exemple 4.4.6 (Méthode RK_4)**.** La **méthode de Runge–Kutta classique** est donnée par le tableau :

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array}$$

Elle correspond à la méthode de quadrature principale :

$$\begin{aligned} \int_0^1 g(u)\,\mathrm{d}u &\approx \frac{1}{6}g(0) + \frac{2}{6}g\left(\tfrac{1}{2}\right) + \frac{2}{6}g\left(\tfrac{1}{2}\right) + \frac{1}{6}g(1) \\ &= \frac{1}{2}\left(\frac{1}{3}g(0) + \frac{4}{3}g\left(\tfrac{1}{2}\right) + \frac{1}{3}g(1)\right) \qquad \text{(Simpson)} \end{aligned}$$

et aux méthodes de quadratures intermédiaires :

$$\begin{aligned} &\text{pour } i = 2 : && \int_0^{\frac{1}{2}} g(u)\,\mathrm{d}u \approx \frac{1}{2}g(0) && \text{(rectangles gauches)} \\ &\text{pour } i = 3 : && \int_0^{\frac{1}{2}} g(u)\,\mathrm{d}u \approx \frac{1}{2}g\left(\tfrac{1}{2}\right) && \text{(rectangles droits)} \\ &\text{pour } i = 4 : && \int_0^{1} g(u)\,\mathrm{d}u \approx g\left(\tfrac{1}{2}\right) && \text{(point milieu)} \end{aligned}$$

Le schéma s'écrit :

$$\begin{cases} p_{n,1} = f(t_n, y_n) \\ p_{n,2} = f\left(t_n + \dfrac{h_n}{2}, y_n + \dfrac{h_n}{2} p_{n,1}\right) \\ p_{n,3} = f\left(t_n + \dfrac{h_n}{2}, y_n + \dfrac{h_n}{2} p_{n,2}\right) \\ p_{n,4} = f\left(t_n + h_n, y_n + h_n p_{n,3}\right) \\ y_{n+1} = y_n + h_n\left(\dfrac{1}{6} p_{n,1} + \dfrac{2}{6} p_{n,2} + \dfrac{2}{6} p_{n,3} + \dfrac{1}{6} p_{n,4}\right) \end{cases}$$

Cette méthode est parfois appelée « la » méthode de Runge–Kutta. Elle est stable et d'ordre 4. Elle est très utilisée en pratique car elle offre un bon compromis entre l'ordre et le nombre d'évaluations de f à chaque pas (4 évaluations par pas). ◀

4.4.b Consistance et stabilité

Si on considère l'écriture avec les pentes du schéma (voir remarque 4.4.2), la fonction Φ d'une méthode de Runge–Kutta s'écrit :

$$\Phi(t, y, h) = \sum_{i=1}^{q} b_i p_i(t, y, h),$$

où les fonctions p_i sont solution du système :

$$\forall i \in [\![1, q]\!], \quad p_i(t, y, h) = f\left(t + c_i h, y + h \sum_{j=1}^{q} a_{i,j} p_j(t, y, h)\right). \tag{4.4}$$

Théorème 4.4.7. Notons L la constante de Lipschitz de f, et posons :

$$\alpha = \max_{1 \le i \le q} \sum_{j=1}^{q} |a_{i,j}|.$$

Si $LH\alpha < 1$, alors le système (4.4) admet une unique solution, donc la méthode de Runge–Kutta est bien définie. De plus, elle est stable de constante de stabilité $S = \mathrm{e}^{\Lambda T}$ avec :

$$\Lambda := \frac{L}{1 - LH\alpha} \sum_{i=1}^{q} |b_i|.$$

Démonstration. Montrons que la méthode est bien définie. Posons $F\colon \mathbb{R}^q \to \mathbb{R}^q$ la fonction définie par :

$$\forall i \in [\![1, q]\!], \quad \forall x = (x_1, \dots, x_q) \in \mathbb{R}^q, \quad F_i(x) = f\left(t + c_i h, y + h \sum_{j=1}^{q} a_{i,j} x_j\right).$$

Alors (4.4) équivaut à dire que $\big(p_1(t, y, h), \dots, p_q(t, y, h)\big)$ est un point fixe de F. Montrons que F est contractante sur $\mathbb{R}^q$ pour la norme :

$$\forall x \in \mathbb{R}^q, \quad \|x\|_\infty := \max_{1 \le i \le q} |x_i|.$$

Soient $x, x' \in \mathbb{R}^q$, on a :

$$\forall i \in [\![1, q]\!], \quad \big|F_i(x) - F_i(x')\big| \le Lh \sum_{j=1}^{q} |a_{i,j}||x_j - x'_j| \le LH\left(\sum_{j=1}^{q} |a_{i,j}|\right)\|x - x'\|_\infty.$$

En prenant le maximum sur i, on obtient :

$$\|F(x) - F(x')\|_\infty \le LH\alpha \|x - x'\|_\infty.$$

Puisque $LH\alpha < 1$ par hypothèse, l'application F est contractante. Par le théorème de point fixe de Banach–Picard, F admet un unique point fixe. Les fonctions p_i sont donc bien définies. On admet qu'elles sont continues.

Montrons que la méthode est stable. Soient $y_1, y_2 \in \mathbb{R}$, on a :

$$\big|\Phi(t,y_1,h)-\Phi(t,y_2,h)\big| \le \sum_{i=1}^{q} |b_i| \big|p_i(t,y_1,h)-p_i(t,y_2,h)\big|.$$

Puisque les p_i sont solution de (4.4), on a :

$$\begin{aligned}\big|p_i(t,y_1,h)-p_i(t,y_2,h)\big| &\le L|y_1-y_2| + Lh\sum_{j=1}^{q}|a_{i,j}|\big|p_j(t,y_1,h)-p_j(t,y_2,h)\big| \\ &\le L|y_1-y_2| + LH\left(\sum_{j=1}^{q}|a_{i,j}|\right)\max_{1\le j\le q}\big|p_j(t,y_1,h)-p_j(t,y_2,h)\big|.\end{aligned}$$

En prenant le maximum sur i :

$$\max_{1\le i\le q}\big|p_i(t,y_1,h)-p_i(t,y_2,h)\big| \le L|y_1-y_2| + LH\alpha \max_{1\le i\le q}\big|p_i(t,y_1,h)-p_i(t,y_2,h)\big|,$$

ce qu'on réécrit :

$$\max_{i}\big|p_i(t,y_1,h)-p_i(t,y_2,h)\big| \le \frac{L}{1-LH\alpha}|y_1-y_2|,$$

car $LH\alpha < 1$. Finalement, on obtient :

$$\big|\Phi(t,y_1,h)-\Phi(t,y_2,h)\big| \le \sum_{i=1}^{q}|b_i|\big|p_i(t,y_1,h)-p_i(t,y_2,h)\big| \le \frac{L}{1-LH\alpha}\left(\sum_{i=1}^{q}|b_i|\right)|y_1-y_2|.$$

On conclut grâce au théorème 4.3.9. □

Remarque 4.4.8. Si les méthodes de quadrature sont d'ordre 0 (hypothèse (4.3)), et que leurs poids sont positifs, on a :

$$\sum_{i=1}^{q}|b_i| = \sum_{i=1}^{q} b_i = 1, \quad \text{et} \quad \sum_{j=1}^{q}|a_{i,j}| = \sum_{j=1}^{q} a_{i,j} = c_i,$$

donc $\alpha = \max_i c_i \le 1$. La constante de stabilité s'écrit alors $S = \mathrm{e}^{\Lambda T}$ avec :

$$\Lambda = \frac{L}{1-LH}.$$

Si H est petit devant L^{-1}, alors la constante de stabilité est essentiellement $S \approx \mathrm{e}^{LT}$ et la méthode de Runge–Kutta est aussi stable que la méthode d'Euler, quel que soit le nombre d'étages q. ◀

Remarque 4.4.9. Dans le cas d'une méthode explicite, celle-ci est toujours bien définie et stable, sans la condition $LH\alpha < 1$. En reprenant la preuve ci-dessus, on peut montrer que la constante de stabilité s'écrit $S = \mathrm{e}^{\Lambda T}$ avec :

$$\Lambda = L\sum_{i=1}^{q}|b_i|\sum_{k=0}^{i-1}(LH\alpha)^k,$$

voir Demailly (2016). Si $LH\alpha \ge 1$, alors la constante Λ croît avec q. Dans le cas où les b_i et les $\alpha_{i,j}$ sont positifs, et sous l'hypothèse (4.3), on peut simplifier :

$$\Lambda = L\sum_{k=0}^{q-1}(LH)^k.$$ ◀

Proposition 4.4.10. La méthode de Runge–Kutta est consistante si et seulement si $\sum_{i=1}^{q} b_i = 1$. ◀

Démonstration. Calculons $\Phi(t,y,0)$. Les $p_i(t,y,h)$ sont l'unique solution du système :

$$\forall i \in [\![1,q]\!], \quad p_i(t,y,h) = f\left(t + c_i h, y + h\sum_{j=1}^{q} a_{i,j} p_j(t,y,h)\right).$$

Pour $h=0$, on a $p_i(t,y,0) = f(t,y)$ pour tout $i \in [\![1,q]\!]$. Par conséquent :

$$\Phi(t,y,0) = \sum_{i=1}^{q} b_i p_i(t,y,0) = \left(\sum_{i=1}^{q} b_i\right) f(t,y).$$

D'après le théorème 4.3.6, la méthode est consistante si et seulement si $\sum_{i=1}^{q} b_i = 1$. □

Si on utilise des méthodes de quadratures d'ordre 0 et à poids positifs, alors la méthode de Runge–Kutta est bien définie pour $h_{\max}$ assez petit, elle est très stable, et elle est consistante. La méthode de Runge–Kutta est donc convergente (théorème 4.3.14). Il ne reste qu'à déterminer l'ordre de la méthode.

4.4.c Ordre

Proposition 4.4.11. On suppose que $f \in \mathscr{C}^1(I \times \mathbb{R})$. Sous l'hypothèse (4.3), la méthode de Runge–Kutta est consistante d'ordre 2 si et seulement si :

$$\sum_{i=1}^{q} b_i c_i = \frac{1}{2}.$$ ◀

Démonstration. Sous l'hypothèse (4.3), on a vu que la méthode était d'ordre 1 :

$$\forall (t,y) \in [t_0, t_0+T] \times \mathbb{R}, \quad \Phi(t,y,0) = f(t,y).$$

Montrons que de plus, la méthode est d'ordre 2 si et seulement si $\sum_i b_i c_i = \frac{1}{2}$. On donne la preuve dans le cas d'une méthode explicite[4], c'est-à-dire que la matrice des $a_{i,j}$ est triangulaire inférieure stricte. Puisque la méthode est explicite, on a :

$$\Phi(t,y,h) = \sum_{i=1}^{q} b_i p_i(t,y,h),$$

où les fonctions p_i sont définies par récurrence par :

$$\forall i \in [\![2,q]\!], \quad p_i(t,y,h) = f\left(t + c_i h, y + h\sum_{j=1}^{i-1} a_{i,j} p_j(t,y,h)\right),$$

et $p_1(t,y,h) = f(t,y)$. Par récurrence immédiate, les p_i admettent une dérivée partielle par rapport à h et les $\frac{\partial p_i}{\partial h}$ sont continues. De plus, on a :

$$\begin{aligned}\frac{\partial p_i}{\partial h}(t,y,h) = c_i \frac{\partial f}{\partial t}\left(t + c_i h, y + h\sum_{j=1}^{i-1} a_{i,j} p_j(t,y,h)\right) \\ + \frac{\partial f}{\partial y}\left(t + c_i h, y + h\sum_{j=1}^{i-1} a_{i,j} p_j(t,y,h)\right)\left(\sum_{j=1}^{i-1} a_{i,j} p_j(t,y,h) + h\sum_{j=1}^{i-1} a_{i,j}\frac{\partial p_j}{\partial h}(t,y,h)\right).\end{aligned}$$

4. pour une méthode implicite, la difficulté est de justifier que les fonctions p_i sont dérivables par rapport à h. Les outils nécessaires pour le faire seront vus au cours *espaces vectoriels normés et calcul différentiel*. Le reste de la preuve est identique.

Pour $h = 0$:

$$\begin{aligned}\frac{\partial p_i}{\partial h}(t,y,0) &= c_i \frac{\partial f}{\partial t}(t,y) + \frac{\partial f}{\partial y}(t,y) \sum_{j=1}^{i-1} a_{i,j} p_j(t,y,0) \\ &= c_i \frac{\partial f}{\partial t}(t,y) + \frac{\partial f}{\partial y}(t,y) \sum_{j=1}^{i-1} a_{i,j} f(t,y) \\ &= c_i \frac{\partial f}{\partial t}(t,y) + c_i \frac{\partial f}{\partial y}(t,y) f(t,y) \\ &= c_i f^{[1]}(t,y).\end{aligned}$$

Finalement :

$$\frac{\partial \Phi}{\partial h}(t,y,0) = \sum_{i=1}^{q} b_i \frac{\partial p}{\partial h}(t,y,0) = \left(\sum_{i=1}^{q} b_i c_i \right) f^{[1]}(t,y).$$

Par le théorème 4.3.16, la méthode est d'ordre 2 si et seulement si $\sum_i b_i c_i = \frac{1}{2}$. □

Avec des calculs similaires (mais très fastidieux) on peut montrer le résultat suivant.

Proposition 4.4.12. On se place sous l'hypothèse (4.3).

- Si $f \in \mathscr{C}^2(I \times \mathbb{R})$, la méthode est d'ordre 3 si et seulement si :

$$\sum_{i=1}^{q} b_i c_i = \frac{1}{2}, \quad \sum_{i=1}^{q} b_i c_i^2 = \frac{1}{3}, \quad \sum_{i,j=1}^{q} b_i a_{i,j} c_j = \frac{1}{6}.$$

- Si $f \in \mathscr{C}^3(I \times \mathbb{R})$, la méthode est d'ordre 4 si et seulement si elle satisfait les conditions pour être d'ordre 3, et de plus :

$$\sum_{i=1}^{q} b_i c_i^3 = \frac{1}{4}, \quad \sum_{i,j=1}^{q} b_i a_{i,j} c_j^2 = \frac{1}{12}, \quad \sum_{i,j=1}^{q} b_i c_i a_{i,j} c_j = \frac{1}{8}, \quad \sum_{i,j,k=1}^{q} b_i a_{i,j} a_{j,k} c_k = \frac{1}{24}. \quad ◀$$

On pourra vérifier que la méthode RK_4 classique satisfait toutes ces conditions, c'est donc une méthode d'ordre 4.

Index