Analyse numérique et équations différentielles ordinaires

MAIN 3
Polytech Sorbonne

Année 2023-2024

Fabien Vergnet

fabien.vergnet@sorbonne-universite.fr

Table des matières

Introduction

Ce cours traite de deux sujets des mathématiques appliquées qui sont l'analyse numérique et la résolution numérique des équations différentielles ordinaires. Néanmoins, à travers leur enseignement l'objectif de ce cours est également d'introduire la lectrice ou le lecteur au monde de la **modélisation mathématique** et de la **simulation numérique**, qui ont une importance majeure dans tous les domaines scientifiques et les applications industrielles. La modélisation mathématique est la science de représenter un phénomène physique, économique, biologique, etc. par des modèles abstraits qui permettent l'analyse et le calcul. La simulation numérique est quant à elle le procédé qui permet de calculer à l'aide d'un ordinateur les solutions de ces modèles et donc de simuler le phénomène étudié.

On rencontre souvent en mathématique des problèmes pour lesquels la solution ne peut pas être obtenue analytiquement. Néanmoins, il est souvent possible de mettre en œuvre des **méthodes numériques** fournissant une **approximation** de la solution, avec une **erreur** qu'il faut pouvoir **quantifier**. **L'analyse numérique** se définie alors comme l'étude mathématique de ces méthodes numériques.

C'est en particulier le cas pour les équations différentielles ordinaires (EDO). Pour rappel, considérons $d \in \mathbb{N}^\star$, $I = [0, T]$ un intervalle de $\mathbb{R}$, $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ une fonction continue et $\mu_0 \in \mathbb{R}^d$ une donnée initiale. Nous nous intéressons à la résolution du problème de Cauchy suivant : trouver une fonction u définie sur I et à valeur dans $\mathbb{R}^d$ telle que

$$\left\{ \begin{array}{rcl} u'(t) & = & f(t, u(t)), \quad t \in I = [0, T], \\ u(0) & = & \mu_0 \in \mathbb{R}^d. \end{array} \right. \tag{1}$$

Pour rappel, un problème de Cauchy est défini comme la donnée d'une EDO, (ou d'un système de plusieurs EDO) ainsi que d'une (ou plusieurs) condition(s) initiale(s). Dans le problème (1), nous considérons un système de d EDO d'ordre 1. Nous avons donc besoin d'une seule condition initiale qui s'écrit : $u(0) = \mu_0 \in \mathbb{R}^d$. Pour ce problème, le Théorème de Cauchy-Lipschitz, nous assure l'existence et l'unicité d'une solution globale, sous certaines hypothèses sur la fonction f.

Théorème (Cauchy-Lipschitz)**.** *Soit $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ une application continue. Si f est globalement lipschitizienne par rapport à la deuxième variable, i.e. si*

$$\exists L > 0, \quad \forall t \in I, \quad \forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d \quad \|f(t, x) - f(t, y)\| \leq L\|x - y\| \tag{2}$$

(pour une norme quelconque de $\mathbb{R}^d$, toutes les normes sur $\mathbb{R}^d$ étant équivalentes), alors pour toute condition initiale $\mu_0 \in \mathbb{R}^d$, il existe une unique solution u définie sur I et de classe $\mathcal{C}^1$ au problème de Cauchy (1)*.*

Si, de plus, f est de classe $\mathcal{C}^r(I \times \mathbb{R}^d; \mathbb{R}^d)$ avec $r \geq 1$, alors u est de classe $\mathcal{C}^{r+1}(I; \mathbb{R}^d)$.

Il est ensuite facile de généraliser ce résultat à des EDO d'ordre quelconque en se ramenant à un système d'EDO d'ordre 1, comme c'est le cas dans l'exemple ci-dessous.

Exemple (Équation du pendule)**.** L'évolution de l'angle entre un pendule soumis à la pesanteur, de longueur L, et la verticale est régie par le système d'équations dites du pendule,

$$\begin{cases} \theta''(t) + \frac{g}{L}\sin(\theta(t)) = 0, \quad t \geq 0, \\ \theta(0) = \theta_0 \in \mathbb{R}, \\ \theta'(0) = \theta_0' \in \mathbb{R}. \end{cases} \tag{3}$$

En posant

$$u(t) = \begin{pmatrix} \theta(t) \\ \theta'(t) \end{pmatrix} \qquad \text{et} \qquad f : \left(t, \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) \mapsto \begin{pmatrix} x_2 \\ -\frac{g}{L}\sin(x_1) \end{pmatrix}$$

l'équation (3) peut se mettre sous la forme d'un problème de Cauchy,

$$u'(t) = f(t, u(t)).$$

De plus, la fonction f est continue de $\mathbb{R} \times \mathbb{R}^2$ dans $\mathbb{R}^2$ et vérifie l'hypothèse (2) : pour tout $x = (x_1, x_2)^t \in \mathbb{R}^2$ et tout $y = (y_1, y_2)^t \in \mathbb{R}^2$

$$\|f(t,x) - f(t,y)\|_2 = \| \begin{pmatrix} x_2 - y_2 \\ \frac{g}{L}(\sin(y_1) - \sin(x_1)) \end{pmatrix} \|_2 \leq \max(1, \frac{g}{L})\|x - y\|_2,$$

donc il existe une unique solution u de classe $\mathcal{C}^1(\mathbb{R}^+; \mathbb{R}^2)$ à ce problème de Cauchy.

Remarque. C'est la donnée de la condition initiale qui implique l'unicité de la solution. Si on considère l'EDO seule, il peut exister une infinité de solutions. Cependant, il est possible de considérer d'autres conditions que les conditions initiales, ce sont les conditions aux limites. Pour fixer les idées, considérons une EDO scalaire d'ordre 2, $y''(t) = f(t, y(t)$, définie sur un intervalle réel $[0, T]$. On peut alors considérer, par exemple,

- les conditions aux limites de Dirichlet : $y(0) = \mu_1$ et $y(T) = \mu_2$,
- les conditions aux limites de Neumann : $y'(0) = \mu_1$ et $y'(T) = \mu_2$,
- une condition de Dirichlet d'un côté et de Neumann de l'autre.

L'étude des ces problèmes est d'une grande importance dans notre société car de nombreux phénomènes peuvent se modéliser sous la forme d'un système d'EDO muni de conditions initiales et/ou de conditions limites. C'est le cas de la plupart des problèmes en mécanique de point, mais cela concerne également l'étude de l'évolution d'une population au cours du temps, la modélisation de la propagation d'une épidémie, l'étude des intentions de vote d'une population et bien d'autres problèmes encore.

Comme annoncé plus haut, la résolution "à la main" de ces problèmes peut être compliquée. La question qui se pose maintenant est donc la suivante : **comment résoudre le problème de Cauchy** (1) **à l'aide d'un ordinateur ?** C'est pour répondre à cette question que nous allons introduire des méthodes numériques et analyser leurs propriétés. C'est cela l'analyse numérique : la construction et l'étude mathématique de méthodes numériques pour résoudre des problèmes mathématiques.

Reprenons le problème de Cauchy (1) et supposons qu'il existe une unique solution de régularité $\mathcal{C}^1$. Faisons alors un développement limité de la solution autour d'un point $t \in I$. Pour tout $h > 0$, il vient

$$u(t+h) = u(t) + hu'(t) + o(h) = u(t) + hf(t, u(t)) + o(h)$$

car, d'après l'EDO satisfaite par u, nous savons que $u'(t) = f(t, u(t))$ pout tout $t \in I$. Ainsi, nous avons écrit u au temps $t + h$ en fonction de $u(t)$ uniquement. Cette égalité est vraie

pour tout $t \in I$. En particulier, on peut calculer des valeurs approchées particulières de u de proche en proche,

$$\begin{array}{rcl} u(h) & \approx & u(0) + hf(h, u(0)) = \mu_0 + hf(h, \mu_0), \\ u(2h) & \approx & u(h) + hf(2h, u(h)), \\ & \vdots & \\ u((n+1)h) & \approx & u(nh) + hf(nh, u(nh)). \end{array}$$

C'est sur ce principe de récurrence que nous pouvons construire une approximation de la solution u. Soit $N > 0$, on découpe l'intervalle $I = [0, T]$ en N sous-intervalles de taille constante $h = T/N$. Nous construisons ensuite la suite (u_n) définie par

$$\left\{ \begin{array}{rcl} u_0 & = & \mu_0, \\ t_n & = & nh, \\ u_{n+1} & = & u_n + hf(t_n, u_n) \quad \forall n \in \{0, \ldots, N-1\}. \end{array} \right. \tag{4}$$

La suite (t_n) est une **discrétisation** de l'intervalle $[0, T]$. Le réel h est le **pas** de la discrétisation. La formule de récurrence (4) permettant d'exprimer u_{n+1} en fonction de u_n est appelé **schéma numérique**. C'est même un schéma numérique particulier, appelé **schéma d'Euler explicite**, dont nous étudierons au chapitre 3.

Remarque. Attention, il est important de ne pas confondre la solution exacte de l'EDO, notée u, avec son approximation numérique au temps t_n, notée u_n, qui dépend du pas de discrétisation h. En particulier, une question primordiale en analyse numérique (et qui nous intéressera tout au long du cours) est celle de la quantification de l'erreur d'approximation et de son comportement lorsque le pas h tend vers 0.

Définition. On appelle erreur de convergence d'un schéma numérique la quantité définie par $\max_{0 \leq n \leq N} \|u(t_n) - u_n\|$. Un schéma numérique est dit convergent si

$$\lim_{h \to 0} \max_{0 \leq n \leq N} \|u(t_n) - u_n\| = 0.$$

La convergence d'un schéma numérique assure que la solution approchée tend vers la solution exacte lorsque le pas de discrétisation h tend vers 0.

Dans ce cours, nous allons donc nous intéresser à la construction de schémas numériques pour l'approximation de solutions d'EDO et de l'étude de leur convergence. Pour cela, nous allons mettre au point un procédé systématique pour la définition de schémas numériques. Une autre façon de construire le schéma d'Euler explicite (4) peut être obtenue à partir de la formule,

$$u(t_{n+1}) = u(t_n) + \int_{t_n}^{t_{n+1}} u'(t) dt,$$

et en approchant le calcul de l'intégrale, c'est-à-dire l'aire sous la courbe $(t, u'(t))$, par un rectangle de hauteur $u'(t_n)$,

$$\int_{t_n}^{t_{n+1}} u'(t) dt \approx \int_{t_n}^{t_{n+1}} u'(t_n) dt = hu'(t_n) = hf(t_n, u(t_n)).$$

On retrouve bien l'approximation précédente

$$u(t_{n+1}) \approx u(t_n) + hf(t_n, u(t_n)).$$

Par cette technique, la construction d'un schéma numérique revient alors à un problème d'approximation du calcul d'une intégrale (appelé problème d'intégration numérique), dans

laquelle nous avons approché la fonction u' par une fonction constante (un polynôme de degré 0). Il est possible de généraliser cette approche en considérant des polynômes de degrés plus élevés et en utilisant d'autres formules d'intégration numérique.

Ce cours est donc divisé en trois chapitres correspondant aux problèmes de l'approximation polynomiale de fonctions, de l'intégration numérique et de l'approximation de solutions d'EDO.

Problème 1 : On cherche à reconstruire une fonction dont on ne connait les valeurs qu'en un certain nombre de points. Pour ce faire, on va déterminer le polynôme de plus petit degré qui passe par ces points. C'est ce qu'on appelle le **polynôme d'interpolation** de la fonction. Nous chercherons également à évaluer l'erreur commise lorsque l'on remplace la fonction par son polynôme d'interpolation. Ce problème est l'objet du chapitre 1.

Problème 2 : Le calcul de l'intégrale

$$\int_a^b f(x)dx$$

pour un fonction f intégrable sur un intervalle $[a, b]$ de $\mathbb{R}$ peut parfois être difficile si on ne connait pas explicitement une primitive de la fonction f. Les **méthodes de quadrature**, que nous verrons dans le chapitre 2, sont des méthodes numériques permettant d'approcher le calcul de ce type d'intégrales. Comme dans le chapitre 1, une attention particulière sera apporté à l'estimation de l'erreur commise avec cette approximation.

Problème 3 : Soit $d \in \mathbb{N}^\star$, $I = [T_1, T_2]$ un intervalle de $\mathbb{R}$ et $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ une fonction continue. On s'intéresse au problème de Cauchy suivant : trouver une fonction u définie sur $[T_1, T_2]$ et à valeurs dans $\mathbb{R}^d$ telle que

$$\begin{cases} u'(t) &= f(t, u(t)), \quad t \in I = [T_1, T_2], \\ u(t = T_1) &= \mu_0 \in \mathbb{R}^d. \end{cases}$$

Encore une fois, la résolution analytique d'un tel problème n'est pas toujours évidente mais il existe des méthodes numériques pour approcher la solution u de cette **équation différentielle ordinaire (EDO)**. Ceci sera l'objet du chapitre 3.

Références : Ce cours a été rédigé avec le recours aux ouvrages suivants, dont il n'est pas explicitement fait mention dans la suite ; les lecteurs et lectrices pourront s'y rapporter pour plonger plus avant dans les démonstrations des résultats présentés ici :

1. *Analyse numérique et équations différentielles.* Jean-Pierre Demailly. Collection Sciences Grenoble.
2. *Analyse numérique - algorithme et étude mathématique. Cours et exercices corrigés.* Francis Filbet. Dunod.
3. *Analyse numérique. Exercices et problèmes corrigés.* Bernard Héron, Françoise Issard-Roch, Colette Picard. Dunod.
4. *Analyse numérique : Une approche mathématique.* Michèle Schatzman. Dunod.

Chapitre 1

Approximation de fonctions

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

La théorie de l'approximation étudie de façon rigoureuse le compromis entre la complexité donnée par le nombre de paramètres n et la précision que l'on peut obtenir entre f et g. Elle s'intéresse aussi à la manière dont on construit en pratique l'approximation g à partir de f.

Dans ce chapitre nous allons nous intéresser à deux méthodes d'approximation couramment utilisées en analyse numérique : l'interpolation polynomiale et la méthode des moindres carrés.

On ne considèrera ici uniquement que des fonctions d'une seule variable réelle, la généralisation à l'approximation des fonctions à plusieurs variables dépassant le cadre de ce cours.

1.1 Interpolation polynomiale

Dans cette partie, nous cherchons à approcher une fonction régulière par une fonction polynomiale : c'est la théorie de l'interpolation polynomiale. En calcul scientifique, nous avons rarement accès à une infinité de points pour évaluer une fonction donnée. Au contraire, nous ne connaissons souvent cette fonction qu'au travers d'un jeux de données qui représente les valeurs de cette fonction en un certain nombre de points. L'interpolation (et en particulier l'interpolation polynomiale qui nous intéresse dans ce chapitre) permet alors de représenter, de façon approchée, une fonction dont on ne connait pas une expression explicite. En analyse numérique, comme nous allons le voir, cette méthode est à la base de nombreuses techniques pour le calcul scientifique.

Nous nous intéressons dans cette section principalement à deux problématiques : comment construire en pratique un polynôme d'interpolation ? Comment estimer et améliorer la qualité de cette approche ?

1.1.1 Notations et pré-requis

Dans toute cette partie, nous notons $\mathcal{P}_n(\mathbb{R})$ **l'espace vectoriel des fonctions polynômiales sur $\mathbb{R}$, à coefficients réels, de degrés inférieur ou égal à n.**

Soit a, b deux réels. L'espace des fonctions continues sur l'intervalle $[a, b]$ à valeurs réelles est noté $\mathcal{C}([a, b]; \mathbb{R})$. Sur cet espace, on définit les normes L^p, pour tout entier naturel $p \geq 1$, par

$$\|f\|_{p,[a,b]} = \left(\int_a^b |f(x)|^p dx \right)^{\frac{1}{p}}.$$

Pour $p = \infty$, on définit la norme infinie (ou *norme sup* ou *norme de la convergence uniforme*) par

$$\|f\|_{\infty,[a,b]} = \sup_{x \in [a,b]} |f(x)|.$$

Pour $p \in \mathbb{N}^* \cup \{\infty\}$, si $\|f\|_{p,[a,b]} < \infty$, on dira que $f \in L^p([a, b]; \mathbb{R})$.

Pour certaines démonstrations, nous aurons besoin d'utiliser un des théorèmes de base de l'analyse réelle : le Théorème de Rolle, dont nous rappelons l'énoncé ci-dessous.

Théorème 1.1. (Théorème de Rolle) *Soit a et b deux réels, $a < b$, et f une fonction continue sur $[a, b]$ à valeurs réelles, dérivable sur $]a, b[$, telle que $f(a) = f(b)$. Alors il existe (au moins) un réel $c \in]a, b[$ vérifiant $f'(c) = 0$.*

De même, nous utiliserons le théorème d'algèbre linéaire fondamental suivant, conséquence du théorème du rang.

Théorème 1.2. (Théorème fondamental d'algèbre linéaire) *Soit ϕ une application linéaire entre deux espaces vectoriels E et F de dimensions finies. Si $\dim(E) = \dim(F)$ et si ϕ est injective, alors ϕ est bijective.*

1.1.2 Polynômes de Lagrange

Soit $f : [a, b] \to \mathbb{R}$ connue uniquement en $n + 1$ points distincts ξ_1, ξ_2, ..., ξ_{n+1} de l'intervalle $[a, b]$. Il s'agit de construire un polynôme p_n de dégré inférieur ou égal à n tel que

$$\boxed{p_n(\xi_i) = f(\xi_i) \qquad \forall i \in 1, 2, \ldots, n + 1.} \tag{1.1}$$

On dit alors que le polynôme p_n est un polynôme d'interpolation de la fonction f aux points $(\xi_i)_{1 \leq i \leq n+1}$.

Existence et unicité

Nous commençons par montrer qu'un tel polynôme existe et qu'il est unique.

Théorème 1.3. *Il existe un et un seul élément de $\mathcal{P}_n(\mathbb{R})$ vérifiant* (1.1)*. On l'appelle polynôme d'interpolation de Lagrange associée aux points $(\xi_i, f(\xi_i))_{i \in \{1,\ldots,n+1\}}$*

Démonstration. Considérons tout d'abord l'application

$$\phi_n : \begin{array}{rl} \mathcal{P}_n \mapsto & \mathbb{R}^{n+1} \\ p \mapsto & (p(\xi_1), \ldots, p(\xi_{n+1})). \end{array}$$

Pour démontrer l'existence est l'unicité de ce polynôme d'interpolation, il suffit de montrer que l'application ϕ_n est bijective. En effet, on aura alors par bijectivité de l'application ϕ_n que pour tout n+1-uplet $(f(\xi_1), \ldots, f(\xi_{n+1}))$ de $\mathbb{R}^{n+1}$ il existe un unique polynôme p dans $\mathcal{P}_n$ vérifiant (1.1).

Montrons que l'application ϕ_n est linéaire. Soient p et q deux éléments de $\mathcal{P}_n$ et et $\lambda \in \mathbb{R}$.

$$\begin{aligned} \phi_n(p + \lambda q) &= ((p + \lambda q)(\xi_1), \ldots, (p + \lambda q)(\xi_n)), \\ &= (p(\xi_1) + \lambda q(\xi_1), \ldots, p(\xi_n) + \lambda q(\xi_n)), \\ &= \phi_n(p) + \lambda \phi_n(q). \end{aligned}$$

Donc ϕ_n est linéaire.

L'application ϕ_n est également injective car si $\phi_n(p_n) = (0, \dots, 0)$ avec $p_n \in \mathcal{P}_n$ on en déduit que p_n est un polynome de degré inférieur ou égal à n possèdant $n+1$ racines ce qui implique que $p_n = 0$.

Enfin, $\dim(\mathcal{P}_n) = n+1 = \dim(\mathbb{R}^{n+1})$, ce qui permet de conclure à l'aide du théorème 1.2 que l'application ϕ_n est bijective (c'est un isomorphisme entre deux espaces vectoriels de même dimension).

□

Définition 1.4. On appelle polynôme de base de Lagrange associé au point ξ_i l'élément l_i de $\mathcal{P}_n(\mathbb{R})$ définie par

$$\boxed{l_i(x) = \prod_{k=1,\, k\neq i}^{n+1} \frac{(x-\xi_k)}{(\xi_i - \xi_k)}.}$$

En particulier ces polynômes vérifient la propriété suivante :

$$l_i(\xi_j) = \delta_{i,j} = \begin{cases} 0 & \text{si } j \neq i, \\ 1 & \text{si } j = i. \end{cases}$$

Proposition 1.5. *Le polynôme d'interpolation de Lagrange de f aux points ξ_1, …, ξ_{n+1} s'écrit*

$$\boxed{p_n(x) = \sum_{i=1}^{n+1} f(\xi_i) l_i(x).} \tag{1.2}$$

L'expression (1.2) *s'appelle la forme de Lagrange du polynôme d'interpolation p_n.*

Démonstration. Il est facile de remarque que pour tout $i \in \{1, \dots, n+1\}$, l_i est un polynôme de degré n et que

$$\sum_{i=1}^{n+1} f(\xi_i) l_i(\xi_i) = f(\xi_i).$$

Or, d'après le théorème d'existence et d'unicité précédent, il existe un unique polynôme de degré n vérifiant cette égalité pour tout i. Ce qui démontre le résultat. □

Remarque 1.6. La famille $(l_i)_{1\leq i\leq n+1}$ est une base de $\mathcal{P}_n$.

Exercice 1.

Soient ξ_1 et ξ_2 des éléments de $\mathbb{R}$ et f une fonction définie dans $\mathbb{R}$ à valeurs dans $\mathbb{R}$. Tracer les polynômes de base de Lagrange associés aux points ξ_1 et ξ_2 et donner l'expression du polynôme d'interpolation de Lagrange de f aux points ξ_1 et ξ_2.

Remarque 1.7. La forme de Lagrange du polynôme d'interpolation est peu intéressante d'un point de vue pratique. Elle a en effet un caractère relativement peu algorithmique et son évaluation requiert trop d'opérations élémentaires. Dans la pratique, nous préférerons donc la formule de Newton, introduite dans la prochaine section.

Forme de Newton

La forme de Newton d'un polynôme d'interpolation de Lagrange aux points $\xi_1, \dots \xi_n$ consiste à écrire le polynôme p_n sous la forme

$$p_n(x) = a_0 + a_1(x - \xi_1) + \cdots + a_n \prod_{i=1}^{n} (x - \xi_i),$$

où les coefficients $(a_i)_{0 \leq i \leq n}$ sont des réels à déterminer.

Observons d'abord que tout polynôme de degré inférieur ou égal à n peut s'écrire sous cette forme du moment que les points ξ_i sont tous distincts. Ensuite, cette formule présente un intérêt puisque elle fournit naturellement une récurrence. En effet, la partie tronquée $a_0 + \cdots + a_{n-1} \prod_{i=1}^{n-1} (x - \xi_i)$ n'est rien d'autre que le polynôme d'interpolation p_{n-1} écrit aux points $\xi_1, \dots, \xi_{n-1}$. En effet, p_{n-1} est un polynôme de degré inférieur ou égal à $n-1$ et tel que pour tout $i \in \{1, \dots, n\}$, $p_{n-1}(\xi_i) = f(\xi_i)$.

Ainsi connaissant p_{n-1}, le calcul de p_n s'effectue en déterminant le coefficient a_n assurant que $p_n(\xi_n) = f(\xi_n)$. Les coefficients $(a_i)_{0 \leq i \leq n}$ sont donnés par la formule de Newton suivante.

Théorème 1.8. *Le polynôme d'interpolation de Lagrange d'une fonction f aux $n+1$ points distincts $\xi_1, \dots, \xi_{n+1}$ est donné par*

$$\boxed{p_n(x) = f[\xi_1] + f[\xi_1, \xi_2](x - \xi_1) + \sum_{i=3}^{n+1} f[\xi_1, \dots, \xi_i] \prod_{j=1}^{i-1} (x - \xi_j),} \tag{1.3}$$

où $f[\,]$ désigne les différences divisées de f définies par

$$\begin{aligned}
f[\xi_i] &= f(\xi_i) \\
f[\xi_i, \xi_j] &= \frac{f[\xi_j] - f[\xi_i]}{\xi_j - \xi_i} \\
f[\xi_i, \xi_j, \xi_k] &= \frac{f[\xi_j, \xi_k] - f[\xi_i, \xi_j]}{\xi_k - \xi_i} \\
&\vdots \\
f[\xi_1, \dots, \xi_{n+1}] &= \frac{f[\xi_2, \dots, \xi_{n+1}] - f[\xi_1, \xi_n]}{\xi_{n+1} - \xi_1}
\end{aligned}$$

Démonstration. Si $n = 0$, alors pour tout x, $p_0(x) = f(\xi_1) = f[\xi_1]$ et l'égalité (1.3) est vérifiée.

Soit $n \in \mathbb{N}^*$. On note p_{n-1} le polynôme d'interpolation de Lagrange de f en $\xi_1, \dots, \xi_n$. Comme $p_n - p_{n-1} \in \mathcal{P}_n$ et s'annule en $\xi_1, \dots, \xi_n$, alors $p_n(x) - p_{n-1}(x)$ peut s'écrire

$$p_n(x) - p_{n-1}(x) = \tau_n \prod_{j=1}^{n} (x - \xi_j),$$

où τ_n est le coefficient dominant de p_n. Montrons par récurrence sur le nombre de points d'interpolation que $\tau_n = f[\xi_1, \dots, \xi_{n+1}]$.

- $n = 1 : p_1(x) = f(\xi_1) + \frac{f(\xi_2) - f(\xi_1)}{\xi_2 - \xi_1}(x - \xi_1)$, d'où $\tau_1 = \frac{f(\xi_2) - f(\xi_1)}{\xi_2 - \xi_1} = f[\xi_1, \xi_2]$
- Supposons l'hypothèse de récurrence vraie pour tout $k \leq n-1$ points. En particulier on a $\tau_k = f[\xi_1, \dots, \xi_k]$ pour tout $k \leq n-1$. On pose

$$q_n(x) = \frac{(x - \xi_1) r_{n-1}(x) - (x - \xi_{n+1}) p_{n-1}(x)}{\xi_{n+1} - \xi_1}$$

où r_{n-1} est le polynôme d'interpolation de f en $\xi_2, \ldots, \xi_{n+1}$. On remarque que $q_n \in \mathcal{P}_n$, $q_n(\xi_1) = f(\xi_1)$ et $q_n(\xi_{n+1}) = f(\xi_{n+1})$, et

$$q_n(\xi_i) = \frac{(\xi_i - \xi_1 - \xi_i + \xi_{n+1})}{\xi_{n+1} - \xi_1} f(\xi_i) = f(\xi_i) \qquad \text{pour} \quad 2 \leq i \leq n.$$

Ainsi par unicité du polynôme d'interpolation de Lagrange, on en déduit que $q_n = p_n$. Or, d'après l'hypothèse de récurrence appliquée à r_{n-1} et p_{n-1} on voit que le coefficient dominant de q_n est

$$\frac{f[\xi_2, \ldots, \xi_{n+1}] - f[\xi_1, \ldots, \xi_n]}{\xi_{n+1} - \xi_1} = f[\xi_1, \ldots, \xi_{n+1}].$$

□

Calcul pratique des termes $f[\xi_1, \ldots, \xi_k]$
En pratique, pour calculer un terme de différence divisée d'ordre i, on a besoin de nombreux termes de différence divisée d'ordres inférieurs. Le tableau suivant schématise le calcul par induction de ces termes :

	Etape 1	Etape 2	Etape 3	...	Etape $n+1$
d_1	$f[\xi_1]$				
d_2	$f[\xi_2]$	$f[\xi_1, \xi_2]$			
d_3	$f[\xi_3]$	$f[\xi_2, \xi_3]$	$f[\xi_1, \xi_2, \xi_3]$		
	$\vdots$	$\vdots$	$\vdots$		
d_{n+1}	$f[\xi_{n+1}]$	$f[\xi_n, \xi_{n+1}]$	$f[\xi_{n-1}, \xi_n, \xi_{n+1}]$	...	$f[\xi_1, \ldots, \xi_{n+1}]$

Exercice 2.
Étant donné trois points $\{(0,1), (2,5), (4,17)\}$, déterminer le polynôme d'interpolation de Newton de degré 2 passant par ces points.

Numériquement, cette méthode de calcul des termes $d_i = f[\xi_1, \ldots, \xi_i]$ peut être implémentée simplement et on obtient l'algorithme (en pseudo langage) suivant très efficace :

- *pour i de* 1 *à* $n+1$ *:* $d_i := f(\xi_i)$,
- *pour k de* 2 *à* $n+1$ *:*
  - ⋆ *pour i de n+1 à k :* $d_i \leftarrow \dfrac{d_i - d_{i-1}}{\xi_i - \xi_{i-k+1}}$.

Algorithme de calcul de $p_n(x)$
Après avoir calculé les termes d_i, on peut mettre en œuvre le calcul de l'évalutation de p_n en un réel x en utilisant l'algorithme de Hörner. Celui-ci s'utilise en écrivant le polynôme sous la forme

$$p_n(x) = d_1 + (x - \xi_1)\left[d_2 + (x - \xi_2)\left(d_3 + \cdots + (x - \xi_n)d_{n+1}\right)\right].$$

On obtient l'algorithme suivant.

- $p := d_{n+1}$,
- *pour k de n à* 1 *:* $p \leftarrow d_k + (x - \xi_k)p$.

1.1.3 Erreur d'interpolation

Théorème 1.9. *Soit f de classe $\mathcal{C}^{n+1}$ sur $[a,b]$, et p_n son polynôme d'interpolation de Lagrange en $n+1$ points distincts de $[a,b]$, notés $\xi_1,\dots,\xi_{n+1}$. Alors*

$$\boxed{\forall x \in [a,b], \qquad \exists\, y_x \in]a,b[, \qquad E(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(y_x)}{(n+1)!}\prod_{i=1}^{n+1}(x-\xi_i)} \tag{1.4}$$

La quantité $E(x)$ est appellée erreur d'interpolation de Lagrange au point x.

Pour montrer le Théorème 1.9 on utilise le Lemme suivant (dont la preuve s'obtient facilement en appliquant m fois le théorème de Rolle).

Lemme 1.10. *Soit g une fonction de classe $\mathcal{C}^m$ sur un intervalle $I \subset \mathbb{R}$, qui s'annule en $m+1$ points de I. Alors il existe $z \in \mathring{I}$ tel que $g^m(z) = 0$.*

Démonstration. Soit $\xi_1,\dots,\xi_{n+1}$ les points d'interpolation et $x \in [a,b]$ fixé.

Si $x \in \{\xi_1,\dots,\xi_{n+1}\}$, alors l'égalité est vraie.

Supposons alors que $x \notin \{\xi_1,\dots,\xi_{n+1}\}$. On introduit la fonction g définie par

$$g(t) = f(t) - p_n(t) - \mu\prod_{i=1}^{n+1}(t-\xi_i),$$

où μ est une constante que l'on fixera ultérieurement. Comme $p_n(\xi_i) = f(\xi_i)$ pour tout $i \in \{1,\dots,n+1\}$, la fonction g s'annule aux $n+1$ points $\xi_1,\dots,\xi_{n+1}$. On choisit μ de sorte que g s'annule aussi en x, ce qui amène à prendre

$$\mu = \frac{f(x) - p_n(x)}{\prod_{i=1}^{n+1}(x-\xi_i)}$$

D'après le Lemme 1.10, il existe alors y_x vérifiant $g^{(n+1)}(y_x) = 0$, ce qui s'écrit, comme $p_n \in Pn$,

$$f^{(n+1)}(y_x) - 0 - (n+1)!\mu = 0,$$

On en déduit alors l'égalité (1.4). □

Corollaire 1.11. *On peut obtenir une majoration de l'erreur en x par*

$$|f(x) - p_n(x)| \le \frac{M_{n+1}}{(n+1)!}\,|N_{n+1}(x)|\,,$$

en notant

$$M_{n+1} = \|f^{(n+1)}\|_{\infty,[a,b]} = \sup_{x\in[a,b]} |f^{(n+1)}(x)|$$

et

$$N_{n+1}(x) = \prod_{i=1}^{n+1}(x-\xi_i).$$

Une majoration uniforme de l'erreur est également obtenue par

$$\boxed{\|f - p_n\|_{\infty,[a,b]} \le \frac{M_{n+1}}{(n+1)!}\|N_{n+1}\|_{\infty,[a,b]}.}$$

On peut alors se poser les questions suivantes : l'erreur $E(x)$ diminue-t-elle lorsque le nombre de points augmente ? Tend-elle vers 0 lorsque $n \to \infty$? A-t-on convergence uniforme ? Le coefficient $\frac{1}{(n+1)!}$ est plutôt favorable à la convergence, sous réserve que M_{n+1} n'explose pas trop vite lorsque n augmente (ce qui dépend de la fonction interpolée), et que $\|N_{n+1}\|_{\infty,[a,b]}$ non plus (ce qui dépend seulement du choix des points d'interpolation).

1.1.4 Convergence uniforme

Un exemple de convergence

Pour $f(x) = e^x$, on a aisément une majoration de $f^{(n+1)}$ sur l'intervalle $[a, b]$:

$$M_{n+1} = e^b.$$

On obtient alors

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

ce qui prouve que $\|f - p_n\|_{\infty,[a,b]} \to 0$ lorsque $n \to +\infty$ (le terme $\frac{\alpha^{n+1}}{(n+1)!}$ étant le terme général d'une série convergente, quelque soit $\alpha \in \mathbb{R}$, il tend donc vers 0 lorsque $n \to \infty$). Autrement dit, le polynôme d'interpolation de f converge uniformément vers f lorsque $n \to \infty$, **quelque soit le choix des points d'interpolation**.

Plus généralement, ce résultat est vrai pour la classe de fonctions suivante.

Théorème 1.12. *Soit $f \in \mathcal{C}^\infty([a, b])$ dont les dérivées n-ème sont bornées **uniformément par rapport à** n. Alors, le polynôme d'interpolation de f converge uniformément vers lorsque n tend vers $+\infty$.*

Démonstration. Il suffit de remarquer qu'il existe $C > 0$ telle que $\|f^{(n)}\|_{\infty,[a,b]} \leq C$, pour tout $n \in \mathbb{N}$. Ainsi on a

$$\|f - p_n\|_{\infty,[a,b]} \leq C \frac{(b-a)^{n+1}}{(n+1)!} \underset{n\to\infty}{\to} 0.$$

□

Un résultat de non convergence

Théorème 1.13 (résultat négatif dans $\mathcal{C}^0$)**.** *Pour toute famille de points d'interpolation $(\xi_i)_{1\leq i\leq n}$, il existe une fonction f continue sur $[a, b]$ telle que la suite des polynômes d'interpolation aux points $(\xi_i)_{1\leq i\leq n}$ ne converge pas uniformément vers f lorsque n tend vers $+\infty$.*

Le phénomène de Runge

Définition 1.14. Le phénomène de Runge est un phénomène de non-convergence uniforme que l'on peut observer lorsque l'on considère des points d'interpolations équidistants, même pour des fonctions $\mathcal{C}^\infty([a, b]; \mathbb{R})$.

On présente ici un exemple de fonction pour laquelle la suite des polynômes d'interpolation pour des points équidistants n'est pas convergente. On considère la fonction

$$f(x) = \frac{1}{1+x^2}$$

et son polynôme d'interpolation aux points $\xi_i = -5 + (i-1)\frac{10}{n}$, $1 \leq i \leq n+1$ équidistants sur $[-5, 5]$. La figure 1.1 montre les fortes oscillations qui apparaissent au bord du domaine lorsque n augmente.

1.1.5 Choix des points d'interpolation

Dans cet section, on se place sur un intervalle $[a, b]$ et on examine l'influence du choix des points d'interpolation sur la majoration que l'on peut obtenir pour $\|N_{n+1}\|_{\infty,[a,b]}$.

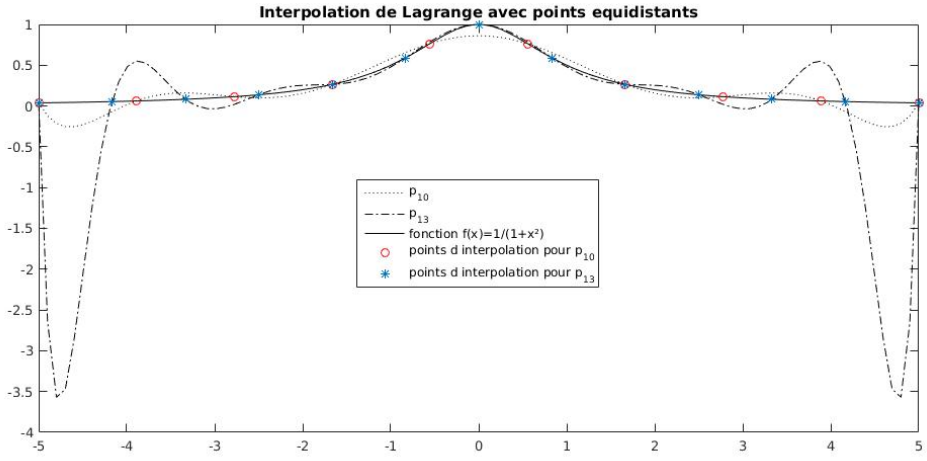

Figure 1.1 – Phénomène de Runge

Points quelconques Une première estimation grossière peut être obtenue lorsqu'on ne suppose aucune répartition particulière des points d'interpolation $(\xi_i)_{1\leq i\leq n+1}$. En effet, pout tout i dans $\{1,\ldots,n+1\}$ et pour tout x dans l'intervalle $[a,b]$,

$$|x-\xi_i|\leq b-a,$$

ce qui conduit à l'estimation

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

En passant au sup on a alors,

$$\|N_{n+1}\|_{\infty,[a,b]}\leq (b-a)^{n+1}.$$

Points équidistants On prend ici $\xi_i=a+(i-1)h$, pour $1\leq i\leq n+1$, avec $h=\frac{b-a}{n}$. On peut alors montrer qu'on obtient avec ce choix

$$\|N_{n+1}\|_{\infty,[a,b]}\leq \frac{C}{\sqrt{n}}\left(\frac{b-a}{e}\right)^{n+1}$$

où C est une constante indépendante de n.

En effet, notons $s=\frac{x-a}{h}\in[0,n]$. Pour tout $x\in[a,b]$, on a

$$\prod_{i=1}^{n+1}(x-\xi_i)=\prod_{i=1}^{n+1}h(s-(i-1))=h^{n+1}\prod_{i=0}^{n}(s-i).$$

Par récurrence, montrons alors que, pour tout $n\in\mathbb{N}$,

$$\prod_{i=0}^{n}(s-i)\leq n!,\quad \forall s\in[0,n].$$

En effet, pour $n=0$, on a bien $s\leq 1$, pour tout $s\in[0,1]$. Par suite, supposons que l'assertion est vraie pour un certain $n\in\mathbb{N}$ et montrons qu'elle reste vraie au rang $n+1$. On a, pour

tout $s \in [0, n+1]$,

$$\begin{array}{rcl} \prod_{i=0}^{n+1}(s-i) - (n+1)! & = & \left(\prod_{i=0}^{n}(s-i)\right)(s-n-1) - (n+1)!, \\ & \leq & (s-n-1)n! - (n+1)n!, \\ & \leq & (s-2(n+1))n!, \\ & \leq & 0, \quad \text{car } s \in [0, n+1]. \end{array}$$

D'où

$$|N_{n+1}(x)| \leq h^{n+1} n!.$$

En utilisant alors la formule de Stirling, on a que $n! \leq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$, ce qui implique que

$$|N_{n+1}(x)| \leq \left(\frac{b-a}{n}\right)^{n+1}\left(\frac{n}{e}\right)^n \sqrt{2\pi n} \leq \frac{\sqrt{2\pi}e}{\sqrt{n}}\left(\frac{b-a}{e}\right)^{n+1}.$$

D'où le résultat annoncé.

Ainsi, si $b-a < e$, $\|N_{n+1}\|_{\infty,[a,b]}$ converge vers 0 quand $n \to \infty$. Par contre, on ne peut rien conclure si $b-a > e$.

Points de Tchebychev

Définition 1.15. Soit $n \in \mathbb{N}$. Les $n+1$ points d'interpolation de Tchebychev sur un intervalle $[a,b]$ de $\mathbb{R}$ sont les points $(\xi_i)_{i\in\{1,\ldots,n+1\}}$ définis par

$$\boxed{\xi_i = \frac{a+b}{2} + \frac{b-a}{2}\cos\left(\frac{(2i-1)\pi}{2(n+1)}\right).}$$

Pour $[a,b] = [-1,1]$, les points de Tchebychev sont également définis comme les racines du polynôme

$$T_{n+1}(x) = \cos((n+1)\arccos x),$$

qu'on appelle le $(n+1)$-ième *polynôme de Tchebychev de première espèce.*

Nous allons montrer que les points de Tchebychev sont les points qui minimisent la norme infinie $\|N_{n+1}\|_{\infty,[a,b]}$. Commençons par énoncer un lemme technique.

Lemme 1.16. *Pour tout $x \in [-1,1]$ et tout entier $n \geq 0$,*

$$\cos((n+1)\arccos(x)) = 2^n \prod_{i=1}^{n+1}\left(x - \cos\left(\frac{(2i-1)\pi}{2(n+1)}\right)\right).$$

Démonstration. Commençons par vérifier que pour tout $n \geq 0$ et tout $t \in \mathbb{R}$, il existe un polynôme $p_n \in \mathcal{P}_n$ tel que

$$\cos((n+1)t) = 2^n(\cos(t))^{n+1} + p_n(\cos(t)),$$

c'est-à-dire que $\cos((n+1)t)$ est un polynôme en $\cos(t)$ de degré $n+1$ et de coefficient dominant 2^n. On fait une démonstration par récurrence.

La relation est vraie pour $n=0$ avec $p_0 = 0$ et pour $n=1$ avec $p_1 = -1$.

Soit $m \geq 2$, on suppose que la relation est vraie pour tout $n \in \{0,\ldots,m\}$. De la formule trigonométrique

$$\cos((m+1)t) + \cos((m-1)t) = 2\cos(t)\cos(mt),$$

on déduit que

$$\begin{aligned}\cos((m+1)t) &= 2\cos(t)\cos(mt) - \cos((m-1)t),\\ &= 2\cos(t)[2^{m-1}(\cos(t))^m + p_{m-1}(\cos(t))] - [2^{m-2}(\cos(t))^{m-1} + p_{m-2}(\cos(t))],\\ &= 2^m(\cos(t))^{m+1} + p_m(\cos(t)),\end{aligned}$$

en notant

$$p_m(x) = 2xp_{m-1}(x) - 2^{m-2}x^{m-2}p_{m-2}(x),$$

qui est bien un polynôme de degré 2 à coefficients réels. Cela conclut la preuve par récurrence.

Par suite, on en déduit que pour tout $x \in [-1,1]$

$$\cos((n+1)\arccos(x)) = 2^n x^{n+1} + p_n(x),$$

ce qui signifie que $\cos(n\arccos(x))$ est bien un polynôme de degré $n+1$, de coefficient dominant 2^n. De plus, $\cos(n\arccos(x))$ a les mêmes racines que le polynôme

$$2^n \prod_{i=1}^{n+1} \left(x - \cos\left(\frac{(2i-1)\pi}{2(n+1)}\right)\right)$$

donc ces deux polynômes sont bien égaux. □

Théorème 1.17. *Pour tous points d'interpolation $(\xi_i)_{i=1,\dots,n+1}$ distincts dans $[-1,1]$, on a*

$$\|N_{n+1}\|_{\infty,[-1,1]} \geq \frac{1}{2^n},$$

avec égalité si, pour tout i, $\xi_i = \cos\left(\frac{(2i-1)\pi}{2(n+1)}\right)$, c'est-à-dire si les ξ_i sont les racines du n-ième polynôme de Tchebychev de première espèce.

Démonstration. Posons

$$P(x) = \prod_{i=1}^{n+1} \left(x - \cos\left(\frac{(2i-1)\pi}{2(n+1)}\right)\right) = \frac{1}{2^n}\cos((n+1)\arccos(x)).$$

On obtient directement que

$$\|P\|_{\infty,[-1,1]} = \frac{1}{2^n}.$$

Supposons alors qu'il existe une famille de points d'interpolation $(\xi_i)_{i=1,\dots,n+1}$ telle que

$$\|N_{n+1}\|_{\infty,[-1,1]} < \|P\|_{\infty,[-1,1]},$$

c'est-à-dire

$$-\frac{1}{2^n} < \|N_{n+1}\|_{\infty,[-1,1]} < \frac{1}{2^n}.$$

Pour $k \in \{0,\dots,n+1\}$, les points

$$x_k = \cos\left(\frac{k}{n+1}\pi\right)$$

réalisent le maximum de $|P|$ sur $[-1,1]$, car

$$P(x_k) = \frac{1}{2^n}\cos((n+1)\arccos(x_k)) = \frac{1}{2^n}\cos(k\pi) = (-1)^k\|P\|_{\infty,[-1,1]}.$$

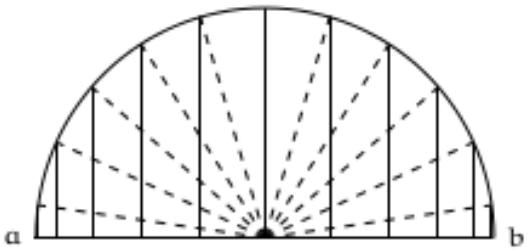

Figure 1.2 – Représentation des points de Tchebychev.

Ainsi, pour tout $k \in \{0, \dots, n+1\}$, on a

$$(P - N_{n+1})(x_k) = (-1)^k \|P\|_{\infty,[-1,1]} - N_{n+1}(x_k)$$

. Si k est pair, on obtient

$$(P - N_{n+1})(x_k) = \|P\|_{\infty,[-1,1]} - N_{n+1}(x_k) > \frac{1}{2^n} - \frac{1}{2^n} = 0,$$

et si k est impair, on obtient

$$(P - N_{n+1})(x_k) = -\|P\|_{\infty,[-1,1]} - N_{n+1}(x_k) < -\frac{1}{2^n} + \frac{1}{2^n} = 0.$$

Finalement, on a, pour tout $k \in \{0, \dots, n\}$,

$$(P - N_{n+1})(x_k)(P - N_{n+1})(x_{k+1}) < 0.$$

De plus, les polynômes P et N_{n+1} sont unitaires et de degré $n+1$. Cela signifie que le polynôme $P - N_{n+1}$ est de degré n et change $n+1$ fois de signe. Il admet donc $n+1$ racines et c'est forcément le polynôme nul. On conclut donc que $P = N_{n+1}$, ce qui contredit l'hypothèse précédente. □

Théorème 1.18 (Convergence uniforme)**.** *Si $f \in \mathcal{C}^1([a,b])$ alors la suite $(p_n)_{n\geq 0}$ des polynômes d'interpolation de f aux points de Tchebychev converge uniformément vers f sur $[a,b]$.*

Remarque 1.19. Les points d'interpolation de Tchebychev sont répartis symétriquement autour du milieu de l'intervalle $[a,b]$, de façon plus dense près des bords a et b (voir Figure 1.2).

1.1.6 Interpolation par morceaux

Afin d'éviter le type d'instabilité que l'on peut observer avec le phénomène de Runge, on peut, plutôt que d'augmenter le nombre de points d'interpolation, procéder à de l'interpolation par morceaux sur des intervalles de plus petite taille. On considère pour cela une subdivision de l'intervalle $[a,b]$ définie par $a = \alpha_1 < \alpha_1 < \cdots < \alpha_{m+1} = b$, de pas h_α :

$$h_\alpha = \max_{1\leq i\leq m} |\alpha_{i+1} - \alpha_i|$$

On réalise alors une interpolation de Lagrange de degrés n de la fonction f sur chacun des intervalles $[\alpha_i, \alpha_{i+1}]$, pour $i = 1, \dots, m$, en choisissant des points d'interpolation $(\xi_j^i)_{1\leq j\leq n+1}$ qui vérifient

$$\xi_1^i = \alpha_i, \qquad \xi_{n+1}^i = \alpha_{i+1}. \tag{1.5}$$

Cela définit alors un polynôme p_n^i, et on pose

$$p_{n,\alpha}(x) = p_n^i(x) \qquad \text{pour } x \in [\alpha_i, \alpha_{i+1}].$$

On remarque, notamment à cause de la condition (1.5), que la fonction $p_{n,\alpha}$ obtenue est continue (mais pas nécessairement dérivable). En utilisant les résultats précédents, on obtient :

Proposition 1.20. *On suppose f de classe $\mathcal{C}^{n+1}([a,b])$. La fonction polynômiale par morceaux $p_{n,\alpha}$ vérifie*

$$\boxed{\|f - p_{n,\alpha}\|_{\infty,[a,b]} \leq \frac{M_{n+1}}{(n+1)!} h_\alpha^{n+1}.}$$

En particulier, pour un entier n fixé, $p_{n,\alpha}$ converge uniformément vers f lorsque h_α tend vers 0.

Démonstration. Il suffit d'appliquer le Corollaire 1.11 sur chacun des sous-intervalle de $[a,b]$. □

Exercice 3.
Définir l'approximation linéaire par morceaux d'une fonction passant par les points $\{(-1,0),(0,1),(1,0)\}$.

1.1.7 Interpolation de Hermite

Dans cette section, nous avons approché une fonction f par un polynôme de degré n satisfaisant, en tout les points d'interpolation $(\xi_i)_{1\leq i\leq n+1}$ d'un intervalle $[a,b]$, l'égalité suivante :

$$p_n(\xi_i) = f(\xi_i).$$

C'est l'interpolation de Lagrange, qui est dite "par morceaux" si elle est faite sur des sous-intervalles de $[a,b]$.

Il existe également d'autres formes d'interpolation polynomiale. Par exemple l'interpolation de Hermite, qui consiste à contrainte également la dérivée du polynôme p_n aux points d'interpolation.

*c*f. TD

1.2 Polynôme de meilleure approximation

Dans la section précédente nous avons considéré l'approximation d'une fonction f par un procédé d'interpolation à l'aide des valeurs de cette foncion en des points d'interpolations $(\xi_i)_{1\leq i\leq n+1}$ soigneusement choisis. Cependant, cela suppose que ces valeurs soient connues exactement, ce qui n'est pas le cas dans de nombreuses situations. En particulier lorsque les valeurs de f proviennent de mesures expérimentales.

Le résultat de mesures tel que celui présenté sur la Figure 1.3 conduit à penser que f doit être une fonction affine $f(x) = a_1 x + a_0$. Dans ce cas, il ne semble pas très raisonnable de remplacer $f(x)$ par un polynôme d'interpolation aux points $(\xi_i)_{1\leq i\leq n+1}$ dont le calcul dépendrait des valeurs manifestement erronées $(f(\xi_i))_{1\leq i\leq n+1}$.

En effet, une analyse "statistique" succincte du phénomène ci-dessus indique que ces valeurs $(f(\xi_i))_{1\leq i\leq n+1}$ contiennent une information juste (variant lentement) mais aussi un certain "bruit" (c'est un signal parasite variant rapidement mais de faible amplitude). L'ajustement de données consiste à éliminer ce bruit. Dans cet exemple, le principe de la méthode va

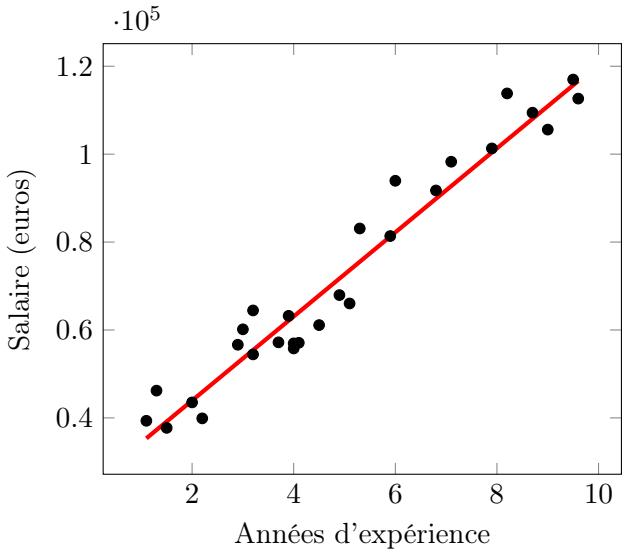

Figure 1.3 – Exemple de régression linéaire

consister à chercher la fonction f, non plus sous la forme d'un polynôme de degré n mais plutôt sous la forme $f(x) = a_1 x + a_0$ où a_0 et a_1 sont calculées de manière à rendre le bruit le plus faible possible. Il reste à définir correctement la notion de plus faible possible, nous pourrions par exemple rechercher les valeurs a_0 et a_1 dans $\mathbb{R}$ solution du problème de minimisation de la quantité suivante,

$$\max_{1 \leq i \leq n+1} |f(\xi_i) - a_1 \xi_i - a_0|. \tag{1.6}$$

Cependant, le plus souvent nous préférons minimiser la quantité suivante,

$$\sum_{i=1}^{n+1} |f(\xi_i) - a_1 \xi_i - a_0|^2, \tag{1.7}$$

car ce problème peut être résolu plus facilement. C'est la méthode des moindres carrés.

Avant de passer à la description de la méthode des moindres carrés, démontrons un résultat d'existence de solutions qui s'applique à la fois aux problèmes (1.6) et (1.7) mais ne donne pas l'unicité de la solution, ce qui est un inconvénient d'un point de vue de la mise en oeuvre de l'algorithme.

1.2.1 Existence d'une meilleure approximation

Soit $(E, \|\cdot\|)$ un espace vectoriel normé, $f \in E$, G un sous-espace vectoriel de E **de dimension finie**.

Définition 1.21. On appelle erreur de meilleure approximation d'une fonction f dans G la quantité

$$\inf_{g \in G} \|f - g\|.$$

Dans le cas particulier où $E = C^0([a,b]; \mathbb{R})$ muni de la norme $\|\cdot\|_\infty$, avec $[a,b]$ un intervalle de $\mathbb{R}$, et $G = \mathcal{P}_n$ avec n un entier naturel fixé, on a le Théorème d'approximation polynomiale suivant.

Théorème 1.22 (Théorème d'approximation polynômiale)**.** *Soit $f \in C^0([a,b];\mathbb{R})$, alors il existe $p^* \in \mathcal{P}_n$ tel que*

$$\|f - p^*\|_\infty = \inf_{p\in\mathcal{P}_n} \|f - p\|_\infty.$$

Autrement dit, l'infimum est atteint, et on a donc $\|f - p^*\|_\infty = \min_{p\in\mathcal{P}_n} \|f - p\|_\infty$, c'est-à-dire que p^* est le polynôme de degré n de meilleure approximation pour la fonction f.

Démonstration. Notons d la distance induite par la norme $\|\cdot\|_\infty$ de l'espace vectoriel considéré, $E = C^0([a,b];\mathbb{R})$ et $G = \mathcal{P}_n$ le sous-espace vectoriel de E de dimension finie contenant les polynômes de degré inférieur ou égal à n.

$$d(f,G) = \inf_{g\in G} \|f - g\|_\infty, \qquad \forall f, g \in E.$$

Par définition de l'infimum, il existe une suite minimisante, c'est-à-dire une suite $(g_k)_{k\geq 0}$ de G telle que

$$\lim_{k\to\infty} \|f - g_k\|_\infty = d(f,G).$$

Montrons alors que la suite $(g_k)_{k\geq 0}$ est bornée. La suite $(\|f - g_k\|_\infty)_{k\geq 0}$ étant convergente, elle est bornée. Il existe donc $M \geq 0$ tel que $\sup_{k\geq 0} \|f - g_k\|_\infty \leq M$ et on a

$$\|g_k\|_\infty \leq \|g_k - f\|_\infty + \|f\|_\infty \leq M + \|f\|_\infty,$$

car la fonction f est continue sur un intervalle borné donc est bornée. Ainsi, la suite $(g_k)_{k\geq 0}$ est une suite bornée d'un espace vectoriel de dimension finie, elle admet donc une valeur d'adhérence $g^* \in G$, c'est-à-dire qu'il existe une sous-suite $(g_{\varphi(k)})_{k\geq 0}$ qui converge vers g^* et on a alors, par continuité, que

$$\lim_{k\to\infty} \|f - g_{\varphi(k)}\|_\infty = \|f - g^*\|_\infty.$$

Or $(\|f - g_{\phi(k)}\|_\infty)_{k\geq 0}$ est une sous-suite de $(\|f - g_k\|_\infty)_{k\geq 0}$ qui est en convergente. Il en résulte donc que

$$\|f - g^*\|_\infty = d(f,G) = \inf_{g\in G} \|f - g\|_\infty.$$

□

Ce résultat est également vrai dans le cadre plus général où $(E, \|\cdot\|_\infty)$ est un espace vectoriel normé et G un sous-espace vectoriel de E de dimension finie. Avant d'énoncer ce résultat, rappelons quelques propriétés importantes des espaces vectoriels de dimension finie.

Lemme 1.23. *Soit $(E, \|\cdot\|)$ un espace vectoriel normé et G un sous-espace vectoriel de dimension finie de E, alors*

- *l'ensemble G est un fermé de E,*
- *l'intersection de deux fermés de E est également fermée,*
- *un sous-ensemble fermé et borné de G est compact.*

Nous avons alors le Théorème d'approximation suivant.

Théorème 1.24 (Théorème d'approximation)**.** *Soit $f \in E$, G un sous-espace vectoriel de E de dimension finie. Alors il existe $g^* \in G$ tel que*

$$\inf_{g\in G} \|f - g\| = \|f - g^*\|.$$

Autrement dit, l'infimum est atteint, et on a donc $\|f - g^*\| = \min_{g\in G} \|f - g\|$.

Démonstration. Soit $f \in E$. Considérons un élément $\bar{g} \in G$ et définissons l'ensemble

$$B = \{g \in G; \|f - g\| \leq \|f - \bar{g}\|\},$$

qui est l'intersection entre le boule de E centrée en f et de rayon $\|f-\bar{g}\|$ avec le sous-espace G. L'ensemble B est évidemment borné et il est également fermé (comme intersection de deux fermés de E). De plus, comme G est de dimension finie, on en déduit que B est un compact de G.

Définissons maintenant l'application $\varphi : g \in B \mapsto \|f - g\| \in \mathbb{R}$. L'application φ est continue sur B (par continuité de la norme sur E), qui est un compact de G. Elle atteint donc ses bornes sur B et en particulier sa borne inférieure. Il existe donc $g^* \in B$ tel que, pour tout $g \in B$, $\|f - g^*\| \leq \|f - g\|$.

Soit $g \in G$ quelconque. Si $g \notin B$ alors $\|f - g\| \geq \|f - \bar{g}\| \geq \|f - g^*\|$. Si $g \in B$ on a bien également que $\|f - g\| \geq \|f - g^*\|$. Ainsi, nous avons trouvé un $g^* \in G$ tel que pour tout $g \in G$, $\|f - g^*\| \leq \|f - g\|$, c'est-à-dire que $\|f - g^*\| = \min_{g\in G} \|f - g\|$. Ce g^* est donc la meilleure approximation de f dans G.

□

Remarque 1.25. L'élément g^* réalisant le minimum n'est pas nécessairement unique.

Exemple 1.26. Pour un intervalle $[a, b]$ de $\mathbb{R}$, prenons $E = \mathcal{C}^0([a, b), \mathbb{R})$ muni de la norme infinie $\|\cdot\|_\infty$ et prenons $\mathcal{P}^1 \subset E$ l'espace vectoriel des polynôme de degré inférieur ou égal à un, qui est de dimension 2. En appliquant les Théorèmes 1.22 ou 1.24, nous démontrons alors qu'il existe au moins une solution au problème de minimisation de la quantité

$$\|f - p^*\|_\infty = \inf_{p\in\mathcal{P}^1} \|f - p\|_\infty,$$

c'est-à-dire un polynôme de degré un qui minimise la norme infinie $\|\cdot\|_\infty$ de $\mathcal{C}^0([a, b), \mathbb{R})$.

1.2.2 Approximation polynômiale uniforme

On se place ici dans l'espace $E = \mathcal{C}^0([a, b], \mathbb{R})$ des fonctions continues sur un intervalle $[a, b]$ et à valeurs dans $\mathbb{R}$, muni de la norme $\|\cdot\|_\infty$:

$$\|f\|_\infty := \sup_{x\in[a,b]} |f(x)|.$$

Définition 1.27. Cette norme est appelée norme de la convergence uniforme. et on dit qu'une suite de fonctions $(f_n)_{n\in\mathbb{N}}$ converge uniformément vers la fonction f sur un intervalle $[a, b]$ si

$$\lim_{n\to\infty} \|f - f_n\|_\infty = 0.$$

Ici, nous cherchons à appocher un élément de E par une fonction polynômiale. Nous considérons donc l'espace vectoriel $\mathcal{P}_n$ des fonctions polynômiales de degré inférieur où égal à n à coefficients dans $\mathbb{R}$, engendré par la famille $\{x \mapsto x^k; k = 0, \ldots, n\}$, comme sous-espace vectoriel de E.

On sait d'après le Théorème 1.24 qu'à n fixé il existe (au moins) un élément de $\mathcal{P}_n$ vérifiant $\|f - p_n\|_\infty = \inf_{g\in\mathcal{P}_n} \|f - g\|_\infty$.

Néanmoins, nous n'avons pas d'indication sur le comportement de l'approximation p_n lorsque n tend vers l'infini. Cette information est donnée par le théorème de Weierstrass suivant.

Théorème 1.28 (Théorème de Weierstrass)**.** *Si f est une fonction continue sur un intervalle $[a,b]$ de $\mathbb{R}$, alors*

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

On en déduit donc en particulier qu' il existe une suite $(f_n)_{n\in\mathbb{N}}$, de polynômes de $\mathcal{P}_n$, qui converge uniformément vers f lorsque $n \to +\infty$. Il est ici aussi possible de quantifier la vitesse de cette convergence

Théorème 1.29. *Si f est de classe $\mathcal{C}^m$ sur $[a,b]$, on a*

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

où C_m est indépendante de n et de f mais dépend à priori de m et de $b-a$.

Remarque 1.30. Pour comparer avec le polynôme d'interpolation de Lagrange $p_n \in \mathcal{P}_n$ étudié dans la section 1.1, il est évident que

$$\inf_{g\in\mathcal{P}_n} \|f-g\|_\infty \leq \|f-p_n\|_\infty.$$

De plus, le phénomène de Runge met en évidence le fait que le polynôme d'interpolation de Lagrange ne converge pas toujours uniformément vers la fonction f que l'on souhaite approcher. Par contre, ici on n'a pas unicité du polynôme de meilleure approximation et on n'a pas de méthode a priori pour construire le polynôme de meilleure approximation, contrairement au polynôme d'interpolation de Lagrange, qui est unique et que l'on sait construire simplement.

1.2.3 Le problème des moindres carrés continu

La méthode des moindres carrés continue consiste à approcher une fonction f sur un intervalle $[a,b]$ de $\mathbb{R}$ par un polynôme de degré inférieur ou égal à n qui minimise l'erreur L^2. Plus précisément, on cherche à minimiser la quantité

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

parmi tous les polynômes $q \in \mathcal{P}_n$.

On se place ici dans l'espace vectoriel $\mathcal{C}([a,b],\mathbb{R})$, où $[a,b]$ est un intervalle borné de $\mathbb{R}$, muni de la norme

$$\|f\|_{L^2} = \left(\int_a^b |f(x)|^2 dx\right)^{1/2}$$

qui dérive du produit scalaire de L^2

$$\langle f,g\rangle_{L^2} = \int_a^b f(x)g(x)dx.$$

On recherche ici le polynôme $q_n \in \mathcal{P}_n$ solution de

$$\|f-q_n\|_{L^2} = \min_{q\in\mathcal{P}_n} \|f-q\|_{L^2}.$$

Le Théorème d'approximation 1.24 nous garantit l'existence d'un tel polynôme, mais on peut ici compléter ce théorème en démontrant son unicité.

Théorème 1.31. *Soit $n \in \mathbb{N}$ et $p_n \in \mathcal{P}_n$ un polynôme de meilleure approximation de $f \in \mathcal{C}([a,b],\mathbb{R})$ muni de la norme $\|\cdot\|_{L^2}$. Alors p_n est unique et est caractérisé par la relation*

$$\langle f - p_n, q\rangle_{L^2} = 0 \qquad \forall q \in \mathcal{P}_n.$$

Autrement dit, p_n est la projection orthogonale de f sur sous-espace vectoriel $\mathcal{P}_n$.

Démonstration. Soit $q \in \mathcal{P}_n$ et $t \in \mathbb{R}^*$. Comme $p_n + tq \in \mathcal{P}_n$, on a par définition de p_n

$$\|f - p_n\|_{L^2}^2 \leq \|f - (p_n + tq)\|_{L^2}^2,$$

Or,

$$\begin{aligned}\|f - (p_n + tq)\|_{L^2}^2 &= \langle f - p_n - tq, f - p_n - tq\rangle,\\ &= \|f - p_n\|_{L^2} - 2t\langle f - p_n, q\rangle_{L^2} + t^2\|q\|_{L^2}^2.\end{aligned}$$

Nous en déduisons donc que

$$0 \leq t^2\|q\|_{L^2}^2 - 2t\langle f - p_n, q\rangle_{L^2}.$$

Lorsque $t > 0$, le passage à la limite $t \to 0^+$ dans la relation $2\langle f - p_n, q\rangle_{L^2} \leq t\|q\|_{L^2}^2$ permet d'obtenir que $\langle f - p_n, q\rangle_{L^2} \leq 0$, tandis que pour $t < 0$, le passage à la limite $t \to 0^-$ dans la relation $2\langle f - p_n, q\rangle_{L^2} \geq t\|q\|_{L^2}^2$ permet d'obtenir que $\langle f - p_n, q\rangle_{L^2} \geq 0$, ce qui prouve le résultat annoncé.
Déduisons alors de cette propriété l'unicité de p_n. Soit alors p_1 et p_2 deux polynômes de meilleure approximation de f et prenons $q = p1 - p2$, on a

$$\begin{aligned}\langle f - p_1, p_1 - p_2\rangle &= 0\\ \langle f - p_2, p_1 - p_2\rangle &= 0\end{aligned}$$

d'où l'on déduit par soustraction que $\|p_1 - p_2\|^2 = 0$. Le polynôme de meilleure approximation de f est donc unique. □

En utilisant les Théorèmes 1.28 et 1.29, on obtient facilement le résultat suivant.

Corollaire 1.32. *Pour toute fonction $f \in \mathcal{C}([a,b],\mathbb{R})$, l'erreur de meilleure approximation de f dans $\mathcal{P}_n$ pour la norme $\|\cdot\|_{L^2}$ vérifie*

$$\lim_{n\to\infty} \|f - p_n\|_{L^2} = 0.$$

De plus si f est de classe $\mathcal{C}^m$ sur $[a,b]$, on a

$$\|f - p_n\|_{L^2} \leq C_m \frac{\|f^{(m)}\|}{n^m},$$

où la constante C_m dépend de m et de $b-a$ et est indépendante de n et de f.

1.2.4 Le problème des moindres carrés discret

Dans le procédé d'interpolation, on a besoin des valeurs de f en $n+1$ points pour construire un polynôme de degré n. Mais si l'on dispose des valeurs de f en $m+1$ avec $m > n$, on peut également construire un polynôme de degré n qui approche f par le procédé dit *des moindres carrés*. Plus précisement, étant donné un vecteur $y = (y_1, \ldots, y_{m+1}) \in \mathbb{R}^{m+1}$, on cherche un polynôme q_n de degré n qui minimise la quantité

$$\sum_{i=1}^{m+1} |q(x_i) - y_i|^2.$$

parmi tous les polynômes $q \in \mathcal{P}_n$. Ce procédé est intuitivement lié à la méthode des moindres carrés continue en remarquant que si on choisit des points $a = x_1 < \cdots < x_{m+1} = b$ équidistants, la quantité

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

qui est minimisée par le polynôme aux points $x_0, \ldots, x_m$ est alors une somme de Riemann qui approche l'intégrale de la section précédente lorsque le nombre de points m augmente.

En écrivant q_n dans la base canonique de $\mathcal{P}_n$: $q_n(x) = \sum_{k=0}^{n} \alpha_k x^k$, alors on voit que la recherche de q_n est équivalente à celle du vecteur $a^* = (\alpha_0, \ldots, \alpha_n) \in \mathbb{R}^{n+1}$ tel que

$$\boxed{\|Va^* - y\|_2 = \min_{a \in \mathbb{R}^{n+1}} \|Va - y\|_2,} \tag{1.8}$$

où $\|\cdot\|_2$ désigne ici la norme euclidienne de $\mathbb{R}^{n+1}$ et où $V \in \mathcal{M}_{m+1,n+1}$ a pour coefficients $V_{ij} = x_i^{j-1}$, pour tout $i \in \{1, \ldots, m+1\}$ et tout $j \in \{1, \ldots, n+1\}$. On utilise alors le résultat suivant.

Théorème 1.33. *Soient $n, m \in \mathbb{N}$ et $y \in \mathbb{R}^{m+1}$. Un vecteur $a^* \in \mathbb{R}^{n+1}$ est solution du problème des moindres carrés*

$$\|Va^* - y\|_2 = \min_{a \in \mathbb{R}^{n+1}} \|Va - y\|_2$$

si et seulement si a^* est l'unique solution du système linéaire (carré)

$$V^T V a^* = V^T y.$$

Afin de démontrer ce Théorème, nous aurons besoin du Lemme suivant.

Lemme 1.34. *Soient $m, n \in \mathbb{N}^*$ tels que $m > n$ et $A \in \mathcal{M}_{m,n}$. On dit que A est injective si $\ker A = \{x \in \mathbb{R}^n \ ; \ Ax = 0_{\mathbb{R}^m}\} = \{0_{\mathbb{R}^n}\}$. Alors, A est injective si et seulement si la matrice $A^T A \in \mathcal{M}_{n,n}$ est inversible.*

Démonstration. La matrice $A^T A$ est une matrice carré de taille $n \times n$, elle est donc inversible si est seulement si elle est injective. En effet, d'après le Théorème du rang, nous savons que

$$n = \dim \ker A^T A + rang\ A^T A = \dim \ker A^T A + \dim \operatorname{im}\ A^T A.$$

Que $A^T A$ est injective et donc équivalent à ce que $\dim \operatorname{im}\ A^T A = n$. Or, le seul sous-espace vectoriel de $\mathbb{R}^n$ de dimension n est $\mathbb{R}^n$ lui-même, donc pour tout $y \in \mathbb{R}^n$ il existe un unique $x \in \mathbb{R}^n$ tel que $A^T Ax = y$, i.e. la matrice est inversible.

Montrons alors que $\ker A = \ker A^T A$. Soit $x \in \ker A^T A$, nous avons donc que $A^T Ax = 0$, ce qui est équivalent à $\langle A^T Ax, y\rangle = 0$ pour tout $y \in \mathbb{R}^n$. En particulier, en prenant $y = x$ nous avons

$$\langle A^T Ax, x\rangle = 0 \Leftrightarrow \langle Ax, Ax\rangle = 0 \Leftrightarrow Ax = 0 \Leftrightarrow x \in \ker A.$$

Réciproquement, si $x \in \ker A$, alors $\langle Ax, y\rangle = 0$ pour tout $y \in \mathbb{R}^m$. Puis, en posant $y = Az$ pour tout $z \in \mathbb{R}^n$, nous avons

$$\langle Ax, Az\rangle = 0 \Leftrightarrow \langle A^T Ax, z\rangle = 0 \Leftrightarrow A^T Ax = 0 \Leftrightarrow x \in \ker A^T A.$$

Ainsi, nous concluons que

$$\ker A = \{0\} \Leftrightarrow \ker A^T A = \{0\} \Leftrightarrow A^T A \text{ est inversible.}$$

□

Nous pouvons maintenant démontrer le Théorème énoncé.

Démonstration du Théorème 1.33. Introduisons l'ensemble $F = \{u \in \mathbb{R}^{m+1}; \exists a \in \mathbb{R}^{n+1}, u = Va\}$. Le problème des moindres carrés (1.8) est donc équivalent à trouver $u^* \in F$ tel que

$$\|u^* - y\|_2 = \min_{u \in F} \|u - y\|_2.$$

Or, on montre facilement que l'espace F est un sous-espace vectoriel de $\mathbb{R}^{m+1}$, donc d'après le Théorème d'approximation 1.24, il existe une solution à ce nouveau problème de minimisation. De plus, comme dans le cas du problème des moindres carrés continu, on peut montrer que cette solution est unique et caractérisée par la relation suivante, pour tout $u \in F$,

$$\langle y - u^*, u \rangle_{\mathbb{R}^{m+1}} = 0.$$

Soit $a \in \mathbb{R}^{m+1}$, notons $u = Va$, qui appartient à l'ensemble F par définition. Comme $u^* \in F$, il existe $a^* \in \mathbb{R}^{m+1}$ tel que $u^* = Va^*$. Ainsi, pour tout $a \in \mathbb{R}^{n+1}$, l'égalité précédente devient

$$\langle y - Va^*, Va \rangle_{\mathbb{R}^{n+1}} = 0,$$

ou de manière équivalente

$$\langle V^T y - V^T V a^*, a \rangle_{\mathbb{R}^{n+1}} = 0,$$

qui est vraie pour tout $a \in \mathbb{R}^{m+1}$. Cela implique que le vecteur $V^T y - V^T V a^*$ est orthogonal à tout vecteur de $\mathbb{R}^{n+1}$, c'est donc le vecteur nul. Ainsi, a^* est solution du système linéaire carré $V^T V a^* = V^T y$. Pour autant, ceci n'assure pas l'unicité. En effet, nous savons que u^* est unique mais il peut exister plusieurs a^* tel que $Va^* = u^*$. L'unicité de a^* est, en fait, liée au caractère injectif de V : si $a = (a_0, \dots, a_m)$,

$$Va = 0 \qquad \Rightarrow q(x_i) = \sum_{k=0}^{n} a_k x_i^k = 0 \quad \forall i \in \{1, \dots, m+1\} \qquad \Rightarrow q = 0 \qquad \Rightarrow a = 0.$$

On en déduit alors, d'après le Lemme 1.34, que la matrice $V^T V$ est inversible. Il existe donc une unique solution au problème des moindres carrés et on peut la déterminer par résolution d'un système linéaire.

□

Nous pouvons alors introduire la définition suivante.

Définition 1.35. Étant donné un vecteur $y = (y_1, \dots, y_{n+1}) \in \mathbb{R}^{n+1}$, on définit le polynôme des moindres carrés de degré $m \leq n$ aux points $(x_1, y_1), \dots, (x_{n+1}, y_{n+1})$ le polynôme $q_m = a_0 + a_1 x + \cdots + a_m x^m \in \mathcal{P}_m$ dont le vecteur des coefficients $a = (a_0, \dots, a_m)$ est solution du système linéaire suivant

$$V^t V a = V^t y.$$

Dans le cas où les y_i sont les valeurs d'une fonction f aux points x_i, l'approximation des moindres carrés de degré m de f aux points $x_1, \dots, x_{n+1}$ est l'unique polynôme $q_m \in \mathcal{P}_m$ qui minimise la quantité $\sum_{i=1}^{n+1} |q(x_i) - f(x_i)|^2$.

Remarque 1.36.

- Ici on a

$$V^t V = \left(\sum_{k=1}^{n+1} x_k^{i+j-2} \right)_{1 \leq i,j \leq m+1} \qquad \text{et} \qquad V^t y = \left(\sum_{k=1}^{n+1} x_k^{j-1} y_k \right)_{1 \leq j \leq m+1}$$

- Dans le cas $n = 0$, la solution constante du problème des moindres carrés est donnée par la moyenne des valeurs y_k

$$a_0 = \frac{1}{n+1} \sum_{i=0}^{n+1} y_i$$

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

Exercice 4.
Déterminer par la méthode des moindres carrés le polynôme de meilleure approximation de degré 1 pour les points $\{(1, 2), (2, 5), (3, 4), (4, 7)\}$.

Chapitre 2

Méthodes de quadrature

Soient (a, b) un intervalle de $\mathbb{R}$ (borné ou non) et f une fonction intégrable sur (a, b). Les *méthodes de quadratures* (aussi appelées méthodes d'intégration numérique) sont des méthodes qui permettent de calculer de façon approchée l'intégrale

$$\int_a^b f(x)dx.$$

En effet, le calcul exact de cette intégrale est parfois impossible, en particulier dans les cas suivants :

- lorsque l'expression de la primitive de f n'est pas connue (comme par exemple la fonction $x \mapsto e^{-x^2}$)
- lorsque la fonction f n'est pas connue explicitement : on peut obtenir ou calculer ses valeurs en des points, mais pas son expression.

2.1 Principe des méthodes de quadrature

2.1.1 Méthodes simples et composées

Les méthodes de quadrature dites *simples* consistent à considérer les valeurs de la fonction f en $n+1$ points $(\xi_j)_{1\leq j\leq n+1}$ sur un intervalle $[a, b]$, afin d'approcher l'intégrale $\int_a^b f(x)dx$ par une expression de la forme

$$I(f) = \sum_{j=1}^{n+1} \alpha_j f(\xi_j),$$

où les coefficients $(\alpha_j)_{j\leq i\leq n+1}$ sont à déterminer. En générale, les points $(\xi_j)_{1\leq j\leq l+1}$ et les coefficients $(\alpha_j)_{1\leq j\leq l+1}$ sont choisis de façon à ce que l'erreur

$$E(f) = |\int_a^b f(x)dx - I(f)|$$

soit la plus petite possible (et en tenant compte des contraintes éventuelles du problème).

Les méthodes de quadrature dites *composées* consistent à découper l'intervalle $[a, b]$ en N sous-intervalles $[a_i, a_{i+1}]$, avec $a = a_1 < a_1 < \cdots < a_{N+1} = b$ et d'appliquer une méthode de quadrature simple sur chaque sous-intervalle $[a_i, a_{i_1}]$ en choisissant $n+1$ points $(\xi_j^i)_{1\leq j\leq n+1}$ sur chaque sous-intervalle $[a_i, a_{i+1}]$. En utilisant le Théorème de Chasles pour les intégrales on obtient alors

$$\int_a^b f(x)dx = \sum_{i=1}^{N} \int_{a_i}^{a_{i+1}} f(x)dx$$

que l'on approche par

$$I(f) = \sum_{i=1}^{N} \sum_{j=1}^{n+1} \alpha_j^i f(\xi_j^i).$$

2.1.2 Intervalle de référence

En général, on définit une méthode de quadrature sur l'intervalle de référence $[-1, 1]$. On peut ensuite se ramener facilement sur n'importe quel intervalle $[a, b]$ de $\mathbb{R}$ par le changement de variable affine suivant :

$$\begin{array}{rccc} \phi: & [-1,1] & \to & [a,b] \\ & t & \to & \frac{a+b}{2} + \frac{(b-a)}{2} t. \end{array}$$

On a alors

$$\int_a^b f(x)dx = \int_{\phi(-1)}^{\phi(1)} f(x)dx = \int_{-1}^{1} f(\phi(t))\phi'(t)dt = \frac{b-a}{2} \int_{-1}^{1} f(\phi(t))dt.$$

2.1.3 Utilisation du polynôme d'interpolation de Lagrange

Sur l'intervalle $[-1, 1]$, une méthode de quadrature simple pour l'intégrale d'une fonction f consiste à choisir $n+1$ points d'interpolation $\xi_1 < \ldots, < \xi_{n+1}$ sur $[-1, 1]$, puis à approcher f par son polynôme d'interpolation de Lagrange p_n en ces points afin d'approcher le calcul de $\int_{-1}^{1} f(t)dt$ par

$$\int_{-1}^{1} f(x)dx \approx \int_{-1}^{1} p_n(t)dt = \int_{-1}^{1} \sum_{j=1}^{n+1} f(\xi_j) l_j(t)dt = \sum_{j=1}^{n+1} f(\xi_j) \int_{-1}^{1} l_j(t)dt = \sum_{j=1}^{n+1} \alpha_j f(\xi_j)$$

où les polynômes $l_j(t)$ sont les polynômes de base de Lagrange pour les points d'interpolation $\xi_1 < \ldots, < \xi_{n+1}$, qui s'écrivent pout tout $j \in \{1, \ldots, n+1\}$

$$l_j(t) = \prod_{k=1, k\neq j}^{n+1} \frac{t - \xi_j}{\xi_k - \xi_j}.$$

On notera également que l'on a définit pour tout $j \in \{1, \ldots, n+1\}$

$$\alpha_j = \int_{-1}^{1} l_j(t)dt.$$

On obtient une méthode simple sur $[a, b]$ après le changement de variable $x = \phi(t)$:

$$\int_a^b f(x)dx \sim \frac{(b-a)}{2} \sum_{j=1}^{n+1} \alpha_j f(\phi(\xi_j))$$

où la fonction ϕ a été définie ci-dessus. On peut aussi définir une méthode composée à partir d'une subdivision $a = a_1 < \cdots < a_N + 1 = b$ de l'intervalle $[a, b]$:

$$\int_a^b f(x)dx \sim \sum_{i=1}^{N} \frac{h_i}{2} \sum_{j=1}^{n+1} \alpha_j f(\xi_j^i)$$

avec

$$h_i = a_{i+1} - a_i, \qquad \xi_j^i = \frac{a_i + a_{i+1}}{2} + \frac{h_i}{2} t_j.$$

2.2 Méthodes de Newton - Cotes

Lorsque l'on choisit une subdivision régulière ($h_i = h$ pour tout $1 \leq i \leq N$) et les points ξ_j^i équidistants, on parle de **méthodes de Newton-Cotes**.

2.2.1 Méthodes classiques

- **Méthode des rectangles à gauche**
  On approche f par la constante $f(a)$ sur $[a, b]$. On obtient la méthode simple

$$\int_a^b f(x)dx \sim \int_a^b f(a)dx = (b-a)f(a)$$

  et la méthode composée

$$\boxed{\int_a^b f(x)dx \sim h\sum_{i=1}^{N} f(a_i).}$$

- **Méthode des rectangles à droite**
  On approche f par la constante $f(b)$ sur $[a, b]$. On obtient la méthode simple

$$\int_a^b f(x)dx \sim (b-a)f(b)$$

  et la méthode composée

$$\boxed{\int_a^b f(x)dx \sim h\sum_{i=2}^{N+1} f(a_i).}$$

- **Méthode du point milieu**
  On approche f par la constante $f(\frac{a+b}{2})$ sur $[a, b]$. On obtient la méthode simple

$$\int_a^b f(x)dx \sim (b-a)f\left(\frac{a+b}{2}\right)$$

  et la méthode composée

$$\boxed{\int_a^b f(x)dx \sim h\sum_{i=1}^{N} f\left(a_{i,1/2}\right), \qquad \text{où} \quad a_{i,1/2} = \frac{a_i + a_{i+1}}{2}.}$$

- **Méthode des trapèzes**
  On approche f par un polynôme de degré inférieur ou égal à 1 sur $[-1, 1]$: le polynôme interpolateur de Lagrange de f aux points -1 et 1

$$p_1(t) = f(-1) + \frac{f(1) - f(-1)}{2}(t+1)$$

  et on a

$$\int_{-1}^{1} f(t)dt \sim \int_{-1}^{1} p_1(t)dt = f(1) + f(-1)$$

  On obtient la méthode simple,

$$\int_a^b f(x)dx \sim \frac{(b-a)}{2}\left(f(a) + f(b)\right)$$

et la méthode composée

$$\boxed{\int_a^b f(x)dx \sim \frac{h}{2}\sum_{i=1}^{N}\left(f(a_i)+f(a_{i+1})\right) = h\sum_{i=2}^{N-1} f(a_i) + \frac{h}{2}\left(f(a)+f(b)\right)}$$

- **Méthode de Simpson**
  On approche f par un polynôme de degré inférieur ou égal à 2 sur $[-1,1]$: le polynôme interpolateur de Lagrange de f aux points $\{-1,0,1\}$. On obtient la méthode simple pour f

$$\boxed{\int_a^b f(x)dx \sim \frac{(b-a)}{6}\left(f(a)+4f(\frac{a+b}{2})+f(b)\right)}$$

  et la méthode composée

$$\boxed{\int_a^b f(x)dx \sim \frac{h}{6}\left(f(a)+f(b)+2\sum_{i=2}^{N-1} f(a_i)+4\sum_{i=1}^{N} f(a_{i,1/2})\right)}$$

Remarque 2.1. Les coefficients $(\frac{1}{2},\frac{1}{2})$ pour la méthode des trapèzes, et $(\frac{1}{6},\frac{2}{3},\frac{1}{6})$ pour la méthode de Simpson sont appelés les poids de la méthode. On remarque que la somme de ces poids est égal à 1. Cela provient du fait que par construction, on doit avoir $\int_a^b \lambda dx = I(\lambda)$ pour toute constante λ.

2.2.2 Analyse de l'erreur

Exemple de la méthode composée des rectangles à gauche

On suppose ici f de classe $\mathcal{C}^1$. En remplaçant f par une formule de Taylor avec reste intégral à l'ordre 1, on peut écrire sur un intervalle $[a_i, a_{i+1}]$

$$\int_{a_i}^{a_{i+1}} f(x)dx = \int_{a_i}^{a_{i+1}}\left[f(a_i)+\int_{a_i}^{x} f'(t)dt\right]dx = hf(a_i)+\int_{a_i}^{a_{i+1}}\int_{a_i}^{x} f'(t)dtdx$$

et

$$\left|\int_{a_i}^{a_{i+1}} f(x)dx - hf(a_i)\right| \leq \|f'\|_\infty \int_{a_i}^{a_{i+1}}\int_{a_i}^{x} dtdx \leq \|f'\|_\infty \frac{h^2}{2}$$

En additionnant cette erreur sur tous les sous-intervalles, on obtient alors

$$\begin{aligned}\left|\int_a^b f(x)dx - h\sum_{i=1}^{N} f(a_i)\right| &= \left|\sum_{i=1}^{N}\left(\int_{a_i}^{a_{i+1}} f(x)dx - hf(a_i)\right)\right| \\ &\leq \sum_{i=1}^{N}\left|\int_{a_i}^{a_{i+1}} f(x)dx - hf(a_i)\right| \\ &\leq \|f'\|_\infty \sum_{i=1}^{N}\frac{h^2}{2}\end{aligned}$$

et finalement

$$\boxed{\left|\int_a^b f(x)dx - h\sum_{i=1}^{N-1} f(a_i)\right| \leq \|f'\|_\infty (b-a)\frac{h}{2}.}$$

Exercice 5.
De la même manière, analyser l'erreur de la méthode des rectangles à droite.

Cas général

Considérons à présent le cas général où la fonction f à intégrer sur un interval réel $[a,b]$ est approchée par son polynôme interpolateur de Lagrange p_n, aux nœuds $\xi_1,\dots,\xi_{n+1}$, inclus dans l'intervalle $[a,b]$. La formule de quadrature simple est donc donnée par

$$I(f)=\int_a^b p_n(x)dx=\sum_{j=1}^{n+1}\alpha_j f(\xi_j),$$

où les poids α_j sont définis par

$$\alpha_j=\int_a^b l_j(x)dx,\quad \forall j\in\{1,\dots,n+1\},$$

avec l_j le polynôme de base de Lagrange au point ξ_j. Nous avons alors l'estimation d'erreur de quadrature suivante.

Théorème 2.2. *Soient $[a,b]$ un intervalle de $\mathbb{R}$ et $f\in\mathcal{C}^{n+1}([a_i,a_{i+1}];\mathbb{R})$, avec $n\in\mathbb{N}$. Considérons de plus la formule de quadrature simple $I(f)=\sum_j\alpha_j f(\xi_j)$ associée au polynôme interpolateur de Lagrange de la fonction f aux nœuds $\{\xi_1,\dots,\xi_{n+1}\}$. Alors*

$$E(f)=\left|\int_a^b f(x)dx-I(f)\right|\leq\frac{(b-a)^{n+2}}{(n+1)!}\|f^{(n+1)}\|_{\infty,[a,b]}.$$

Démonstration. Tout d'abord, on a que

$$\begin{aligned}E(f)&=\left|\int_a^b f(x)dx-I(f)\right|,\\&=\left|\int_a^b (f(x)-p_n(x)dx)\right|,\\&\leq\int_a^b|f(x)-p_n(x)|\,dx.\end{aligned}$$

Or, d'après le Théorème 1.9, on a l'estimation suivante sur l'erreur d'interpolation entre f et p_n en tout point x de $[a,b]$

$$|f(x)-p_n(x)|\leq\frac{\|f^{(n+1)}\|_{\infty,[a,b]}}{(n+1)!}\left|\prod_{i=1}^{n+1}(x-\xi_i)\right|\leq\frac{\|f^{(n+1)}\|_{\infty,[a,b]}}{(n+1)!}(b-a)^{n+1}.$$

Il vient alors

$$\begin{aligned}E(f)&\leq\int_a^b\frac{\|f^{(n+1)}\|_{\infty,[a,b]}}{(n+1)!}(b-a)^{n+1}dx,\\&\leq\frac{\|f^{(n+1)}\|_{\infty,[a,b]}}{(n+1)!}(b-a)^{n+2}.\end{aligned}$$

□

2.3 Ordre d'une méthode

2.3.1 Définition de l'ordre d'une méthode simple

Définition 2.3. On dit qu'une méthode d'intégration numérique (ou méthode de quadrature) simple (du type $I(f) = \sum_{n=1}^{l+1} \alpha_j f(\xi_j)$) est d'ordre au moins m si elle est exacte pour tous les polynômes de degré inférieur ou égal à m, c'est -à-dire

$$E(p) = \int_a^b p(x)dx - I(p) = 0 \qquad \forall p \in \mathcal{P}_m$$

Par linéarité, cela équivaut à dire qu'elle est exacte sur tous les polynômes $x \mapsto x^k$, pour $0 \leq k \leq m$. On voit donc que par construction, les méthodes simples utilisant $n+1$ points de quadrature (et remplaçant donc f par un polynôme de degrés inférieur ou égal à n) sont d'ordre au moins n.
On dit que la méthode est d'ordre exactement m si elle est d'ordre au moins m et n'est pas d'ordre $m+1$.

Remarque 2.4. Si une méthode de quadrature simple sur un intervalle $[a, b]$ est d'ordre $m \geq 0$ alors, en particulier, pour tout polynome constant $p(x) = C$, on a

$$E(p) = 0 \Leftrightarrow \int_a^b C dx = \sum_i \alpha_i C \Leftrightarrow \sum_i \alpha_i = b - a.$$

Exercice 6.
Montrer que la méthode du point milieu est exactement d'ordre 1.

2.3.2 Lien avec l'estimation de l'erreur

Théorème 2.5. *Soient $f \in \mathcal{C}^{m+1}([a, b]; \mathbb{R})$, pour $m \in \mathbb{N}$ et $I(f)$ une formule de quadrature simple sur l'intervalle $[a, b]$ de la forme $I(f) = \sum_{i=1}^{n+1} \alpha_i f(\xi_i)$, pour $n \in \mathbb{N}$ et où les (α_i) sont les poids de la formule de quadrature et les (ξ_i) sont les points de quadrature. Nous supposons de plus que les poids (α_i) sont tous positifs. Si la méthode de quadrature simple $I(f)$ est d'ordre au moins m, alors il existe une constante $C > 0$ telle que*

$$\left| \int_a^b f(x)dx - I(f) \right| \leq C \frac{(b-a)^{m+2}}{(m+1)!} \|f^{(m+1)}\|_{\infty,[a,b]}$$

Démonstration. On applique la formule de Taylor-Lagrange à l'ordre m à la fonction f : il existe $\alpha \in [a, b]$ tel que

$$f(x) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(m)}(a)}{m!}(x-a)^m + \frac{f^{(m+1)}(\alpha)}{(m+1)!}(x-a)^{m+1},$$

que l'on écrit

$$f(x) = P_m(x) + Q(x)$$

avec P_m un polynôme dede degré m et Q le reste.

L'erreur s'écrit alors

$$\begin{aligned}
\left|\int_a^b f(x)dx - I(f)\right| &= \left|\int_a^b (P_m(x)+Q(x))dx - I(P_m+Q)\right| \\
&\leq \left|\int_a^b P_m(x)dx - I(P_m)\right| + \left|\int_a^b Q(x)dx - I(Q)\right| \\
&\leq 0 + \left|\int_a^b \frac{f^{(m+1)}(\alpha)}{(m+1)!}(x-a)^{m+1}dx - \sum_j \alpha_j Q(\xi_j)\right| \\
&\leq \left|\int_a^b \frac{f^{(m+1)}(\alpha)}{(m+1)!}(x-a)^{m+1}dx - \sum_j \alpha_j \frac{f^{(m+1)}(\alpha)}{(m+1)!}(\xi_j-a)^{m+1}\right| \\
&\leq \frac{\|f^{(m+1)}\|_{\infty,[a,b]}}{(m+1)!}\left|(b-a)^{m+2} - \sum_j \alpha_j(\xi_j-a)^{m+1}\right| \\
&\leq \frac{\|f^{(m+1)}\|_{\infty,[a,b]}}{(m+1)!}\left(|b-a|^{m+2} + |b-a|^{m+1}\sum_j |\alpha_j|\right) \\
&\leq 2\frac{\|f^{(m+1)}\|_{\infty,[a,b]}}{(m+1)!}(b-a)^{m+2}
\end{aligned}$$

□

Corollaire 2.6. *Pour une méthode de quadrature composée d'ordre m et de pas h on a alors l'estimation d'erreur suivante*

$$E(f) \leq \frac{C(b-a)\|f^{(m+1)}\|_{\infty,[a,b]}}{(m+1)!}h^{m+1},$$

où C est une constante réelle strictement positive.

2.3.3 Tableau recapitulatif pour les méthodes de Newton-Cotes classiques (avec $h_i = h$)

On note ici $M_k = \sup_{x\in[a,b]} |f^{(k)}(x)|$

Méthode	n	Formule	$\lvert E(f)\rvert \leq$	ordre
Rect. à gauche	0	$h\sum_{i=1}^{N-1} f(a_i)$	$\frac{M_1}{2}(b-a)h$	0
Rect. à droite	0	$h\sum_{i=2}^{N} f(a_i)$	$\frac{M_1}{2}(b-a)h$	0
Point milieu	0	$h\sum_{i=1}^{N-1} f\left(a_{i,1/2}\right)$	$\frac{M_2}{24}(b-a)h^2$	1
Trapèzes	1	$h\left(\sum_{i=2}^{N-1} f(a_i) + \frac{f(a)+f(b)}{2}\right)$	$\frac{M_2}{12}(b-a)h^2$	1
Simpson	2	$\frac{h}{6}\left(f(a)+f(b)+2\sum_{i=2}^{N-1} f(a_i) + 4\sum_{i=1}^{N-1} f(a_{i,1/2})\right)$	$\frac{M_4}{2880}(b-a)h^4$	3

Remarque 2.7. Si $n > 1$, on peut montrer que les méthodes de Newton-Cotes utilisant $n+1$ points de quadrature sont d'ordre n si n est impair, et d'ordre $n+1$ si n est pair.

2.4 Méthodes de Gauss

On a vu que les méthodes de quadrature de Newton-Cotes de degrés n (qui utilisent $n+1$ points de quadrature régulièrement répartis sur chaque sous-intervalle) sont d'ordre n ou d'ordre $n+1$. La recherche de la quadrature ayant l'ordre le plus élevé possible pour un nombre de points donné conduit à la méthode de Gauss.

2.4.1 Polynômes orthogonaux

Soit $]a,b[$ un intervalle, borné ou non, de $\mathbb{R}$.

Définition 2.8. On appelle fonction poids une fonction $\omega \in \mathcal{C}^0(]a,b[;\mathbb{R}_+^*)$ telle que

$$\forall n \in \mathbb{N}, \qquad \int_a^b |x|^n \omega(x)dx < +\infty.$$

Notons $\mathcal{E}$ l'ensemble des fonctions f continues sur $]a,b[$ à valeurs dans $\mathbb{R}$, telles que

$$\|f\|_{\omega,2} := \left(\int_a^b |f(x)|^2 \omega(x)dx\right)^{1/2} < +\infty.$$

Remarque 2.9. L'espace $\mathcal{E}$ est l'espace de Hilbert des fonctions continues et de carré intégrable sur $[a,b]$ par rapport au poid ω. C'est un espace vectoriel normé (dont chaque élément est une fonction), complet, muni du produit scalaire

$$\langle f,g\rangle_\omega = \int_a^b f(x)g(x)\omega(x)dx.$$

De plus, $\mathcal{E}$ contient l'ensemble des fonctions polynômiales sur $[a,b]$.

Définition 2.10. Une suite $(p_n)_{n\in I\subset\mathbb{N}}$ de polynômes est une suite de polynômes orthogonaux par rapport au produit scalaire $\langle\cdot,\cdot\rangle_\omega$ si les polynômes p_n sont deux à deux orthogonaus dans $\mathcal{E}$, i.e. s'ils vérifient

$$\langle p_n, p_m\rangle_\omega = 0 \quad \text{si } n \neq m.$$

Théorème 2.11. *Soit ω une fonction poids sur un intervalle $[a,b]$ de $\mathbb{R}$. Il existe une unique suite de polynômes unitaires $(p_n)_{n\in\mathbb{N}}$ orthogonaux par rapport au produit scalaire $\langle\cdot,\cdot\rangle_\omega$ et tels que $deg(p_n) = n$. Plus précisement, la suite $(p_n)_{n\in\mathbb{N}}$ vérifie la relation de récurrence suivante*

$$\begin{aligned} p_0(x) &= 1, \\ p_1(x) &= x - \lambda_0, \\ p_{n+1}(x) &= (x-\lambda_n)p_n(x) - \mu_n p_{n-1}(x), \forall n \geq 1, \end{aligned}$$

avec

$$\lambda_n = \frac{\langle xp_n, p_n\rangle_\omega}{\|p_n\|_{\omega,2}^2} \qquad \forall n \in \mathbb{N}, \quad \mu_n = \frac{\|p_n\|_{\omega,2}^2}{\|p_{n-1}\|_{\omega,2}^2} \quad \forall n \in \mathbb{N}^*.$$

Démonstration. Construisons la suite de polynômes $(p_n)_{n\in\mathbb{N}}$ par le procédé de Gram-Schmidt. Posons alors $p_0(x) = 1$ et $p_1(x) = 1 - \lambda_0$, où λ_0 est déterminé de sorte que la condition d'orthogonalité entre p_0 et p_1 soit vérifiée.

$$\begin{aligned} \langle p_0, p_1\rangle_\omega = 0 &\Leftrightarrow \int_a^b \omega(x)(x-\lambda_0)dx = 0, \\ &\Leftrightarrow \lambda_0 = \frac{\langle xp_0, p_0\rangle_\omega}{\|p_0\|\omega, 2}. \end{aligned}$$

Posons ensuite $p_2(x) = (x - \lambda_1)p_1(x) - \mu_1 p_0(x0$ où λ_1 et μ_1 sont à déterminer, de sorte que le polynôme p_2 soit orthogonal à p_0 et p_1. En observant en particulier que

$$\|p_1\|^2_{\omega,2} = \langle xp_1, p_0\rangle_\omega,$$

on obtient que

$$\lambda_1 = \frac{\langle xp_1, p_1\rangle_\omega}{\|p_1\|^2_{\omega,2}}, \quad \mu_1 = \frac{\|p_1\|^2_{\omega,2}}{\|p_0\|^2_{\omega,2}}.$$

Procédons ensuite par récurrence et suppons que la proposition est vraie en rang n, c'est-à-dire qu'il existe une famille de polynômes unitaires $\{p_0, \ldots, p_{n+1}\}$ orthognaux par rapport au produit scalaire induit par ω. Posons alors $p_{n+2}(x) = (x - \lambda_{n+1})p_{n_1}(x) - \mu_{n+1}p_n(x)$.

D'une part, démontrons que pout tout $i \in \{0, \ldots, n-1\}$, le polynôme p_{n+2} est orthogonal au polynôme p_i. On a que

$$\langle p_{n+2}, p_i\rangle_\omega = \langle xp_{n+1}, p_i\rangle_\omega - \lambda_{n+1}\langle p_{n+1}, p_i\rangle_\omega - \mu_{n+1}\langle p_n, p_i\rangle_\omega.$$

Or, comme le polynôme $xp_i(x)$ est de degré $i+1$ et la famille $(p_k)_{0\leq k\leq i+1}$ forme une base de l'ensemble des polynôme de degré $i+1$, on peut écrire $xp_i(x)$ comme une combinaison linéaire des polynômes de cette famille. On a alors que

$$\langle p_{n+2}, p_i\rangle_\omega = \sum_{k=0}^{i+1} c_k\langle p_{n+1}, p_k\rangle_\omega = 0.$$

D'autre part, en imposant les condition $\rangle p_{n+2}, p_{n+1}\langle_\omega = 0$ et $\rangle p_{n+2}, p_n\langle_\omega = 0$ on trouve,

$$\lambda_{n+1} = \frac{\langle xp_{n+1}, p_{n+1}\rangle_\omega}{\|p_{n+1}\|^2_{\omega,2}}, \quad \mu_{n+1} = \frac{\langle xp_{n+1}, p_n\rangle_\omega}{\|p_n\|^2_{\omega,2}}.$$

Il ne reste plus qu'à montrer que $\langle xp_{n+1}, p_n\rangle_\omega = \|p_{n+1}\|^2_{\omega,2}$. En effet, puisque l'ensemble des polynômes $\{p_0, \ldots, p_{n+1}\}$ forme une base de l'espace vectoriel des polynômes de degré $n+1$, on peut écrire $xp_{n+1}(x)$ dans cette base et en utilisant le fait que ces polynômes sont unitaires, nous avons que $xp_n(x) = p_{n+1}(x) + \sum_{i=0}^n c_i p_i(x)$ et nous en déduisons l'égalité souhaitée.

□

Exemple 2.12. *Polynômes de Legendre*
On se place dans $]a, b[=]-1, 1[$ et on considère le poids $\omega(x) = 1$. On définit les polynômes de Legendre par

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

Alors les polynômes p_n définis par $p_n = \frac{\mu_n}{L}{}_n(x)$, où μ_n est une constante définie pour que p_n soit unitaire, correspondent à la suite de polynômes orthogonaux définie dans la Théorème 2.11. De plus, le polynôme q_n qui minimise $\|f - q\|_2$ parmi tous les $q \in \mathcal{P}_n$ est donné par

$$q_n = \sum_{k=0}^{n} \langle f, L_k\rangle_\omega L_k.$$

Exemple 2.13. *Polynômes de Tchebychev*
Pour $n \geq 0$ on définit la fonction t_n par

$$\cos(nt) = t_n(\cos(t)), \qquad \text{pour } t \in \mathbb{R}.$$

On peut montrer que les fonctions t_n sont bien définies et sont des polynômes de degré n, de coefficient dominant 2^{n-1}. De plus, la suite $(p_n)_{n\in\mathbb{N}}$ définie par $p_n = \frac{1}{2^{n-1}}t_n$ correspond à la la suite de polynômes orthogonaux définie dans le Théorème 2.11 pour $]a, b[=]-1, 1[$ et le poids $w(x) = 1/\sqrt{1-x^2}$.

Autre exemples

- Pour $]a,b[=]0,+\infty[$ et $\omega = e^{-x}$, ce sont les polynômes de Laguerre
- Pour $]a,b[=]-\infty,+\infty[$ et $\omega = e^{-x^2}$, ce sont les polynômes de Hermite

Par ailleurs, on a la propriété suivante concernant les racines des polynômes p_n.

Proposition 2.14. *Le polynôme p_n possède n zéros réels distincts, inclus dans $]a,b[$.*

2.4.2 Méthode de Gauss

On s'intéresse ici à une méthode de calcul de l'intégrale

$$\int_a^b f(x)\omega(x)dx$$

Théorème 2.15. *Il existe une unique formule de quadrature sur $[-1,1]$ de la forme*

$$\int_a^b f(x)\omega(x)dx \sim I^G(f) := \sum_{j=1}^{n+1} \alpha_j f(\xi_j) \tag{2.1}$$

qui soit d'ordre au moins $2n+1$.

Démonstration.

Unicité

Montrons déjà l'unicité de cette formule. Supposons qu'il existe des points $(\xi_j)_{1\leq j\leq n+1}$ et des coefficients $(\alpha_j)_{1\leq j\leq n+1}$ telle que la formule de quadrature définie par (2.1) soit d'ordre $2n+1$. On pose alors

$$q_{n+1}(x) = \prod_{j=1}^{n+1}(x-\xi_j) \in \mathcal{P}_{n+1}$$

et soit $p \in \mathcal{P}_n$. Comme $d^\circ(q_{n+1}\,p) \leq 2n+1$, la formule de quadrature est exacte pour ce polynôme et on a alors

$$\langle q_{n+1}, p\rangle = \int_a^b q_{n+1}(x)\,p(x)\omega(x)dx = I^G(q_{n+1}\,p) = \sum_{j=1}^{n+1}\alpha_j q_{n+1}(\xi_j)p(\xi_j) = 0$$

car les ξ_j sont les racines de q_{n+1}. On en déduit alors que q_{n+1} est égal (à un coefficient multiplicatif près) au polynôme unitaire p_{n+1} de degré $n+1$ de la base orthogonale introduite dans le Théorème 2.11. Les points $(\xi_j)_{1\leq j\leq n+1}$ correspondent donc aux racines de p_{n+1}. Cela nous donne donc l'unicité des points de quadrature (ξ_j).

On introduit ensuite

$$l_j(x) = \prod_{k=1,\,k\neq j}^{n+1}\frac{x-\xi_k}{\xi_j-\xi_k}$$

les polynômes de base de Lagrange associés aux points $(\xi_j)_{1\leq j\leq n+1}$. Comme les polynômes l_j sont de degré n on a alors que

$$\int_a^b l_j(x)\omega(x)dx = I^G(l_j) = \sum_{k=1}^{n+1}\alpha_k l_j(\xi_k) = \alpha_j$$

ce qui montre que les coefficients $(\alpha_j)_{1\leq j\leq n+1}$ sont également déterminés de façon unique.

Existence
Montrons maintenant qu'avec ce choix des points $(\xi_j)_{1\leq j\leq n+1}$ et les coefficients $(\alpha_j)_{1\leq j\leq n+1}$ la formule de quadrature obtenue est d'ordre $2n+1$. Soit $p \in \mathcal{P}_{2n+1}$. On effectue la division euclidienne de p par p_{n+1} (le $(n+1)$-ème polynôme unitaire de la base orthogonale définie précédemment) : il existe $q, r \in \mathcal{P}_n$ tels que

$$p = qp_{n+1} + r$$

Ainsi

$$\int_a^b p(x)\omega(x)dx = \int_a^b q(x)p_{n+1}(x)\omega(x)dx + \int_a^b r(x)\omega(x)dx = \int_a^b r(x)\omega(x)dx$$

car p_{n+1} est orthogonal au sous-espace vectoriel $\mathcal{P}_n$. Par ailleurs

$$I^G(p) = \sum_{j=1}^{n+1} \alpha_j q(\xi_j)p(\xi_j) + \sum_{j=1}^{n+1} \alpha_j r(\xi_j) = \sum_{j=1}^{n+1} \alpha_j r(\xi_j)$$

car les points de quadrature $(\xi_j)_{1\leq j\leq n+1}$ sont les racines de p_{n+1}. Comme la formule est exacte pour $r \in \mathcal{P}_n$, on en déduit que $\int_a^b p(x)\omega(x)dx = I^G(p)$. La formule est donc exacte pour tout polynôme de $p \in \mathcal{P}_{2n+1}$.

Par ailleurs, comme $d^\circ(l_j^2) = 2n$, on a

$$\int_a^b l_j^2(x)\omega(x)dx = I^G(l_j^2) = \sum_{k=1}^{n+1} \alpha_k l_j(\xi_k)^2 = \alpha_j$$

ce qui montre que les coefficients α_j sont positifs strictement. □

Remarque 2.16. Il ressort de la preuve précédente que

- les $n+1$ points de quadrature $(\xi_j)_{1\leq j\leq n+1}$ correspondent aux racines du polynôme $\mathbf{p}_{n+1}$, où $(\mathbf{p}_n)_{n\geq 0}$ est la suite de polynômes orthogonaux deux-à-deux pour le produit scalaire associée à la fonction poids ω et tels que $\deg(\mathbf{p}_n) = n$.
- le coefficient α_j est égal à l'intégrale du polynôme de base de Lagrange associé à ξ_j contre la fonction poids :

$$\alpha_j = \int_a^b l_j(x)\omega(x)dx.$$

2.4.3 Méthode

La démarche pour la mise en œuvre d'une méthode de Gauss à $n+1$ points pour le calcul de l'intégrale

$$\int_a^b f(x)\omega(x)dx$$

est donc la suivante

1. On détermine la suite des polynômes orthogonaux $p_0, \ldots, p_{n+1}$ pour le produit scalaire associé au poids ω.
2. On détermine (éventuellement de façon numérique) les racines $\xi_1, \ldots, \xi_{n+1}$ de p_{n+1}

3. On calcule
$$l_j(x) = \prod_{k=1,\, k\neq j}^{n+1} \frac{x-\xi_k}{\xi_j-\xi_k}$$
et les coefficients
$$\alpha_j = \int_a^b l_j(x)\omega(x)dx$$

4. On obtient la formule de quadrature
$$\int_a^b f(x)\omega(x)dx \sim \sum_{j=1}^{n+1} \alpha_j f(\xi_j)$$

2.4.4 Exemples

Méthode de Gauss-Legendre
On considère ici l'intervalle $]a,b[=]-1,1[$ et le poids $\omega(x)=1$. Les $n+1$ points de quadrature $(\xi_j)_{1\leq j\leq n+1}$ correspondent ici aux racines du polynômes de Legendre L_{n+1}, et les coefficients α_j à l'intégrale des polynômes de Lagrange associés à ces points :
$$\alpha_j = \int_{-1}^{1} l_j(x)dx.$$

La suite des polynômes de Legendre peut être déterminée grâce à la formule de récurrence
$$L_n(x) = \frac{2n-1}{n} x L_{n-1}(x) - \frac{n-1}{n} L_{n-2}(x),$$
initialisée par $L_0(x)=1$ et $L_1(x)=x$. On obtient alors
$$P_2(x) = \frac{3}{2}(x^2 - \frac{1}{3}), \qquad P_3(x) = \frac{5}{2}x\left(x^2 - \frac{3}{5}\right).$$

Les formules de quadrature simple de Gauss-Legendre s'écrivent alors

- pour la formule à deux points (d'ordre 3)
$$\boxed{\int_{-1}^{1} g(t)dt \sim g\left(\frac{1}{\sqrt{3}}\right) + g\left(\frac{-1}{\sqrt{3}}\right)}$$

- pour la formule à trois points (d'ordre 5)
$$\boxed{\int_{-1}^{1} g(t)dt \sim \frac{5}{9}g\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9}g(0) + \frac{5}{9}g\left(\sqrt{\frac{3}{5}}\right)}$$

On peut déduire de cette méthode de quadrature sur $[-1,1]$ une méthode de quadrature sur n'importe quel intervalle $[a,b]$ à partir du changement de variable $x=\phi(t)=\frac{a+b}{2}+\frac{b-a}{2}t$:
$$\int_a^b f(x)dx = \frac{b-a}{2}\int_{-1}^{1} f(\phi(t))\,dt \sim \frac{b-a}{2}\sum_{j=1}^{n+1} \alpha_j f(\phi(\xi_j)).$$

Méthode de Gauss-Tchebychev
Les polynômes orthogonaux associés au poids $\omega(x) = \frac{1}{\sqrt{1-x^2}}$ sur $[-1, 1]$ sont les polynômes de Tchebychev, dont on sait exprimer les racines (ce sont les points de Tchebychev, *cf.* chapitres 1, 2 et TD). On peut donc construire une formule de quadrature de Gauss pour les intégrales du type $\displaystyle\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}} dx$.

2.4.5 Intérêts des méthodes de Gauss

Comparativement aux méthodes de quadrature de Newton Cotes (où les points de quadrature sont équidistants) les méthodes de Gauss présentent les avantages suivant.

- À nombre de points de quadrature donné, elles sont plus précises. Il est donc intéressant d'utiliser ces méthodes lorsque l'évaluation de la fonction est coûteuse.
- Elles sont mieux adaptées au cas où l'intervalle $[a, b]$ n'est pas borné.
- Elles sont mieux adaptées dans les cas où la fonction à intégrer présente des singularités (les méthodes de Newton-Cotes peuvent être adaptés de façon à éviter l'évaluation de la fonction en ces singularités, mais le résultats est généralement mauvais).

2.4.6 Erreur des méthodes de Gauss

Une estimation de l'erreur des méthodes de Gauss est donnée dans le théorème suivant

Proposition 2.17. *Soient $[a, b]$ un intervalle de $\mathbb{R}$, $f \in \mathcal{C}^{2n+2}([a, b])$ et $I^G_{[a,b]}(f)$ une formule de quadrature simple de Gauss d'ordre $2n + 1$ sur $[a, b]$ pour le calcul de l'intégrale*

$$\int_a^b f(x)\omega(x)dx.$$

Alors, on a

$$\boxed{E(f) = \left| \int_a^b f(x)\omega(x)dx - I^G_{[a,b]}(f) \right| \leq C \frac{\|f^{(2n+2)}\|_{\infty,[a,b]}}{(2n+2)!}(b-a)^{2n+2}}$$

où on a noté $C = \max\left(\displaystyle\int_a^b \omega(x)dx, b-a\right)$, une constante positive.

Démonstration. Si la fonction poids $\omega = 1$, on peut directement appliquer le Théorème 2.5 sur le lien entre l'ordre d'un méthode et l'erreur de quadrature, car les poids d'une méthode de quadrature de Gauss sont toujours positifs.

Si la fonction poids est quelconque, on peut adapter la preuve du Théorème 2.5. □

2.4.7 Méthodes de Gauss composées

Comme pour les méthodes de Newton-Cotes, on peut construire une méthode de Gauss composée en subdivisant un intervalle $[a, b]$ en sous-intervalles $([a_i, a_{i+1}])_{1\leq i\leq N}$. Celle-ci s'écrit sous la forme

$$I^{\mathrm{G,c}}_{[a,b]}(f) = \sum_{i=1}^{N} \frac{h_i}{2} \sum_{j=1}^{n+1} \alpha_j f\left(\theta_{i,j}\right))$$

où $\theta_{i,j} = \dfrac{a_i + a_{i+1}}{2} + \dfrac{h_i}{2}\xi_j$, avec ξ_j les points de quadratures sur $[-1, 1]$.

On déduit ensuite de la Proposition 2.17 l'estimation d'erreur suivante

$$\boxed{E(f) \leq C \frac{M_{2n+2}}{(2n+2)!} h^{2n+2}}$$

où $h = \max_{1 \leq i \leq N} h_i$ et $C > 0$.

Chapitre 3

Résolution numérique d'équations différentielles ordinaires

3.1 Introduction

Soit $d \in \mathbb{N}^\star$, $I = [T_1, T_2]$ un intervalle de $\mathbb{R}$ et $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ une fonction continue. On s'intéresse au problème de Cauchy suivant : trouver une fonction u définie sur $[T_1, T_2]$ et à valeurs dans $\mathbb{R}^d$ telle que

$$\left\{ \begin{array}{rcl} u'(t) & = & f(t, u(t)), \quad t \in I = [T_1, T_2], \\ u(t = T_1) & = & \mu_0 \in \mathbb{R}^d. \end{array} \right. \tag{3.1}$$

La donnée μ_0 est appelée *condition initiale* du problème et on peut montrer que ce problème admet une unique solution, sous certaines conditions de régularité sur la fonction f.

Théorème 3.1. *Soient I un intervalle de $\mathbb{R}$ et $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ une application globalement lipschitizienne par rapport à la deuxième variable, i.e.*

$$\exists L > 0, \quad \forall t \in I, \quad \forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d \quad \|f(t, x) - f(t, y)\| \leq L\|x - y\| \tag{3.2}$$

(pour une norme quelconque de $\mathbb{R}^d$, toutes les normes sur $\mathbb{R}^d$ étant équivalentes). Alors pour toute condition initiale $\mu_0 \in \mathbb{R}^d$, il existe une unique solution u définie sur I et de classe $\mathcal{C}^1$ au problème de Cauchy (3.1).

Si, de plus, f est de classe $\mathcal{C}^r(I \times \mathbb{R}^d; \mathbb{R}^d)$ avec $r \geq 1$, alors u est de class $\mathcal{C}^{r+1}(I; \mathbb{R}^d)$.

Exemple 3.2. L'évolution de l'angle entre un pendule de longueur L et la verticale est régie par le système d'équations dites du pendule :

$$\left\{ \begin{array}{c} \theta''(t) + \frac{g}{L} \sin(\theta(t)) = 0, \quad t \geq 0, \\ \theta(0) = \theta_0 \in \mathbb{R}, \\ \theta'(0) = \theta'_0 \in \mathbb{R}. \end{array} \right. \tag{3.3}$$

En posant

$$u(t) = \begin{pmatrix} \theta(t) \\ \theta'(t) \end{pmatrix} \qquad \text{et} \qquad f : \left(t, \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right) \mapsto \begin{pmatrix} x_2 \\ -\frac{g}{L} \sin(x_1) \end{pmatrix}$$

l'équation (3.3) peut se mettre sous la forme

$$u'(t) = f(t, u(t)).$$

De plus, la fonction f est continue de $\mathbb{R} \times \mathbb{R}^2$ dans $\mathbb{R}^2$ et vérifie l'hypothèse (3.2) : pour tout $x = (x_1, x_2)^t \in \mathbb{R}^2$ et tout $y = (y_1, y_2)^t \in \mathbb{R}^2$

$$\|f(t,x) - f(t,y)\|_2 = \| \begin{pmatrix} x_2 - y_2 \\ \frac{g}{L}(\sin(y_1) - \sin(x_1)) \end{pmatrix} \|_2 \leq \max(1, \frac{g}{L}) \|x - y\|_2,$$

donc il existe une unique solution u de classe $\mathcal{C}^1(\mathbb{R}^+; \mathbb{R}^2)$ à ce problème de Cauchy.

Néanmoins, la plupart du temps, nous ne savons pas résoudre explicitement ce genre de problèmes de Cauchy, c'est pourquoi nous utilisons des méthodes numériques pour approcher leurs solutions.

3.2 Construction de premiers schémas

Nous cherchons à résoudre numériquement le problème de Cauchy (3.1) sur un intervalle $[T_1, T_2]$ de $\mathbb{R}$. Pour cela, nous introduisons une discrétisation de l'intervalle $[T_1, T_2]$ en $N \in \mathbb{N}$ sous-intervalles $([t_n, t_{n+1}])_{n \in \{0,\dots,N\}}$ tels que

$$T_1 = t_0 < t_1 < \cdots < t_N = T_2$$

et l'objectif est alors de construire une suite $(u_n)_{0 \leq n \leq N}$, telle que, pour tout $n \in \{0, \dots, N\}$, l'élément u_n soit *une bonne approximation* de $u(t_n)$. Cette suite est définie par récurrence et la donnée de la formule de récurrence de u_{n+1} en fonction des $(u_k)_{0 \leq k \leq n+1}$ est appelée *schéma numérique*.

Dans toute la suite, nous faisons l'hypothèse que la subdivision $[T_1, T_2]$ est régulière et nous notons h le pas de discrétisation : $h = \frac{T_2 - T_1}{N}$. De plus, dans ce cours, nous considérerons essentiellement des *schéma numérique explicites à un pas*, c'est-à-dire des schéma s'écrivant sous la forme

$$u_{n+1} = u_n + h\Phi(t_n, u_n, h),$$

avec $\Phi : [T_1, T_2] \times \mathbb{R}^d \times [0, \delta] \to \mathbb{R}^d$ une fonction continue et $\delta > 0$.

Définition 3.3.

- Un schéma numérique est dit *explicite* si l'expression de u_{n+1} dépend uniquement des $(u_k)_{0 \leq k \leq n}$ (et ne dépend pas de u_{n+1}). Au contraire, si l'obtention de u_{n+1} dépend de la résolution d'une équation (possiblement non-linéaire) le schéma numérique est dit *implicite*.
- Un schéma numérique est dit à un pas si l'expression de u_{n+1} ne dépend que de u_n (et potentiellement de u_{n+1} s'il est implicite). S'il dépend également des $(u_k)_{0 \leq k \leq n-1}$ (et potentiellement des $(u_k)_{k \geq n+1}$ si le schéma est implicite) le schéma est dit *multi-pas*.

3.2.1 La méthode d'Euler explicite

Nous cherchons à écrire un schéma numérique *simple* pour approcher l'équation

$$\begin{cases} u'(t) = f(t, u(t)), \quad t \in I = [T_1, T_2], \\ u(t = T_1) = \mu_0 \in \mathbb{R}^d. \end{cases} \tag{3.4}$$

Pour cela, nous allons réutiliser les méthodes de quadrature vues au chapitre précédent. L'idée est d'exprimer la solution de (3.4) au temps t_{n+1} sous forme intégrale, en écrivant

$$u(t_{n+1}) = u(t_n) + \int_{t_n}^{t_{n+1}} u'(s)ds = u(t_n) + \int_{t_n}^{t_{n+1}} f(s, u(s))ds,$$

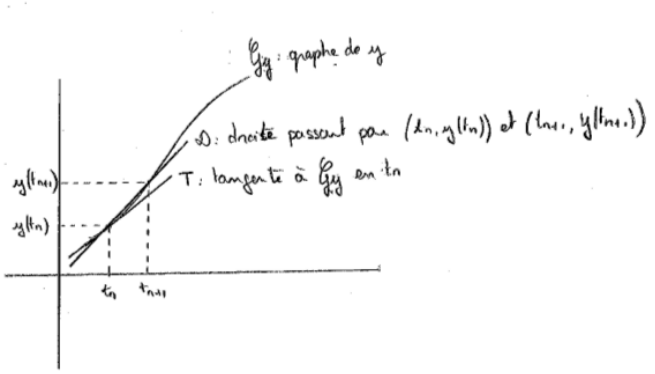

Figure 3.1 – Représentation géométrique de la méthode des différences finies pour le schéna d'Euler explicite.

c'est-à-dire

$$u_{n+1} = u_n + \int_{t_n}^{t_{n+1}} f(s, u(s))ds. \tag{3.5}$$

Ensuite, on utilise la méthode des rectangles à gauche pour approcher l'intégrale qui apparait dans cette formule

$$\int_{t_n}^{t_{n+1}} f(s, u(s))ds \sim (t_{n+1} - t_n) f(t_n, u(t_n)).$$

On trouve alors le schéma numérique suivant, appelé schéma d'Euler explicite, qui dépend de la condition initiale μ_0,

Schéma d'Euler explicite

$$\begin{array}{rcl} u_0 & = & \mu_0, \\ u_{n+1} & = & u_n + hf(t_n, u_n), \quad \forall 1 \leq n \leq N. \end{array} \tag{3.6}$$

Remarque 3.4. Le schéma d'Euler explicite (3.6) est un schéma explicite à un pas.

Remarque 3.5. On peut également construire ce schéma (comme tous les autres schémas d'ailleurs) en utilisant la méthode des *différences finies* qui consiste à approcher la dérivée u' en un point t_n à l'aide d'un dévéloppement de Taylor. En effet, pour tout $n \in \{1, \ldots, N\}$, si on écrit le développement de Taylor de u au point t_{n+1} à l'ordre 1, on a

$$u(t_{n+1}) = u(t_n) + (t_{n+1} - t_n)u'(t_n) + o((t_{n+1} - t_n)).$$

Cela nous donne ainsi une approximation de $u'(t_n)$ par

$$u'(t_n) \approx \frac{u(t_{n+1}) - u(t_n)}{h}.$$

En remplaçant $u'(t_n)$ par cette expression dans l'équation (3.4), on obtient alors de nouveau le schéma d'Euler explicite (3.6).

Géométriquement (pour $d = 1$), cela revient à approcher la tangente du graphe de la fonction u au point $(t_n, u(t_n))$ par la droite reliant les points $(t_n, u(t_n))$ et $(t_{n+1}, u(t_{n+1}))$ (*cf.* Figure 3.1).

3.2.2 La méthode d'Euler implicite

Partant de la relation (3.5), l'utilisation de la méthode des rectangles à droite

$$\int_{t_n}^{t_{n+1}} f(s,u(s))ds \sim (t_{n+1}-t_n)f(t_{n+1},u(t_{n+1})).$$

donne cette fois

$$u_{n+1} = u_n + hf(t_{n+1},u(t_{n+1})). \tag{3.7}$$

On est donc amené ici à construire une suite (u_n) à partir de la relation précédente, pour obtenir le schéma, qui dépend de la condition intiale $\mu_0 \in \mathbb{R}$,

Schéma d'Euler implicite

$$\begin{array}{rcl} u_0 & = & \mu_0, \\ u_{n+1} & = & u_n + hf(t_{n+1},u_{n+1}), \quad \forall 1 \leq n \leq n+1. \end{array} \tag{3.8}$$

Le schéma est dit *implicite* car l'inconnue u_{n+1} est déterminée en résolvant une équation (linéaire si f l'est, non linéaire sinon). Il faut donc éventuellemet utiliser à chaque pas de temps une méthode de résolution d'équation non linéaire (méthode de point fixe, de Newton que nous verrons au deuxième semestre). Grâce au théorème du point fixe de Picard, on peut avoir une condition suffisante pour que cette équation admet une unique solution.

Proposition 3.6. *Si f est globalement L-Lipschitzienne, i.e. s'il existe $L > 0$ tel que $\forall (x,y) \in \mathbb{R}^d \times \mathbb{R}^d$*

$$\|f(t,x) - f(t,y)\| \leq L\|x-y\| \quad \forall t \in I,$$

et si $hL < 1$, alors l'équation d'inconnue x

$$x = u_n + hf(t_{n+1},x)$$

admet une unique solution.

3.2.3 Schéma de Cranck-Nicholson et θ- schémas

Partant de la relation (3.5) l'utilisation de la méthode des trapèzes donne

$$\int_{t_n}^{t_{n+1}} f(s,u(s))ds \sim \frac{h}{2}\left[f(t_n,u(t_n)) + f(t_{n+1},u(t_{n+1}))\right].$$

Le schéma numérique résultant de cette idée est le schéma de Crank-Nicholson, qui s'écrit

Schéma de Cranck-Nicholson

$$\begin{array}{rcl} u_0 & = & \mu_0, \\ u_{n+1} & = & u_n + \frac{h}{2}\left[f(t_n,u_n) + f(t_{n+1},u_{n+1})\right]. \end{array} \tag{3.9}$$

Remarque 3.7. Le schéma de Crank-Nocholson est un schéma implicite à un pas.

Plus généralement, on appelle θ-schéma le schéma

$$u_{n+1} = u_n + h\left[(1-\theta)f(t_n,u_n) + \theta f(t_{n+1},u_{n+1})\right].$$

Les schémas d'Euler explicite, Euler implicite et Cranck-Nicholson font partie de la famille des θ-schémas, pour $\theta = 0$, $\theta = 1$ et $\theta = 1/2$ respectivement.

3.2.4 Schéma du point milieu

Partant de la relation (3.5) l'utilisation de la méthode du point milieu donne

$$\int_{t_n}^{t_{n+1}} f(s,u(s))ds \sim hf\left(t_n+\frac{h}{2},u(t_n+\frac{h}{2})\right).$$

Cependamt, pour construire un schéma à partir de cette relation, il nous faut introduire une approximation de $u(t_n+\frac{h}{2})$, que l'on choisit de prendre ici par le schéma d'Euler explicite :

$$u(t_n+\frac{h}{2}) = u(t_n)+\frac{h}{2}f\left(t_n,u(t_n)\right)+o(h)$$

On obtient alors la méthode du point milieu

Schéma du point milieu

$$\begin{array}{rcl} u_0 & = & \mu_0, \\ u_{n+1} & = & u_n + hf\left(t_n+\frac{h}{2},u_n+\frac{h}{2}f(t_n,u_n)\right). \end{array} \tag{3.10}$$

Exemple 3.8. On considère l'équation différentielle $u'(t)=Au(t)$ où $A\in\mathcal{M}_d(\mathbb{R})$. Le schéma d'Euler explicite donne

$$u_{n+1}=(Id+hA)u_n,$$

et le schéma d'Euler implicite

$$u_{n+1}=(Id-hA)^{-1}u_n.$$

Pour le schéma d'Euler implicite, si h est suffisant petit on peut montrer en particuler que la matrice $Id-hA$ est bien inversible.

3.3 Analyse des schémas numériques explicites à un pas

Dans cette partie, nous nous intéressons à l'analyse des méthodes explicites dites à un pas. Comme toutes les normes sont équivalentes sur R^d, nous considérons une norme quelquonque notée $\|\cdot\|$. On donne ci-dessous une définition d'un schéma numérique explicite à un pas.

Définition 3.9. Méthodes à un pas
Une méthode (ou un schéma numérique) est dite à un pas si elle s'écrit de la façon suivante :

$$u_{n+1}=u_n+h\Phi(t_n,u_n,h) \tag{3.11}$$

avec $\Phi : I\times\mathbb{R}^d\times\mathbb{R}\to\mathbb{R}^d$ une fonction continue.

Exemple 3.10.

- Pour la méthode d'Euler explicite on a $\Phi(t,u,h)=f(t,u)$,
- pour la méthode du point milieu on a

$$\Phi(t,u,h)=f\left(t+\frac{h}{2},u+\frac{h}{2}f(t,u)\right).$$

Soient u la solution exacte du problème de Cauchy et $(u_n)_{0\le n\le N}$ la solution approchée donnée par le schéma explicite à un pas (3.11), avec $N\in\mathbb{N}$. Nous définissons l'erreur de convergence au temps t_n, comme la différence entre les solutions exacte et approchée :

$$e_n=u(t_n)-u_n,\quad \forall n\in\mathbb{N}.$$

La notion de convergence d'un schéma numérique est alors donnée par la définition suivante.

Définition 3.11. Convergence d'une schéma numérique
Un schéma numérique explicite à un pas est dit convergent sur l'intervalle $I = [T_1, T_2]$ si

$$\lim_{h\to 0} \max_{0\leq n\leq N} \|e_n\| = 0,$$

avec $h = \frac{T_2-T_1}{N}$.

En pratique, pour montrer la convergence d'un schéma numérique on s'intéresse d'abord à deux notions importante : la **consistance** qui renseigne sur la cohérence de la discrétisation et la **stabilité** qui signifie le contrôle de l'accumulation des erreurs.

3.3.1 Consistance et stabilité de schémas numériques à un pas

Définition 3.12. Consistance d'une méthode
On appelle erreur de consistance

$$\eta_n = u(t_{n+1}) - u(t_n) - h\phi(t_n, u(t_n), h)$$

où u est la solution exacte du problème de Cauchy (3.1).
Un schéma numérique est alors dit consistant si

$$\lim_{h\to 0} \sum_{n=0}^{N-1} \|\eta_n\| = 0.$$

De plus, il est dit consistant d'ordre $k \geq 1$ si

$$\max_{0\leq n\leq N-1} \|\eta_n\| \leq Ch^{k+1},$$

où $C > 0$ est une constante.

Remarque 3.13. Si le schéma est consistant d'ordre $k \geq 1$, il est bien évidement consistant. On a en effet

$$0 \leq \lim_{h\to 0} \sum_{n=0}^{N-1} \|\eta_n\| \leq \lim_{h\to 0} N \max_{0\leq n\leq N-1} \|\eta_n\| \leq N \lim_{\Delta t\to 0} Ch^{k+1} = 0.$$

Exemple 3.14. Pour le schéma d'Euler explicite, si on suppose f de classe $\mathcal{C}^1([T_1, T_2]\times\mathbb{R}^d; \mathbb{R}^d)$ et globalement lipschitzienne par rapport à sa deuxième variable, alors l'unique solution u est de classe $\mathcal{C}^2([T_1, T_2]; \mathbb{R}^d)$ (d'après le Théorème de Cauchy-Lipschitz), et l'erreur de consistance s'écrit

$$\begin{aligned}\eta_n &= u(t_{n+1}) - u(t_n) - hf(t_n, u(t_n)) \\ &= u(t_{n+1}) - u(t_n) - hu'(t_n).\end{aligned}$$

Or, en utilisant la formule de Taylor-Lagrange à l'ordre 1 pour u au point t_{n+1}, on trouve qu'il existe $\varepsilon \in]0, 1[$ tel que

$$u(t_{n+1}) = u(t_n) + hu'(t_n) + \frac{h^2}{2}u''(t_n + \varepsilon h).$$

L'erreur de consistance s'écrit alors

$$\eta_n = \frac{h^2}{2}u''(t_n + \varepsilon h) \quad \forall 0 \leq n \leq N-1,$$

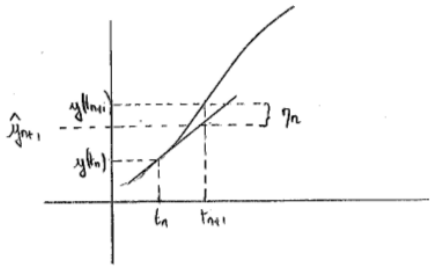

FIGURE 3.2 – Représentation géométrique de l'erreur de consistance.

et on a

$$\max_{0\leq n\leq N}\|\eta_n\| \leq \frac{1}{2}M_2h^2, \qquad \text{où } M_2 := \sup_{t\in[T_1,T_2]}|u''(t)|.$$

Le schéma d'Euler est donc consistant d'ordre 1.

Remarque 3.15. L'erreur de consistance est aussi égale à

$$\eta_n = u(t_{n+1}) - \hat{u}_{n+1},$$

où $\hat{u}_{n+1}$ est défini par

$$\begin{cases}\hat{u}_{n+1} = \hat{u}_n + h\Phi(t_n,\hat{u}_n,h)\\ \hat{u}_n = u(t_n)\end{cases}$$

Ainsi, η_n mesure l'écart entre la valeur exacte au temps t_{n+1} de la solution de l'équation différentielle et la valeur approchée donnée après un pas du schéma en partant de $u(t_n)$ (*cf.* Figure 3.2). La consistance donne alors une indication sur la cohérence de l'approximation Φ de la fonction f. Elle signifie en effet que la fonction $\Phi(t,u(t),h)$ converge vers $f(t,u(t))$ quand le pas h tend vers 0.

Nous pouvons alors établir une condition suffisante sur la fonction Φ pour que le schéma (3.11) soit consistant.

Théorème 3.16. *Considérons le schéma* (3.11) *associé à l'équation différentielle* (3.1). *Si, pour tout* $x\in\mathbb{R}^d$ *et pour tout* $t\in I$, *la fonction* $\Phi\in\mathcal{C}^0(I\times\mathbb{R}^d\times\mathbb{R};\mathbb{R}^d)$ *vérifie*

$$\Phi(t,x,0) = f(t,x)$$

alors le schéma explicite à un pas (3.11) *est consistant.*

Démonstration. Posons $I=[T_1,T_2]$ et prenons $u\in\mathcal{C}^1([T_1,T_2];\mathbb{R}^d)$ la solution exacte de (3.1). Nous pouvons écrire, pour $n\in\{0,\dots,N-1\}$ (avec $N\in\mathbb{N}^\star$),

$$u(t_{n+1}) - u(t_n) = \int_{t_n}^{t_n+h} u'(s)ds = \int_{t_n}^{t_n+h} f(s,u(s))ds.$$

Nous en déduisons alors que

$$\begin{aligned}\eta_n &= u(t_{n+1}) - u(t_n) - h\Phi(t_n,u(t_n),h),\\ &= \int_{t_n}^{t_n+h}\left(f(s,u(s)) - \Phi(t_n,u(t_n),h)\right)ds.\end{aligned}$$

Soit $\varepsilon > 0$, puisque Φ est continue au point $(t_n, u(t_n), 0)$ et que $\Phi(t_n, u(t_n), 0) = f(t_n, u(t_n))$, il existe $\tau_1 > 0$ tel que pour tout $h < \tau_1$

$$\|\Phi(t_n, u(t_n), h) - f(t_n, u(t_n))\| \leq \frac{\varepsilon}{2}.$$

Par inégalité triangulaire, nous avons alors

$$\|\eta_n\| \leq \frac{\varepsilon}{2}h + \int_{t_n}^{t_n+h} \|f(s, u(s)) - f(t_n, u(t_n))\| ds.$$

D'autre part, la fonction $s \mapsto f(s, u(s))$ est continue sur $[T_1, T_2]$, elle est donc uniformément continue et il existe $\tau_2 > 0$ tel que pour tout $|t - s| \leq h \leq \tau_2$

$$\|f(s, u(s)) - f(t, u(t))\| \leq \frac{\varepsilon}{2}.$$

Ainsi, lorsque $h \leq \min(\tau_1, \tau_2)$, l'erreur de consistance satisfait, pour tout $n \in \{0, \ldots, N-1\}$,

$$\|\eta_n\| \leq \varepsilon h$$

et on a alors que, pour tout $\epsilon > 0$ et pour tout $N \in \mathbb{N}^\star$,

$$\sum_{n=0}^{N-1} \|\eta_n\| \leq N\varepsilon h = \varepsilon(T_2 - T_1).$$

ce qui démontre le résultat (car ε est aussi petit qu'on veut). □

Exemple 3.17. D'après ce critère, les θ-schémas et le schéma du point milieu sont consistants.

Ce critère de consistance donne un information sur la précision d'une étape du schéma mais il n'assure pas que l'accumulation des erreurs converge lorsque le nombre d'itérations croît. Pour exprimer cela, nous introduisons la notion de stabilité d'une schéma numérique.

Définition 3.18. Stabilité d'un schéma numérique
Un schéma numérique à un pas est dit stable s'il existe $h^\star > 0$ et $S > 0$ tels que pour tout $0 \leq h < h^\star$ et toutes suites $(u_n)_{0 \leq n \leq N}$ et $(v_n)_{0 \leq n \leq N}$ définies par

$$u_{n+1} = u_n + h\Phi(t_n, u_n, h),$$
$$v_{n+1} = v_n + h\Phi(t_n, v_n, h) + \varepsilon_n,$$

avec $(\varepsilon_n)_{0 \leq n \leq N}$ donnée, on a

$$\|u_n - v_n\| \leq S\left(\|u_0 - v_0\| + \sum_{n=0}^{N-1} \|\varepsilon_n\|\right), \quad \forall n \in \{0, \ldots, N\}.$$

De plus, le schéma est dit *inconditionnellement* stable si $h^\star = \infty$.

Ainsi, une méthode est stable si l'accumulation de petites perturbations sur les calculs à à chaque pas de temps (erreurs d'arrondis par exemple) et sur les données initiales mène à une erreur finale contrôlable. Ce critère est absolument nécessaire pour qu'un schéma numérique soit réaliste. Cependant, en pratique il n'est utilisable que si la constante S n'est *pas trop grande*.

Nous disposons d'une condition suffisante pour garantir la stabilité d'un schéma numérique à un pas.

Théorème 3.19. *Si la fonction Φ est continue et globalement lipschitzienne par rapport à la deuxième variable, c'est-à-dire s'il existe $L > 0$ tel que pour tout $(t, h) \in [T_1, T_2] \times \mathbb{R}$,*

$$\|\Phi(t, u, h) - \Phi(t, v, h)\| \leq L\|u - v\|, \quad \forall (u, v) \in \mathbb{R}^d \times \mathbb{R}^d$$

alors le schéma numérique à un pas est stable, avec la constante de stabilité qui vaut

$$S = e^{L(T_2 - T_1)}.$$

Pour démontrer ce résultat, nous aurons besoin du Lemme de Gronwall discret suivant, qui se montre facilement Nous aurons besoin pour la preuve de ce théorème du Lemme suivant, qui se montre facilement par récurrence.

Lemme 3.20. (Gronwall discret) *Soient $\lambda > 0$ et $(z_n)_{n\geq 0}$ vérifiant la relation*

$$z_{n+1} \leq (1 + \lambda)z_n + \beta_n \qquad \forall 0 \leq n \leq N - 1.$$

Alors pour tout $0 \leq n \leq N$ on a

$$z_n \leq e^{\lambda n} z_0 + \sum_{k=0}^{n-1} e^{\lambda(n-k-1)} \beta_k$$

Démonstration Démonstration du Théorème 3.19.
Soient $(u_n)_{0\leq n\leq N}$ et $(v_n)_{0\leq n\leq N}$ définies par

$$\begin{aligned} u_{n+1} &= u_n + \Delta t \Phi(t_n, u_n, \Delta t) \\ v_{n+1} &= v_n + \Delta t \Phi(t_n, v_n, \Delta t) + \varepsilon_n \end{aligned}.$$

On peut écrire

$$\begin{aligned} \|v_{n+1} - u_{n+1}\| &= \|v_n - u_n + \Delta t \Phi(t_n, v_n, \Delta t) - \Delta t \Phi(t_n, u_n, \Delta t) + \varepsilon_n\| \\ &\leq (1 + \Delta t L)\|v_n - u_n\| + \|\varepsilon_n\| \end{aligned}.$$

En utilisant le Lemme 3.20 on obtient

$$\|v_n - u_n\| \leq e^{Ln\Delta t}\|v_0 - u_0\| + \sum_{k=0}^{n-1} e^{L\Delta t(n-k-1)}\varepsilon_k \quad \leq e^{L(T_2-T_1)} \left(\|v_0 - u_0\| + \sum_{k=0}^{n-1} \varepsilon_k \right).$$

□

Exemple 3.21. Soit $f : [T_1, T_2] \times \mathbb{R}^d \to \mathbb{R}^d$ une fonction continue et Lipschitzienne en la seconde variable, de constante de Lipschitz $L > 0$. Montrons que le schéma d'Euler explicite suivant est stable :

$$\left\{ \begin{array}{rcl} u_{n+1} & = & u_n + \Delta t f(t_n, u_n), \\ u_0 & = & \mu_0 \in \mathbb{R}^d. \end{array} \right.$$

D'après le Théorème 3.19, il suffit pour cela de montrer que la fonction

$$\begin{array}{rccc} \Phi & : & [T_1, T_2] \times \mathbb{R}^d \times \mathbb{R}^+ & \to & \mathbb{R}^d \\ & & (t, u, \Delta t) & \mapsto & f(t, u) \end{array}$$

est continue et Lipschitzienne par rapport à la deuxième variable, ce qui est évidemment vrai ici.

3.3.2 Convergence de schémas numérique à un pas

À partir des notions de consistance et de stabilité, nous pouvons démontrer la convergence de la solution numérique vers la solution exacte du problème (3.1).

Théorème 3.22. *Si un schéma numérique explicite à un pas est consistant et stable alors il est convergent. De plus, si le schéma est consistant d'ordre $k \geq 1$, alors le schéma est dit convergent d'ordre k et il existe une constante $C > 0$ indépendante de Δt telle que l'erreur de convergence vérifie*

$$\|e_n\| \leq C\Delta t^k.$$

Démonstration. Le schéma numérique à un pas s'écrit, pour tout $0 \leq n \leq N-1$

$$u_{n+1} = u_n + \Delta t \Phi(t_n, u_n, \Delta t).$$

De plus, par définition de l'erreur de consistance, on a, pour tout $0 \leq n \leq N-1$

$$u(t_{n+1}) = u(t_n) + \Delta t \Phi(t_n, u(t_n), \Delta t) + \eta_n.$$

En définissant alors la suite $(v_n)_{0 \leq n \leq N}$ par

$$v_0 = u_0, \qquad v_{n+1} = v_n + \Delta t \phi(t_n, v_n, \Delta t) + \eta_n$$

on a

$$v_n = u(t_n) \quad \forall n \in \{0, \dots, N\}$$

et la stabilité de la méthode implique que

$$\|e_n\| = \|v_n - u_n\| \leq S \left(\|u_0 - u_0\| + \sum_{n=0}^{N-1} \|\eta_n\| \right).$$

Or, comme la méthode est consistante $\displaystyle\sum_{n=0}^{N-1} \|\eta_n\| \to 0$ lorsque $N \to +\infty$, ce qui prouve la convergence de la méthode.

De plus, si le schéma est consistant d'ordre k, il existe une constante $C' > 0$ telle que

$$\max_{0 \leq n \leq N-1} \|\eta_n\| \leq C' \Delta t^{k+1},$$

et on obtient alors pour l'erreur de convergence

$$\|e_n\| \leq SN \max_{0 \leq n \leq N-1} \|\eta_n\| \leq SC'(T_2 - T_1)\Delta t^k.$$

Nous définissons alors $C = SC'(T_2 - T_1)$ la constante apparaissant dans l'énoncé du théorème et la preuve est terminée. □

3.4 Notion de stabilité absolue

Un schéma numérique peut être stable en théorie, mais inutilisable en pratique à cause d'une trop grande constante de stabilité qui pourrait entrainer une *explosion* de la solution au bout d'un certain temps. Dans cette partie, nous nous intéressons à une étude plus précise du comportement de la solution en temps long.

Considérons le cas de référence $u'(t) = \lambda u(t)$, avec $\lambda < 0$. Le schéma d'Euler explicite donne

$$u_n = (1 + \lambda \Delta t)^n u_0,$$

et le schéma d'Euler implicite

$$u_{n+1} = \frac{1}{(1 - \lambda \Delta t)^n} u_0.$$

Comme $\lambda < 0$, la solution exacte vérifie

$$u(t) = e^{\lambda t} u_0 \to 0 \quad \text{lorsque } t \to +\infty.$$

Nous nous attendons donc à ce que $\lim_{n \to +\infty} u_n = 0$. Pour le schéma d'Euler implicite, cette proprité est bien vérifiée quelque soit Δt. Par contre, pour le schéma d'Euler explicite, la condition $|\lambda \Delta t| < 2$ doit être vérifiée. Cela nous amène alors à introduire la définition suivante.

Définition 3.23. Un schéma numérique est dit *absolument stable* si, quand il est appliquée au problème de référence

$$u'(t) = \lambda u(t), \qquad \lambda \in \mathbb{C} \text{ avec } \Re(\lambda) < 0, \tag{3.12}$$

il vérifie

$$\forall u_0 \in \mathbb{R}^d, \qquad \lim_{n \to +\infty} |u_n| = 0.$$

En écrivant le schéma numérique considéré appliqué au problème de référence (3.12) sous la forme

$$u_{n+1} = G(\lambda \Delta t) u_n$$

on voit que la méthode est absolument stable si et seulement si $|G(\lambda \Delta t)| < 1$.

Définition 3.24. On appelle région de stabilité absolue le domaine

$$\mathcal{A} = \{z \in \mathbb{C} / |G(z)| < 1\}.$$

Proposition 3.25. *Un schéma numérique est inconditionnellement absolument stable si*

$$\{z \in \mathbb{C}, \Re(z) < 0\} \subset \mathcal{A}.$$

Pour étudier la stabilité absolue d'un schéma, il faut donc déterminer l'ensemble des valeurs Δt pour lesquelles $\lambda \Delta t$ appartient au domaine de stabilité du schéma.

Exemple 3.26. On peut montrer que

- pour le schéma d'Euler implicite le domaine de stabilité absolue est le plan complexe privé du disque de centre $(1, 0)$ et de rayon 1. Le demi-plan complexe formé des des $z \in \mathbb{C}$ tel que $\Re(z) < 0$ appartient donc au domaine de stabilité. On en déduit que le schéma d'Euler implicite est **inconditionnellement absolument stable**.
- pour le schéma d'Euler explicite, le domaine de stabilité est le disque du plan complexe de centre $(-1, 0)$ et de rayon 1. Pour $\lambda = \rho e^{i\theta}$, le schéma est absolument stable si $\Delta t \in]0, -2\frac{\cos(\theta)}{\rho}[$. Le schéma d'Euler explicite est donc seulement **conditionnement absolument stable**.
- les θ-schémas sont inconditionnellement absolument stable si et seulement si $\theta \in [0.5, 1]$. En particulier le schéma de Cranck-Nicholson est inconditionnellement absolument stable.

3.5 Méthodes de Runge-Kutta

3.5.1 Construction

Nous avons vu que la méthode d'Euler explicite est consistante d'ordre 1. Dans la pratique, une méthode d'ordre 1 ou 2 se révèle assez souvent insuffisante, car elle nécessite un pas de temps trop petit pour atteindre une précision donnée. Il est alors nécessaire d'utiliser une méthode d'ordre supérieure : les plus courantes sont celles de Runge-Kutta qui reposent sur des méthodes d'intégration numérique d'ordre supérieur pour approcher le terme

$$\int_{t_n}^{t_{n+1}} f(s,u(s))ds.$$

Pour cela, on introduit une subdivision de l'intervalle $[t_n, t_{n+1}]$, pour tout $n \in \{0,\dots,N-1\}$:

$$t_n \leq t_{n,1} < \cdots < t_{n,q} \leq t_{n+1}, \text{ avec } t_{n,i} = t_n + \Delta t c_i, \quad 0 \leq c_i \leq 1, \quad 1 \leq i \leq q \tag{3.13}$$

et l'utilisation d'une formule de quadrature conduit à une formule du type

$$\int_{t_n}^{t_{n+1}} f(s,u(s))ds \sim \Delta t \left[\sum_{i=1}^{q} b_i f(t_{n,i}, u(t_{n,i}))\right], \tag{3.14}$$

avec les $(b_i)_{0\leq i\leq q}$ des coefficients à determiner.

Les méthodes de Runge-Kutta consistent alors à utiliser cette formule de quadrature pour construire un schéma de la forme

$$u_{n+1} = u_n + \Delta t \left[\sum_{i=1}^{q} b_i f(t_{n,i}, u_{n,i})\right], \tag{3.15}$$

les valeurs $u_{n,i}$ étant elles-mêmes évaluées à l'aide de formules d'intégration numérique utilisant les mêmes points $(t_{n,j})_{1\leq j\leq q}$:

$$u_{n,i} = u_n + \Delta t \left[\sum_{j=1}^{q} a_{i,j} f(t_{n,j}, u_{n,j})\right]. \tag{3.16}$$

Remarque 3.27.

- En prenant $q = 1$, $c_1 = 0$, $b_1 = 1$ et $a_{1,1} = 1$ on retrouve le schéma d'Euler explicite.
- La formule (3.16) définit des valeurs $u_{n,i}$ de façon explicite si la matrice $A = (a_{i,j})_{1\leq i\leq q, 1\leq j\leq q}$ est strictement triangulaire inférieure, sinon elles sont obtenues de façon implicite.
- Si la formule de quadrature (3.14) est au moins d'ordre 0, c'est-à-dire exacte pour les constantes, alors on a les relations

$$\sum_{i=1}^{q} b_i = 1, \qquad \sum_{j=1}^{q} a_{i,j} = c_i, \text{ pour tout } 1 \leq i \leq q. \tag{3.17}$$

3.5.2 Propriétés

Commençons par évoquer la stabilité de ces méthodes.

Théorème 3.28. *Si f est L-lipschiztienne par rapport à sa seconde variable, uniformément par rapport à la première, c'est -à-dire si*

$$\forall (x,y) \in \mathbb{R}^d \times \mathbb{R}^d, \forall t \in [T_1, T_2], \qquad \|f(t,x) - f(t,y)\| \leq L\|x-y\|$$

alors un schéma de Runge-Kutta définie par les relations (3.13)- (3.15)-(3.16) *appliqué à l'EDO* (3.1) *est stable.*

De plus, avec un bon des choix des coefficients, il est possible d'obtenir des méthodes consistantes d'ordres élevés. On a d'ailleurs le résultat suivant.

Théorème 3.29. *Le schéma de Runge-Kutta définie par les relations* (3.13)- (3.15)-(3.16) *est consistant*

- *d'ordre au moins* 1 *si et seulement si*

$$\sum_{i=1}^{q} b_i = 1, \qquad \sum_{j=1}^{q} a_{i,j} = c_i, \text{ pour tout } 1 \leq i \leq q, \tag{3.18}$$

- *d'ordre au moins* 2 *si et seulement les conditions* (3.18) *sont vérifiés et*

$$\sum_{j=1}^{q} b_j c_j = \frac{1}{2}, \tag{3.19}$$

- *d'ordre au moins* 3 *si et seulement les conditions* (3.18) *et* (3.19) *sont vérifiés et*

$$\sum_{i=1}^{q}\sum_{j=1}^{q} b_i a_{i,j} c_j = \frac{1}{6}, \tag{3.20}$$

- *d'ordre au moins* 4 *si et seulement les conditions* (3.18), (3.19) *et* (3.20) *sont vérifiés et*

$$\sum_{i=1}^{q}\sum_{j=1}^{q} b_i a_{i,j} c_j^2 = \frac{1}{12}, \qquad \sum_{i=1}^{q}\sum_{j=1}^{q} b_i c_i a_{i,j} c_j = \frac{1}{8}, \qquad \sum_{i=1}^{q}\sum_{j=1}^{q}\sum_{k=1}^{q} b_i a_{i,j} a_{j,k} c_k = \frac{1}{12}.$$

3.5.3 Schémas de RK explicites

On se place maintenant dans le cas d'une méthode explicite, avec donc une matrice A strictement triangulaire inférieure. On a alors en particulier

$$c_1 = 0, \qquad t_{n,1} = t_n, \qquad u_{n,1} = u_n. \tag{3.21}$$

La méthode de Runge-Kutta est alors définie par l'algorithme

$$\begin{cases} \left|\begin{array}{l} t_{n,1} = t_n \\ u_{n,1} = u_n \\ p_{n,1} = f(t_{n,1}, u_{n,1}) \end{array}\right. \\ \text{Pour } 2 \leq i \leq q \quad \left|\begin{array}{l} t_{n,i} = t_n + \Delta t c_i \\ u_{n,i} = u_n + \Delta t \displaystyle\sum_{1 \leq j < i} a_{i,j} p_{n,j} \\ p_{n,i} = f(t_{n,i}, u_{n,i}) \end{array}\right. \\ t_{n+1} = t_n + \Delta t \\ u_{n+1} = u_n + \Delta t \left[\displaystyle\sum_{i=1}^{q} b_i p_{n,i}\right], \end{cases}$$

et on la représente conventionnellement par le tableau, dit *de Butcher*, suivant :

$$\begin{array}{c|cccc} c_1 & 0 & 0 & \dots & 0 & 0 \\ c_2 & a_{2,1} & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \dots & 0 & 0 \\ c_q & a_{q,1} & a_{q,2} & \dots & a_{q,q-1} & 0 \\ \hline & b_1 & b_2 & \dots & b_{q-1} & b_q \end{array} \tag{3.22}$$

Runge-Kutta 2
Pour $q = 2$, d'après (3.17), (3.21) et le Théorème 3.29 on doit avoir $b_1 + b_2 = 1$, $a_{2,1} = c_2$ et $b_2 c_2 = \frac{1}{2}$. Le tableau de Butcher s'écrit donc (en notant $\alpha = c_2 \in [0, 1]$)

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline & 1 - \frac{1}{2\alpha} & \frac{1}{2\alpha} \end{array}$$

- Pour $\alpha = \frac{1}{2}$ on obtient

$$\begin{cases} p_{n,1} = f(t_n, u_n) \\ t_{n,2} = t_n + \dfrac{\Delta t}{2} \\ u_{n,2} = u_n + \dfrac{\Delta t}{2} p_{n,1} \\ p_{n,2} = f(t_{n,2}, u_{n,2}) \\ t_{n+1} = t_n + \Delta t \\ u_{n+1} = u_n + \Delta t p_{n,2}, \end{cases}$$

  c'est-à-dire

$$u_{n+1} = u_n + \Delta t f\left(t_n + \frac{\Delta t}{2}, u_n + \frac{\Delta t}{2} f(t_n, u_n)\right)$$

  On retrouve le schéma du point milieu !

- Pour $\alpha = 1$ on obtient

$$\begin{cases} p_{n,1} = f(t_n, u_n) \\ t_{n,2} = t_n + \Delta t \\ u_{n,2} = u_n + \Delta t p_{n,1} \\ p_{n,2} = f(t_{n,2}, u_{n,2}) \\ t_{n+1} = t_n + \Delta t \\ u_{n+1} = u_n + \dfrac{\Delta t}{2} \left[p_{n,1} + p_{n,2}\right], \end{cases}$$

  c'est-à-dire

$$u_{n+1} = u_n + \frac{\Delta t}{2} \left[f(t_n, u_n) + f(t_{n+1}, u_n + \Delta t f(t_n, u_n)\right]$$

  ce qui donne un schéma proche du schéma de Crank-Nicholson mais complètement explicite.

Runge-Kutta 4
Il s'agit de la méthode de Runge-Kutta "classique", qui est la méthode "reine" des méthodes

à un pas, utilisée dans de nombreux solveurs d'EDO : elle a un ordre élevé et une grande stabilité. Le tableau s'écrit

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{array} \tag{3.23}$$

et l'algorithme correspondant est

$$\begin{cases} p_{n,1} = f(t_n, u_n) \\ t_{n,2} = t_n + \dfrac{1}{2}\Delta t \\ u_{n,2} = u_n + \dfrac{1}{2}\Delta t p_{n,1} \\ p_{n,2} = f(t_{n,2}, u_{n,2}) \\ t_{n,3} = t_{n,2} \\ u_{n,3} = u_n + \dfrac{1}{2}\Delta t p_{n,2} \\ p_{n,3} = f(t_{n,3}, u_{n,3}) \\ t_{n,4} = t_n + \Delta t \\ u_{n,4} = u_n + \Delta t p_{n,3} \\ p_{n,i} = f(t_{n,4}, u_{n,4}) \\ t_{n+1} = t_{n,4} \\ u_{n+1} = u_n + \dfrac{1}{6}\Delta t \left[p_{n,1} + 2p_{n,2} + 2p_{n,3} + p_{n,4}\right]. \end{cases}$$