

# Introduction à l'analyse numérique des équations différentielles



# Introduction

Ce cours contient trois parties qui vont conduire progressivement à l'analyse numérique des équations différentielles ordinaires, tout en étant chacune d'un intérêt indépendant :

- (i) Interpolation et approximation
- (ii) Intégration et dérivation numérique
- (iii) Analyse numérique des équations différentielles

La première partie traite de l'approximation - en divers sens - des fonctions continues ou plus régulières par des fonctions polynomiales : l'interpolation est une manière particulièrement simple pour essayer d'approcher une fonction par des polynômes mais ses liens avec l'approximation ne sont pas aussi simples que l'on pourrait naïvement le penser.

La seconde partie est consacrée quasi-exclusivement aux méthodes de calculs approchés d'intégrales : méthode des rectangles, des trapèzes ou de Simpson. Les liens entre ces méthodes et l'interpolation sont mis en valeur.

Enfin la troisième partie décrit les éléments de base servant au calcul approché des solutions d'équations différentielles (surtout en dimension 1 d'espace) et les notions importantes pour les schémas numériques (consistance, stabilité,...). Les liens avec les calculs approchés d'intégrales sont décrits.



# Table des matières

<b>1</b>	<b>Interpolation et approximation</b>	<b>7</b>
1.1	Interpolation de Lagrange : existence et unicité du polynôme d'interpolation . . . . .	8
1.2	Interpolation de Lagrange : stabilité et approximation . . . .	8
1.3	Interpolation de Lagrange : erreur d'interpolation dans le cas de fonctions plus régulières . . . . .	12
1.4	Interpolation de Lagrange : calcul pratique du polynôme d'interpolation . . . . .	13
1.4.1	La méthode des différences divisées . . . . .	14
1.4.2	Algorithme de Hörner . . . . .	15
1.4.3	La méthode des différences finies . . . . .	16
1.5	Polynômes de meilleures approximations . . . . .	17
1.5.1	Polynômes de meilleure approximation uniforme . . .	17
1.5.2	Polynômes de meilleure approximation quadratique . .	19
1.6	Exercices . . . . .	23
<b>2</b>	<b>Intégration et dérivation numérique</b>	<b>27</b>
2.1	La méthode classique d'intégration numérique : Newton-Cotes	27
2.1.1	Présentation de la méthode . . . . .	27
2.1.2	Stabilité . . . . .	30
2.1.3	Estimations d'erreur . . . . .	31
2.1.4	Méthode d'accélération de Romberg . . . . .	36
2.2	Un choix optimal de points : les points de Gauss-Legendre . .	36
2.3	Dérivation numérique . . . . .	38
2.4	Exercices . . . . .	39
<b>3</b>	<b>Analyse numérique des équations différentielles</b>	<b>41</b>
3.1	Rappels théoriques . . . . .	41
3.1.1	Théorie générale . . . . .	41
3.1.2	Effets des perturbations . . . . .	43
3.1.3	Régularité de la solution . . . . .	46
3.2	La méthode d'Euler . . . . .	46
3.2.1	Présentation de la méthode . . . . .	46

3.2.2	Étude de l'erreur (I) . . . . .	48
3.2.3	Étude de l'erreur (II) . . . . .	50
3.3	Étude générale des méthodes à un pas . . . . .	51
3.3.1	Propriétés importantes d'une méthode à un pas . . . .	51
3.3.2	Condition nécessaire et suffisante de consistance . . .	52
3.3.3	Condition suffisante de stabilité . . . . .	53
3.3.4	Ordre d'un schéma . . . . .	53
3.3.5	Exemples . . . . .	55
3.4	Quelques élément sur les méthodes de Runge-Kutta . . . .	56
3.5	Exercices . . . . .	58

# Chapitre 1

## Interpolation et approximation

La plupart des calculs numériques (et même théoriques) sont plus simples voire triviaux avec des polynômes : il suffit de penser à la dérivation ou au calcul de primitives. Il est donc tentant d’essayer de remplacer systématiquement une fonction continue, a priori quelconque, par une fonction polynomiale “qui lui ressemble” avec l’idée (par exemple) de remplacer la fonction continue par son approximation polynomiale dans les calculs d’intégrales, voire dans d’autres situations.

Se posent alors plusieurs questions :

1. Pour une fonction donnée, comment choisir la fonction polynomiale qui l’approche (ou qui est censée l’approcher) ? Évidemment la simplicité de la détermination de cette fonction est un des critères à prendre en compte.
2. Comment évaluer la “ressemblance”, c’est-à-dire comment mesurer le degré d’approximation par la fonction polynomiale ? Ceci signifie mathématiquement choisir une norme sur les fonctions continues.

Le cours commence par l’étude de l’interpolation qui apporte une réponse particulièrement simple à la première question ; outre l’existence, l’unicité et (bien sûr) la construction pratique de l’interpolée, on se posera la question du degré d’approximation dans différents contextes.

On s’attaquera ensuite au problème avec une autre approche, en essayant de répondre à la problématique via la deuxième question : existe-t-il des polynômes de meilleure approximation pour une fonction donnée, au moins dans le cas des normes les plus usuelles ? et peut-on les construire directement (et assez simplement) ? Sur l’espace des fonctions continues sur un intervalle  $[a, b]$  de  $\mathbb{R}$  (espace noté  $C([a, b])$ ), on considèrera d’abord la norme “naturelle” sur les fonctions continues, i.e. :

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

puis la norme  $L^2$  :

$$\|f\|_2 = \left( \int_a^b [f(x)]^2 dx \right)^{1/2},$$

où des arguments plus “euclidiens” pourront être utilisés.

## 1.1 Interpolation de Lagrange : existence et unicité du polynôme d’interpolation

**Théorème 1.1.** *Soit  $f : [a, b] \rightarrow \mathbb{R}$  une fonction continue et  $x_0, x_1, \dots, x_n$  ( $n + 1$ ) points distincts de  $[a, b]$ . Il existe une unique fonction polynomiale  $p_n$  de degré inférieur ou égal à  $n$  telle que :*

$$p_n(x_i) = f(x_i) \quad \text{pour tout } i = 0, 1, \dots, n.$$

De plus,  $p_n$  est donné par :

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

où les polynômes  $L_i$  sont définis par :

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

**Preuve :** On note  $\mathcal{P}_n$  l’espace des polynômes de degré inférieur ou égal à  $n$  ;  $\mathcal{P}_n$  est un espace vectoriel de dimension  $n + 1$  (rappel :  $(1, X, X^2, \dots, X^n)$  est la base canonique de  $\mathcal{P}_n$ ).

On considère l’application  $\psi : \mathcal{P}_n \rightarrow \mathbb{R}^{n+1}$  qui, à  $p \in \mathcal{P}_n$ , associe  $(p(x_0), p(x_1), \dots, p(x_n))$ . Cette application est linéaire et l’énoncé du théorème peut être reformulé en disant que  $\psi$  est bijective. Comme  $\psi$  est linéaire et que  $\dim(\mathcal{P}_n) = \dim(\mathbb{R}^{n+1}) = n + 1$ , le Théorème du rang montre qu’il suffit de prouver que  $\psi$  est injective, c’est-à-dire  $\ker(\psi) = \{0\}$ .

Si  $p \in \ker(\psi)$  alors  $\psi(p) = 0$  (le 0 de  $\mathbb{R}^{n+1}$ , bien sûr) et donc  $p(x_0) = 0$ ,  $p(x_1) = 0, \dots, p(x_n) = 0$ . Mais on sait qu’un polynôme non nul de degré inférieur ou égal à  $n$  ne peut avoir qu’au plus  $n$  racines. Donc  $p$  est le polynôme nul,  $\psi$  est injective et donc bijective.  $\square$

## 1.2 Interpolation de Lagrange : stabilité et approximation

En pratique, on commet systématiquement des erreurs car un ordinateur ne travaille qu’avec un nombre limité de chiffres significatifs. Il est donc



important de connaître l'influence sur le résultat final des erreurs commises sur les données. Ici, si on remplace (involontairement) les vraies valeurs  $f(x_i)$  par des valeurs approchées  $f_i$ , quelle est l'incidence sur  $p_n$  ?

Dans ce cas,  $p_n$  est remplacé par  $\tilde{p}_n$  donné par :

$$\tilde{p}_n(x) = \sum_{i=0}^n f_i L_i(x) ,$$

et l'erreur commise est estimée par :

$$\begin{aligned} |\tilde{p}_n(x) - p_n(x)| &= \left| \sum_{i=0}^n (f_i - f(x_i)) L_i(x) \right| \\ &\leq \sum_{i=0}^n |f_i - f(x_i)| |L_i(x)| \\ &\leq \max_i |f_i - f(x_i)| \sum_{i=0}^n |L_i(x)| . \end{aligned}$$

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

$$\|\tilde{p}_n - p_n\|_\infty \leq \Lambda_n \max_i |f_i - f(x_i)| .$$

L'erreur commise sur les  $f(x_i)$  est donc amplifiée (ou atténuée) par la constante  $\Lambda_n$ , appelée constante de Lebesgue associée aux points  $x_0, x_1, \dots, x_n$ . On va voir que cette constante joue un rôle important dans la mesure de l'approximation par interpolation.

Pour aller dans le sens de cette analyse, nous commençons par la :

**Proposition 1.1.** *On introduit l'application linéaire  $\mathcal{L}_n : C([a, b]) \rightarrow \mathcal{P}_n$  qui, à  $f \in C([a, b])$ , associe son unique polynôme d'interpolation de Lagrange aux points  $x_0, x_1, \dots, x_n$ . Alors la norme de l'application linéaire  $\mathcal{L}_n$  est  $\Lambda_n$ . En d'autres termes :*

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

**Preuve :** On montre d'abord que  $\|\mathcal{L}_n\| \leq \Lambda_n$ . On a, par des majorations identiques à celles déjà utilisées pour estimer les erreurs commises sur le

polynôme d'interpolation :

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

Ce qui prouve l'inégalité souhaitée.

Pour obtenir l'égalité, on se demande s'il existe une fonction  $f \in C([a, b])$  telle que  $\|\mathcal{L}_n(f)\|_\infty = \Lambda_n \|f\|_\infty$ . Il n'est pas du tout sûr qu'une telle fonction existe car le "sup" dans la définition de  $\|\mathcal{L}_n\|$  n'est pas forcément atteint. Mais si c'était le cas, cela signifierait que, pour cette fonction, toutes les inégalités ci-dessus sont des égalités. Analysons les contraintes sur une telle fonction :

- (i) L'égalité entre les deux dernières lignes signifie que  $x$  est un point de maximum de la fonction  $y \mapsto \sum_{i=0}^n |L_i(y)|$ . Un tel point existe car cette fonction est continue sur  $[a, b]$ , intervalle fermé borné de  $\mathbb{R}$ .
- (ii) L'égalité entre les deux lignes précédentes signifie que  $|f(x_i)| = \|f\|_\infty$  pour tout  $i$ . Et on peut supposer que  $\|f\|_\infty = 1$ .
- (iii) L'égalité entre les deux premières lignes signifie que les  $f(x_i)L_i(x)$  ont tous le même signe et on peut supposer que ces quantités sont toutes positives.

En combinant (ii) et (iii), on voit que l'on peut prendre  $f(x_i) = 1$  si  $L_i(x) \geq 0$  et  $f(x_i) = -1$  si  $L_i(x) < 0$ . De plus, si on suppose que  $x_0 < x_1 < \dots < x_n$ , on peut choisir  $f$  affine par morceaux (c'est-à-dire affine sur chaque intervalle  $[x_i, x_{i+1}]$ ) et constante pour  $x \leq x_0$  et  $x \geq x_n$ .

On vérifie facilement qu'un tel  $f$  satisfait  $\|\mathcal{L}_n(f)\|_\infty = \Lambda_n \|f\|_\infty$ .  $\square$

Notre première estimation de l'erreur d'interpolation est donnée par le :

**Théorème 1.2.** *Pour tout  $f : [a, b] \rightarrow \mathbb{R}$ , on a :*

$$\|f - \mathcal{L}_n(f)\|_\infty \leq (1 + \Lambda_n) d(f, \mathcal{P}_n) ,$$

où :

$$d(f, \mathcal{P}_n) = \inf_{q \in \mathcal{P}_n} \|f - q\|_\infty .$$

Ce théorème est assez ambigu car on verra plus loin que  $\Lambda_n$  tend vers  $+\infty$  quand  $n \rightarrow +\infty$  alors que, par le Théorème de Weierstass ("toute fonction continue sur  $[a, b]$  est limite uniforme de polynômes"), on a  $d(f, \mathcal{P}_n) \rightarrow 0$

quand  $n \rightarrow +\infty$ . On a donc une “forme indéterminée” et on verra plus loin qu’elle cache une vraie difficulté.

**Preuve :** Par unicité du polynôme d’interpolation (cf. Théorème 1.1),  $\mathcal{L}_n(q) = q$  pour tout  $q \in \mathcal{P}_n$  ; en effet, dans ces conditions,  $q$  est son propre polynôme d’interpolation. On écrit alors :

$$\begin{aligned} \|f - \mathcal{L}_n(f)\|_\infty &= \|f - q + \mathcal{L}_n(q) - \mathcal{L}_n(f)\|_\infty \\ &\leq \|f - q\|_\infty + \|\mathcal{L}_n(q - f)\|_\infty \\ &\leq \|f - q\|_\infty + \Lambda_n \|f - q\|_\infty, \end{aligned}$$

et le résultat en découle en prenant simplement l’infimum sur  $q \in \mathcal{P}_n$ .  $\square$

**Exemple 1.1.** On considère dans cet exemple les deux types de points d’interpolation les plus classiques :

- Les points équidistants : dans ce cas :

$$x_i = a + i \frac{b-a}{n} \quad \text{pour } i = 0, 1, \dots, n.$$

La constante de Lebesgue sur l’intervalle  $[-1, 1]$  est évalué par :

$$\Lambda_n \sim \frac{2^{n+1}}{n \log(n)}.$$

- Les points de Chebyshev : dans ce cas :

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

La constante de Lebesgue sur l’intervalle  $[-1, 1]$  est évalué par :

$$\Lambda_n \sim \frac{2}{\pi} \log(n).$$

L’ordre de grandeur dans les deux cas n’est donc pas du tout le même : la constante de Lebesgue est beaucoup plus petite dans le cas des points de Chebyshev.

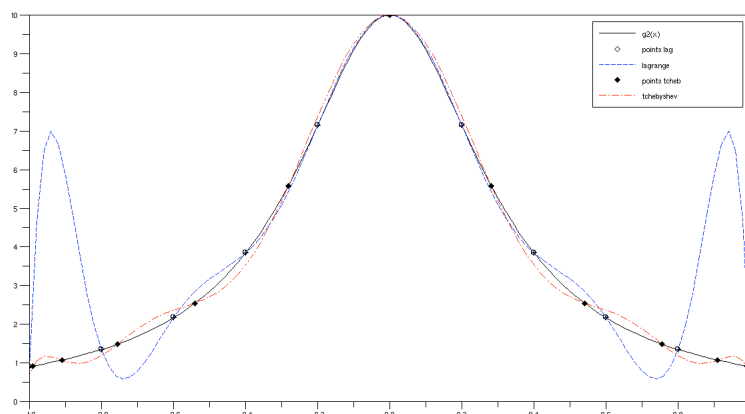
**Effet Runge :** dans le cas de fonctions du type :

$$g_a(x) = \frac{1}{1+a^2x^2},$$

ou :

$$h_a(x) = \frac{1}{a^2+x^2},$$

on constate que l’interpolation via les points équidistants ne donne pas un résultat proche de la fonction interpolée : en particulier des oscillations de grandes amplitudes apparaissent sur le bord de l’intervalle, c’est l’*effet Runge*. On voit donc qu’en général, l’interpolation n’est pas une bonne méthode d’approximation.



L'effet Runge

### 1.3 Interpolation de Lagrange : erreur d'interpolation dans le cas de fonctions plus régulières

Il est naturel de penser que l'approximation devrait être meilleure dans le cas d'une fonction régulière, même si l'effet Runge a lieu dans le cas d'une fonction très régulière. En tout cas, des propriétés convenables sur les dérivées successives d'une fonction devrait conduire à une formule d'erreur. C'est l'objet du résultat suivant :

**Théorème 1.3.** Soit  $f : [a, b] \rightarrow \mathbb{R}$  une fonction  $(n + 1)$ -fois dérivable sur  $]a, b[$  telle que  $f, f', \dots, f^{(n)}$  sont continues<sup>(1)</sup> sur  $[a, b]$ . Alors, pour tout  $x \in [a, b]$ , il existe  $\xi_x \in ]a, b[$  tel que :

$$f(x) - p_n(x) = \frac{1}{(n + 1)!} \Pi_{n+1}(x) f^{(n+1)}(\xi_x),$$

où  $\Pi_{n+1}(x) = (x - x_0) \cdots (x - x_n)$ .

On voit dans ce résultat que l'erreur d'interpolation dépend des valeurs des dérivées de  $f$  : c'est assez intuitif car plus une fonction oscille, plus il est difficile de l'approcher.

Elle dépend aussi des points via le polynôme  $\Pi_{n+1}$  : le corollaire philosophique de ce résultat, c'est que la meilleure interpolation avec  $(n + 1)$  points est celle pour laquelle la valeur de  $\|\Pi_{n+1}\|_\infty$  est minimale.

**Preuve :** On peut d'abord supposer que  $x$  n'est pas l'un des  $x_i$  car sinon les deux membres de l'égalité sont nuls, donc le résultat est acquis.

(1). ou plutôt se prolonge par continuité...

On va appliquer le Théorème de Rolle par récurrence : pour cela, on considère la fonction :

$$\psi(y) := f(y) - p_n(y) - A \cdot \Pi_{n+1}(y) ,$$

où  $A$  est une constante que nous allons choisir. Cette fonction s'annule en chacun des points  $x_i$  par définition de  $p_n$  et  $\Pi_{n+1}$  et nous prenons  $A$  tel que  $\psi(x) = 0$ , i.e.

$$A = \frac{f(x) - p_n(x)}{\Pi_{n+1}(x)} .$$

$A$  est bien défini puisque  $x$  n'est pas l'un des  $x_i$ , donc  $\Pi_{n+1}(x) \neq 0$ .

Grâce à ce choix, la fonction  $\psi$  s'annule au moins  $(n+2)$  fois dans  $[a, b]$  et en appliquant soigneusement le Théorème de Rolle, on voit que  $\psi'$  s'annule au moins  $(n+1)$  fois dans  $]a, b[$ . On procède alors par récurrence pour prouver que  $\psi''$  s'annule au moins  $n$  fois dans  $]a, b[$  et que, pour  $k \leq n+1$ ,  $\psi^{(k)}$  s'annule au moins  $(n+2-k)$  fois dans  $]a, b[$ . En particulier,  $\psi^{(n+1)}$  s'annule au moins une fois dans  $]a, b[$ . Mais :

$$\psi^{(n+1)}(y) := f^{(n+1)}(y) - p_n^{(n+1)}(y) - A \Pi_{n+1}^{(n+1)}(y) ,$$

et  $p_n^{(n+1)}(y) \equiv 0$  puisque  $p_n$  est de degré au plus  $n$  alors que  $\Pi_{n+1}^{(n+1)} \equiv (n+1)!$  car  $\Pi_{n+1}$  est de degré  $n+1$  et le coefficient de  $x^{n+1}$  dans  $\Pi_{n+1}$  est 1.

Il en résulte que, si on note  $\xi_x \in ]a, b[$  le point d'annulation de  $\psi^{(n+1)}$ , on a :

$$0 = f^{(n+1)}(\xi_x) - 0 - A(n+1)! .$$

D'où la conclusion en comparant les deux expressions de  $A$ . □

## 1.4 Interpolation de Lagrange : calcul pratique du polynôme d'interpolation

Comme remarque préliminaire, nous rappelons la méthode de calcul de  $p(x)$  si  $p$  est un polynôme de la forme :

$$p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n .$$

La méthode consistant à calculer chaque terme indépendamment conduit à  $(n-1)$  multiplications pour calculer successivement les  $x^k$  puis encore  $(n-1)$  multiplications pour effectuer les produits avec les  $a_k$  et  $n$  additions : soit  $2(n-1)$  multiplications et  $n$  additions.

On peut aussi calculer de la manière suivante :  $u_0 = a_nx + a_{n-1}$ ,  $u_1 = xu_0 + a_{n-2}$ ,  $u_2 = xu_1 + a_{n-3}$ ...etc. On se convainc facilement que  $u_{n-1} = p(x)$  et on fait seulement  $n$  multiplications et  $n$  additions. Donc on gagne un facteur 2 au niveau des multiplications (plus coûteuses au niveau informatique...).

Le même principe va être utilisée par la méthode des différences divisées.

### 1.4.1 La méthode des différences divisées

On va calculer successivement  $p_0, p_1, \dots, p_n$  où  $p_k \in \mathcal{P}_k$  est le polynôme d'interpolation de  $f$  associé à  $x_0, \dots, x_k$ .

Il est clair que  $p_0(x) = f(x_0)$  car  $p_0$  est de degré 0. On note ensuite  $f[x_0, \dots, x_k]$  le coefficient de plus haut degré du polynôme  $p_k$ . On remarque alors que l'on a :

$$p_k(x) = p_{k-1}(x) + f[x_0, \dots, x_k](x - x_0) \cdots (x - x_{k-1}) .$$

En effet, le polynôme  $q(x) = p_k(x) - f[x_0, \dots, x_k](x - x_0) \cdots (x - x_{k-1})$  est de degré au plus  $k - 1$  car on a retiré à  $p_k$  son monôme de plus haut degré et, pour  $j = 0, \dots, k - 2$ ,  $q(x_j) = p_k(x_j) = f(x_j)$ . Donc, par unicité du polynôme d'interpolation,  $q = p_{k-1}$ .

On obtient donc par récurrence :

$$p_n(x) = f(x_0) + \sum_{k=1}^n f[x_0, \dots, x_k](x - x_0) \cdots (x - x_{k-1}) .$$

C'est la *Formule de Newton*.

Reste à calculer les  $f[x_0, \dots, x_k]$  de manière efficace, c'est l'objet du :

**Lemme 1.1.** *Pour tout  $k \geq 1$  :*

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

L'utilité du lemme est claire : dès que l'on sait calculer des  $f[\dots]$  à  $k$  termes, on calcule facilement les  $f[\dots]$  à  $k + 1$  termes.

**Preuve :** On note  $q_{k-1} \in \mathcal{P}_{k-1}$  l'unique polynôme d'interpolation aux points  $x_1, \dots, x_k$  ; par définition, le coefficient du terme dominant (qui est  $x^{k-1}$ ) est  $f[x_1, \dots, x_k]$ . On considère :

$$\tilde{p}_k(x) := \frac{(x - x_0)q_{k-1}(x) - (x - x_k)p_{k-1}(x)}{x_k - x_0} .$$

Comme  $p_{k-1}, q_{k-1} \in \mathcal{P}_{k-1}$ , il est clair que  $\tilde{p}_k \in \mathcal{P}_k$ . De plus, par les définitions de  $p_{k-1}, q_{k-1}$  :

$$\begin{aligned} \tilde{p}_k(x_0) &= \frac{-(x_0 - x_k)p_{k-1}(x_0)}{x_k - x_0} = f(x_0) , \\ \tilde{p}_k(x_k) &= \frac{(x_k - x_0)q_{k-1}(x_k)}{x_k - x_0} = f(x_k) . \end{aligned}$$

et pour  $1 \leq j \leq k - 1$  :

$$\begin{aligned} \tilde{p}_k(x_j) &= \frac{(x_j - x_0)q_{k-1}(x_j) - (x_j - x_k)p_{k-1}(x_j)}{x_k - x_0} \\ &= \frac{(x_j - x_0)f(x_j) - (x_j - x_k)f(x_j)}{x_k - x_0} \\ &= f(x_j) \frac{(x_j - x_0) - (x_j - x_k)}{x_k - x_0} = f(x_j) . \end{aligned}$$

Donc, par l'unicité du polynôme d'interpolation aux points  $x_0, \dots, x_k$  dans  $\mathcal{P}_k$ ,  $\tilde{p}_k = p_k$  et la conclusion en découle en examinant le coefficient de  $x^k$ .  $\square$

**Remarque :** Le lien montré ci-dessus entre  $p_k$  et  $p_{k-1}$  permet aussi de déduire que  $p_n(y) + f[x_0, \dots, x_n, x]\Pi_{n+1}(y)$  est l'unique polynôme d'interpolation de  $f$  associé aux points  $x_0, \dots, x_n, x$ , d'où :

$$f(x) = p_n(x) + f[x_0, \dots, x_n, x]\Pi_{n+1}(x) .$$

Ceci nous montre, via l'estimation d'erreur, que si  $f$  est assez régulière :

$$f[x_0, \dots, x_n, x] = \frac{1}{(n+1)!} f^{(n+1)}(\xi_x) .$$

Ceci donne une estimation des  $f[\dots]$ .

### 1.4.2 Algorithme de Hörner

Il s'agit maintenant de calculer les  $p_k$  en pratique. Deux étapes : le calcul des  $f[\dots]$  puis la déduction des  $p_k$  via la formule de Newton.

Comme le montre le tableau suivant, on utilise le Lemme 1.1 pour déterminer la valeur de tous les  $f[x_0, \dots, x_k]$ , obtenus en bout de lignes et notés  $T[k]$  :

Tableau					
$T[0]$	$f(x_0)$				
$T[1]$	$f(x_1)$	$\searrow$	$f[x_0, x_1]$		
$T[2]$	$f(x_2)$	$\searrow$	$f[x_1, x_2]$	$\searrow$	$f[x_0, x_1, x_2]$
$\vdots$					
$\vdots$					
$T[n-2]$	$f(x_{n-2})$	$\searrow$	$f[x_{n-3}, x_{n-2}]$	$\searrow$	$f[x_{n-4}, x_{n-3}, x_{n-2}] \dots$
$T[n-1]$	$f(x_{n-1})$	$\searrow$	$f[x_{n-2}, x_{n-1}]$	$\searrow$	$f[x_{n-3}, x_{n-2}, x_{n-1}] \dots$
$T[n]$	$f(x_n)$	$\searrow$	$f[x_{n-1}, x_n]$	$\searrow$	$f[x_{n-2}, x_{n-1}, x_n] \dots$

L'algorithme de Hörner consiste alors à suivre la procédure suivante qui ressemble à celle présentée pour le calcul de la valeur d'un polynôme en un

point :

$$\begin{aligned}
 U[n] &= T[n] \\
 U[n-1] &= T[n-1] + (x - x_{n-1})U[n] \\
 &\vdots \\
 U[k] &= T[k] + (x - x_k)U[k+1] \\
 &\vdots \\
 U[0] &= p_n(x)
 \end{aligned}$$

### 1.4.3 La méthode des différences finies

Lorsque les points sont équi-distants, les différences divisées deviennent des "différences finies" :

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

où  $h$  est la valeur commune des quantités  $x_{i+1} - x_i$ . Si on définit par récurrence :

$$\nabla f_i = f(x_{i+1}) - f(x_i) \quad \text{et} \quad \nabla^k f_i = \nabla(\nabla^{k-1} f_i) = \nabla^{k-1} f_{i+1} - \nabla^{k-1} f_i,$$

on a des formules plus simples :

$$\nabla^k f_i = f[x_i, \dots, x_{i+k}] k! h^k,$$

et :

$$p_n(x) = f_0 + s\nabla f_0 + s(s-1)\nabla^2 f_0 + \dots + \frac{s(s-1)\dots(s-n+1)}{n!} \nabla^n f_0,$$

où :

$$s = \frac{x-a}{h} \quad \text{avec} \quad h = \frac{b-a}{n}.$$

(NB :  $0 \leq s \leq n$ )

En effet :

$$\begin{aligned}
 (x - x_0) \dots (x - x_k) &= sh(sh - h)(sh - 2h) \dots (sh - kh) \\
 &= s(s-1)(s-2) \dots (s-k) h^{k+1}.
 \end{aligned}$$

On obtient alors le polynôme d'interpolation via la formule de Newton en appliquant l'algorithme précédent.



## 1.5 Polynômes de meilleures approximations

Nous commençons par un résultat général :

**Théorème 1.4.** *Soit  $\|\cdot\|$  une norme QUELCONQUE sur  $C([a, b])$ . Pour tout  $f \in C([a, b])$  et pour tout  $n \in \mathbb{N}$ , il existe au moins un polynôme  $\tilde{p}_n \in \mathcal{P}_n$  tel que :*

$$\|f - \tilde{p}_n\| = \min_{q \in \mathcal{P}_n} \|f - q\|.$$

Un résultat aussi général peut paraître surprenant car, d'une part, la norme est quelconque et, d'autre part, on semble minimiser une fonction en dimension infinie, ce qui est a priori délicat. En fait, comme on verra mieux dans la preuve, on minimise en dimension finie (car l'espace qui joue vraiment un rôle est  $\mathcal{P}_n$ ) et seule la continuité de la norme sera importante.

**Preuve :** On considère l'application  $\psi : \mathcal{P}_n \rightarrow \mathbb{R}$  définie par  $\psi(q) = \|f - q\|$  si  $q \in \mathcal{P}_n$ . L'espace  $\mathcal{P}_n$  étant de dimension finie, pour prouver que  $\psi$  atteint son minimum, il suffit de montrer que  $\psi$  est continue et coercive.

• **Continuité :** Grâce à la deuxième inégalité triangulaire, on a :

$$\begin{aligned} |\psi(q_1) - \psi(q_2)| &= \left| \|f - q_1\| - \|f - q_2\| \right| \\ &= \|q_1 - q_2\| \end{aligned}$$

ce qui donne la continuité de  $\psi$  (par rapport à la norme  $\|\cdot\|$  mais comme en dimension finie, toutes les normes sont équivalentes...).

• **Coercivité :** Toujours grâce à la deuxième inégalité triangulaire, on a :

$$\psi(q) = \|f - q\| \geq \|q\| - \|f\|$$

et donc  $\psi(q) \rightarrow +\infty$  quand  $\|q\| \rightarrow +\infty$ .

□

**Remarque :** (i) comme toutes les normes ne sont pas équivalentes sur  $C([a, b])$ , il n'est pas clair que  $\|f - \tilde{p}_n\| \rightarrow 0$  quand  $n \rightarrow +\infty$ , malgré le Théorème de Weierstrass...

(ii) La question de l'unicité est non triviale, en général...

### 1.5.1 Polynômes de meilleure approximation uniforme

On introduit d'abord la définition suivante :

**Définition 1.1.** *On dit qu'une fonction  $g \in C([a, b])$  équioscille sur  $k + 1$  points de  $[a, b]$  s'il existe des points  $x_0 < x_1 < \dots < x_k$  de  $[a, b]$  tels que  $|g(x_i)| = \|g\|_\infty$  pour tout  $0 \leq i \leq k$  et tels que  $g(x_{i+1}) = -g(x_i)$  pour tout  $0 \leq i \leq k - 1$ .*

Cette notion nous permet de prouver le :

**Théorème 1.5.** *Pour tout  $f \in C([a, b])$  et pour tout  $n \in \mathbb{N}$ , il existe un UNIQUE polynôme  $\tilde{p}_n \in \mathcal{P}_n$  tel que :*

$$\|f - \tilde{p}_n\|_\infty = \min_{q \in \mathcal{P}_n} \|f - q\|_\infty .$$

*De plus,  $\tilde{p}_n$  est l'unique polynôme de  $\mathcal{P}_n$  tel que  $f - \tilde{p}_n$  équioscille sur au moins  $n + 2$  points.*

Dans ce théorème, c'est évidemment l'unicité qui est le point important, avec la caractérisation de  $\tilde{p}_n$ .

**Preuve :** Nous n'allons pas faire la preuve complète mais seulement montrer pourquoi la propriété d'équioscillation intervient, sachant que l'existence de  $\tilde{p}_n$  est assurée par le résultat général.

On construit des points  $x_0, x_1, \dots, x_k$  de la manière suivante :

$$x_0 = \min\{t \in [a, b] ; |(f - \tilde{p}_n)(t)| = \|f - \tilde{p}_n\|_\infty\} ,$$

puis par récurrence :

$$x_{i+1} = \min\{t \in [x_i, b] ; |(f - \tilde{p}_n)(t)| = -(f - \tilde{p}_n)(x_i)\} .$$

Si  $f - \tilde{p}_n$  n'équioscille pas sur au moins  $n + 2$  points, cette construction s'arrête à  $x_k$  avec  $k \leq n$ . On choisit alors dans l'intervalle  $[x_i, x_{i+1}]$ , le plus grand réel  $c_i$  tel que  $(f - \tilde{p}_n)(c_i) = 0$  ( $0 \leq i \leq n - 1$ ) et on note :

$$\Pi(x) = \prod_{i=0}^{k-1} (x - c_i) .$$

Il est important de remarquer que  $\Pi \in \mathcal{P}_n$  puisque  $k \leq n$ .

Pour  $\varepsilon > 0$  petit, on examine alors la fonction  $f - \tilde{p}_n \pm \varepsilon \Pi$ , le  $\pm$  étant dicté par le signe de  $(f - \tilde{p}_n)(x_k)$  : on prend “−” si  $(f - \tilde{p}_n)(x_k) > 0$ , “+” sinon.

Commençons justement par examiner  $(f - \tilde{p}_n - \varepsilon \Pi)(x_k)$  en supposant  $(f - \tilde{p}_n)(x_k) > 0$ . On remarque que  $\Pi(x_k) > 0$  car tous les  $x_k - c_i$  sont strictement positifs et donc, si  $\varepsilon$  est choisi suffisamment petit,

$$0 < (f - \tilde{p}_n - \varepsilon \Pi)(x_k) < (f - \tilde{p}_n)(x_k) = \|f - \tilde{p}_n\|_\infty .$$

En examinant ensuite  $x_{k-1}$ , on se rend compte que  $(f - \tilde{p}_n)(x_{k-1})$  a changé de signe puisqu'il vaut  $-\|f - \tilde{p}_n\|_\infty$  mais  $\Pi(x_{k-1}) < 0$  car seul le terme  $(x - c_{k-1})$  a changé de signe dans  $\Pi$ . On a cette fois :

$$0 > (f - \tilde{p}_n - \varepsilon \Pi)(x_{k-1}) > (f - \tilde{p}_n)(x_{k-1}) = -\|f - \tilde{p}_n\|_\infty .$$

Et par récurrence, on voit que la valeur de  $f - \tilde{p}_n$  s'est trouvée diminuée par la soustraction de  $\varepsilon \Pi$  en chacun des points où la norme infinie est atteinte.

En raisonnant plus précisément, on voit que, pour  $\varepsilon$  assez petit,

$$\|f - \tilde{p}_n - \varepsilon \Pi\|_\infty < \|f - \tilde{p}_n\|_\infty,$$

ce qui contredit la définition de  $\tilde{p}_n$  et donc  $f - \tilde{p}_n$  équi oscille sur au moins  $n + 2$  points.  $\square$

### 1.5.2 Polynômes de meilleure approximation quadratique

Pour mettre cette section dans un cadre un peu plus général, on choisit une norme  $L^2$  “à poids”, c’est-à-dire que la norme découle d’un produit scalaire de la forme :

$$(f, g)_\omega = \int_a^b f(t)g(t)\omega(t)dt ,$$

où  $\omega \in L^1([a, b])$  est une fonction strictement positive presque partout. La norme  $L^2$  associée est donc :

$$\|f\|_{2,\omega} = [(f, f)_\omega]^{1/2} .$$

Dans la suite, on remplacera l’indice “ $2, \omega$ ” par un indice “2” pour la norme et on l’omettra dans le produit scalaire, pour avoir des notations plus légères, tout en rappelant qu’il s’agit d’une norme de type “2”.

**Théorème 1.6.** *Pour tout  $f \in C([a, b])$  et pour tout  $n \in \mathbb{N}$ , il existe un UNIQUE polynôme  $\tilde{p}_n \in \mathcal{P}_n$  tel que :*

$$\|f - \tilde{p}_n\|_2 = \min_{q \in \mathcal{P}_n} \|f - q\|_2 .$$

*De plus,  $\tilde{p}_n$  est la projection orthogonale de  $f$  sur  $\mathcal{P}_n$ .*

**Preuve :** Une nouvelle fois, l’existence de  $\tilde{p}_n$  est donné par le résultat général et seule l’unicité nous intéresse ainsi que la caractérisation de  $\tilde{p}_n$ .

Par définition de  $\tilde{p}_n$ , on a :

$$\|f - \tilde{p}_n\|_2^2 \leq \|f - q\|_2^2 ,$$

pour tout  $q \in \mathcal{P}_n$  ; mais, puisque la norme découle d’un produit scalaire, en écrivant  $f - q = f - \tilde{p}_n + \tilde{p}_n - q$ , on a aussi :

$$\|f - q\|_2^2 = \|f - \tilde{p}_n\|_2^2 + 2(f - \tilde{p}_n, \tilde{p}_n - q) + \|\tilde{p}_n - q\|_2^2 .$$

On en déduit que :

$$2(f - \tilde{p}_n, \tilde{p}_n - q) + \|\tilde{p}_n - q\|_2^2 \geq 0 ,$$

pour tout  $q \in \mathcal{P}_n$ . Mais  $\mathcal{P}_n$  étant un espace vectoriel, on peut aussi le réécrire :

$$2(f - \tilde{p}_n, \tilde{q}) + \|\tilde{q}\|_2^2 \geq 0 ,$$

en remplaçant  $q$  par  $\tilde{p}_n - \tilde{q}$  dans l'inégalité ci-dessus.

On remplace enfin  $\tilde{q}$  par  $t\tilde{q}$  pour  $t > 0$ , ce qui conduit à :

$$2t(f - \tilde{p}_n, \tilde{q}) + t^2\|\tilde{q}\|_2^2 \geq 0 ,$$

pour tout  $t > 0$ . En divisant par  $2t$  puis en faisant tendre  $t$  vers 0, on aboutit à :

$$(f - \tilde{p}_n, \tilde{q}) \geq 0 ,$$

pour tout  $\tilde{q} \in \mathcal{P}_n$ . Il suffit alors de changer  $\tilde{q}$  en  $-\tilde{q}$  pour avoir le résultat attendu, i.e. :

$$(f - \tilde{p}_n, \tilde{q}) = 0 ,$$

pour tout  $\tilde{q} \in \mathcal{P}_n$ .

L'unicité en résulte car, si  $\tilde{p}_n^1, \tilde{p}_n^2$  sont deux polynômes de meilleure approximation quadratique alors :

$$(f - \tilde{p}_n^1, \tilde{q}) = 0 ,$$

$$(f - \tilde{p}_n^2, \tilde{q}) = 0 ,$$

pour tout  $\tilde{q} \in \mathcal{P}_n$ . En soustrayant ces deux égalités, on a :

$$(\tilde{p}_n^1 - \tilde{p}_n^2, \tilde{q}) = 0 ,$$

pour tout  $\tilde{q} \in \mathcal{P}_n$ . D'où  $\tilde{p}_n^1 = \tilde{p}_n^2$  en prenant  $\tilde{q} = \tilde{p}_n^1 - \tilde{p}_n^2$ .

On peut aussi avoir l'unicité via le théorème de Pythagore qui est une conséquence immédiate de ce qui précède, i.e. :

$$\|f - q\|_2^2 = \|f - \tilde{p}_n\|_2^2 + \|\tilde{p}_n - q\|_2^2 ,$$

et qui montre que :

$$\|f - q\|_2^2 > \|f - \tilde{p}_n\|_2^2 ,$$

si  $q \neq \tilde{p}_n$ . □

### Calcul de $\tilde{p}_n$ :

Comme  $\tilde{p}_n \in \mathcal{P}_n$ , il suffit de calculer ses coordonnées dans une base et, vu le contexte euclidien, il est naturel de penser qu'il faut choisir une base mieux adaptée que la base canonique  $(1, X, \dots, X^n)$ . Les calculs seront bien plus simples si on possède une base orthogonale (ou même orthonormée) de  $\mathcal{P}_n$  pour le produit scalaire  $(\cdot, \cdot)$  comme on le montre maintenant.

Si on note  $(p_k)_{0 \leq k \leq n}$  une telle base alors :

$$\tilde{p}_n = \sum_{k=0}^n a_k p_k ,$$

et en faisant le produit scalaire avec l'un des  $p_i$ , on a :

$$(\tilde{p}_n, p_i) = \sum_{k=0}^n a_k (p_k, p_i) = a_i (p_i, p_i) .$$

Mais comme on a vu dans la preuve précédente que  $(f, q) = (\tilde{p}_n, q)$  pour tout  $q \in \mathcal{P}_n$ , il en résulte que :

$$a_i = \frac{(f, p_i)}{\|p_i\|_2^2} .$$

Donc le calcul de  $\tilde{p}_n$  se résume à celui de  $2(n+1)$  intégrales car aussi bien les produits scalaires que les normes (au carré) sont des intégrales. En fait de calculs, il s'agit le plus souvent d'évaluations des intégrales car on ne sait pas les calculer explicitement.

L'existence d'une base de polynômes orthogonaux est assurée par le résultat suivant où les polynômes  $p_k$  sont pris "moniques", c'est-à-dire le coefficient de leur monôme de plus haut degré (qui sera  $X^k$  en l'occurrence) sera pris égal à 1.

**Théorème 1.7.** *Il existe une suite de polynômes  $p_0, p_1, \dots, p_k, \dots$  et une seule vérifiant :*

1. *Pour tout  $k$ , le degré de  $p_k$  est  $k$  et  $p_k$  est monique.*
2. *Pour tous  $k \neq j$ ,  $(p_j, p_k) = 0$  (ce qui implique que, pour tout  $k$ ,  $(p_i)_{0 \leq i \leq k}$  est une base de  $\mathcal{P}_k$  et que  $p_k$  est orthogonal à  $\mathcal{P}_{k-1}$ ).*

*De plus, cette suite de polynômes satisfait la relation de récurrence :*

$$p_k(x) = (x - \lambda_k)p_{k-1}(x) - \mu_k p_{k-2}(x),$$

avec :

$$\lambda_k = \frac{(xp_{k-1}, p_{k-1})}{\|p_{k-1}\|_2^2} , \quad \mu_k = \frac{\|p_{k-1}\|_2^2}{\|p_{k-2}\|_2^2} .$$

**Preuve :** On utilise le procédé d'orthogonalisation de Gram-Schmidt, à partir de la base canonique  $(1, x, \dots, x^n, \dots)$ .

Le polynôme  $p_0$  est forcément donné par  $p_0(x) = 1$ . Ensuite  $p_1(x) = x - a$  où la constante  $a$  est à déterminer. En utilisant le fait que  $(p_1, p_0) = 0$ , on voit que :

$$a = \frac{(x, p_0)}{\|p_0\|_2^2} .$$

Plus g n ralement, en proc dant par r currence, on contruit  $p_k$  via la formule :

$$p_k(x) = x^k - \sum_{i=0}^{k-1} a_{i,k} p_i(x) ,$$

o  les  $a_{i,k}$  sont donn s par :

$$a_{i,k} = \frac{(x^k, p_i)}{\|p_i\|_2^2} .$$

On v rifie facilement (par orthogonalit  ou par un argument de degr ) que, pour tout  $k$ ,  $(p_i)_{0 \leq i \leq k}$  est une base de  $\mathcal{P}_k$ .

Il nous reste   montrer la relation de r currence. Comme  $p_k(x) - xp_{k-1}(x) \in \mathcal{P}_{k-1}$  puisque  $p_k$  et  $p_{k-1}$  sont moniques, on peut  crire :

$$p_k(x) - xp_{k-1}(x) = \sum_{i=0}^{k-1} \alpha_i p_i(x) .$$

En faisant le produit scalaire avec  $p_j$ , on obtient pour  $0 \leq j \leq k-1$  :

$$\alpha_j = \frac{(p_k - xp_{k-1}, p_j)}{\|p_j\|_2^2} ,$$

que l'on peut r  crire, en se souvenant que le produit scalaire est une int grale :

$$\alpha_j = \frac{(p_k, p_j)}{\|p_j\|_2^2} - \frac{(p_{k-1}, xp_j)}{\|p_j\|_2^2} .$$

Puisque  $p_k$  est orthogonal    $\mathcal{P}_{k-1}$  et que  $p_{k-1}$  est orthogonal    $\mathcal{P}_{k-2}$ , il est clair que  $\alpha_i = 0$  si  $j \leq k-3$ . Donc il ne nous reste que deux termes :

$$p_k(x) - xp_{k-1}(x) = \alpha_{k-1} p_{k-1}(x) + \alpha_{k-2} p_{k-2}(x) .$$

Pour le premier :

$$\alpha_{k-1} = \frac{(p_k, p_{k-1})}{\|p_{k-1}\|_2^2} - \frac{(p_{k-1}, xp_{k-1})}{\|p_{k-1}\|_2^2} = -\frac{(p_{k-1}, xp_{k-1})}{\|p_{k-1}\|_2^2} ,$$

et on retrouve la valeur de  $-\lambda_k$ .

Pour le second :

$$\alpha_{k-2} = \frac{(p_k, p_{k-2})}{\|p_{k-2}\|_2^2} - \frac{(p_{k-1}, xp_{k-2})}{\|p_{k-2}\|_2^2} = -\frac{(p_{k-1}, xp_{k-2})}{\|p_{k-2}\|_2^2} ,$$

et en utilisant le fait que  $p_{k-1} - xp_{k-2} \in \mathcal{P}_{k-2}$ , on voit que  $(p_{k-1} - xp_{k-2}, p_{k-1}) = 0$  et donc  $(p_{k-1}, xp_{k-2}) = \|p_{k-1}\|_2^2$ ; on retrouve bien la valeur de  $-\mu_k$ .  $\square$

**Exemple 1.2.** *Beaucoup d'exemples de familles de polynômes orthogonaux apparaissent dans la littérature en Mathématiques et encore plus en Physique :*

1. Sur  $[-1, 1]$ ,  $\omega(x) = 1$ , polynômes de Legendre.
2. Sur  $] -1, 1[$ ,  $\omega(x) = \frac{1}{\sqrt{1-x^2}}$ , polynômes de Tchebychev de 1ère espèce.
3. Sur  $] -1, 1[$ ,  $\omega(x) = \sqrt{1-x^2}$ , polynômes de Tchebychev de 2ème espèce.
4. Sur  $\mathbb{R}^+$ ,  $\omega(x) = \exp(-x)$ , polynômes de Laguerre.
5. Sur  $\mathbb{R}$ ,  $\omega(x) = \exp(-x^2)$ , polynômes de Hermite.

Nous terminons par le :

**Théorème 1.8.** *On a  $\tilde{p}_n \rightarrow f$  dans  $L^2_\omega([a, b])$  quand  $n \rightarrow +\infty$ .*

Dans l'énoncé, nous avons fait figurer la fonction  $\omega$  en indice de  $L^2$  pour bien mettre en valeur le fait que ce résultat est vrai pour toute fonction  $\omega \in L^1([a, b])$ .

**Preuve :** C'est une conséquence immédiate du Théorème de Weierstrass : il existe une suite  $(q_n)_n$  de polynômes qui converge uniformément vers  $f$  et quitte à modifier un peu cette suite, on peut supposer que  $q_n \in \mathcal{P}_n$  pour tout  $n$ . Comme :

$$\|f - \tilde{p}_n\|_2 \leq \|f - q_n\|_2,$$

et :

$$\|f - q_n\|_2^2 = \int_a^b |(f - q_n)(t)|^2 \omega(t) dt \leq \|f - q_n\|_\infty^2 \int_a^b \omega(t) dt,$$

la conclusion est immédiate. □

## 1.6 Exercices

- 1) Montrer que les formes linéaires suivantes forment bien une famille libre dans le dual de  $\mathbb{R}[X]$  :

$$\varphi_1 : P \mapsto P(0), \quad \varphi_2 : P \mapsto P(1), \quad \varphi_3 : P \mapsto \int_0^1 P(t) dt,$$

et déterminer la base duale de  $(\varphi_1, \varphi_2, \varphi_3)$  dans  $\mathbb{R}_3[X]$ .

- 2) **Preuve du théorème de Weierstrass par les polynômes de Bernstein, et un zeste de probas.** Soit  $f$  une fonction continue sur  $[0, 1]$  dont on note  $\omega(\cdot)$  le module de continuité uniforme, et pour

$n \in \mathbb{N}_*$  on considère la loi multinomiale de paramètre  $x$ , de taille  $n$ , donnée par les probabilités :

$$\forall 0 \leq k \leq n, \quad p_n(k, x) = C_n^k x^k (1-x)^{n-k}.$$

On introduit alors le polynôme de Bernstein :

$$P_n(X) = \sum_{k=0}^n p_n(k, x) f(k/n).$$

a. Montrer que

$$|f(x) - P_n(x)| \leq \sum_{k=0}^n p_n(k, x) |f(x) - f(k/n)|.$$

b. Montrer que pour les entiers  $k$  tels que  $|x - k/n| \leq \eta$  on a

$$\sum_{|x-k/n| \leq \eta} p_n(k, x) |f(x) - f(k/n)| \leq \omega(\eta).$$

c. Pour la somme complémentaire, il suffit d'estimer la somme  $s(n, x) = \sum_{|k-nx| \geq \eta n} p_n(k, x)$ . En utilisant l'inégalité de Bienaymé-Tchebitchev pour la loi multinomiale, montrer que

$$S(n, x) \leq \frac{x(1-x)}{\eta^2 n} \leq \frac{1}{\eta^2 n}.$$

*Indication : l'espérance et la variance de la loi multinomiale valent respectivement  $nx$  et  $nx(1-x)$ .*

d. Conclure que lorsque  $n \rightarrow \infty$ ,  $P_n$  converge vers  $f$  uniformément sur  $[0, 1]$ .

3) Ecrire explicitement les polynômes d'interpolation de Lagrange correspondant aux points  $a_0 = -1$ ,  $a_1 = 0$ ,  $a_2 = 1$  ; puis écrire la formule d'interpolation correspondante pour une fonction  $f$ . Enfin, donner l'expression obtenue pour le calcul de  $\int_{-1}^1 f(x) dx$  remplaçant  $f$  par son polynôme d'interpolation.

4) Soit  $f$  une fonction continue sur  $I = [a, b]$  et trois points distincts  $a_0, a_1, a_2$  dans  $I$ . On cherche un polynôme d'interpolation  $Q_3$  tel que

$$\forall i = 1..3, \quad Q_3(a_i) = f(a_i), \quad Q_3'(a_i) = f'(a_i).$$

a. Donner la forme explicite de  $Q_3$  en fonction de

$$\Phi_i(x) = \prod_{j \neq i} (x - x_j)^2 \text{ et } L_i^2(x).$$

b. Donner une évaluation de l'erreur  $|f(x) - Q_3(x)|$  analogue à celle obtenue dans le cas de l'interpolation de Lagrange.



c. *Calcul pratique* : Soit  $a_0 = -1$ ;  $a_1 = 0$ ,  $a_2 = 1$  et  $f(x) = \sqrt{x+2}$  sur  $I = [-1; 1]$ . Calculer Le polynôme de Lagrange associé  $P_3$  et le polynôme  $Q_3$  défini ci-dessus. Donner la majoration explicite de l'erreur dans chaque cas.

- 5) En utilisant la méthode de Newton (différences divisées), déterminer le polynôme  $Q$  tel que

$$Q(-1) = 1, \quad Q(0) = 0, \quad Q(2) = -8, \quad Q(3) = -27.$$

- 6) En utilisant les différences non divisées (première, secondes...), déterminer le polynôme  $P$  tel que

$$P(1) = 1, \quad P(2) = 8, \quad P(3) = 27, \quad P(4) = 64.$$

Que peut-on dire des différences quatrièmes de  $P$ ? Calculer  $P(-2)$  et  $P(6)$  à l'aide du tableau obtenu.

- 7) Soit  $E$  l'e.v. des polynômes de degré  $n$  ayant pour coefficient dominant 1. On note  $T_n$  le polynôme de Tchebicheff  $T_n(x) = \cos(n \arccos x)$ .
- Montrer que  $L(x) = \frac{1}{2^{n-1}} T_n(x)$  appartient à  $E$  et donner la norme  $\|L\|_\infty$  dans  $C([-1; 1])$ .
  - Montrer que  $L$  est l'élément minimum de  $E$  pour la norme  $C([-1; 1])$  : pour cela, raisonner par l'absurde en supposant qu'il existe un élément  $Q$  de norme plus petite et soit  $D = L - Q$ . En considérant les points de maximum de  $L$ , Montrer que  $D \equiv 0$  et conclure.
  - Etant donnée  $P \in E$ , quel est le meilleur polynôme de degré  $n-1$  qui approche  $P$  (au sens de la norme  $C([-1; 1])$ ) ?
- 8) Sur l'intervalle  $[-1; 1]$ , on veut construire une famille de polynômes orthogonaux pour le poids  $\omega(x) = |x|$ . Déterminer les 5 premiers éléments  $P_k$  de la famille associée à  $\omega$ , tels que  $\deg(P_k) = k$ .
- 9) Soit  $(P_n)$  une famille de polynômes orthonormée sur un intervalle  $I$  relativement à un poids  $w$ . On suppose  $P_n$  de degré  $n$  et on écrit  $P_n(x) = a_n x^n + b_n x^{n-1} + \dots$ .
- Soit  $Q_{n+1}$  défini par  $Q_{n+1}(x) = x P_n(x)$ . Montrer qu'il existe des coefficients  $c_{n,k}$  tels que

$$Q_{n+1}(x) = \sum_{k=0}^{n+1} c_{n,k} P_k(x),$$

$$\text{avec } c_{n,k} = \int_I x P_n(x) P_k(x) w(x) dx.$$

- b. En déduire la relation

$$c_{n,n+1} P_{n+1}(x) + (c_{n,n} - x) P_n(x) + c_{n,n-1} P_{n-1}(x) = 0.$$

c. Montrer les relations suivantes :

$$c_{n,n+1} = \frac{a_n}{a_{n+1}}, \quad c_{n,n} = \frac{b_n}{a_n} - \frac{b_{n+1}}{a_{n+1}}, \quad c_{n,n-1} = \frac{a_{n-1}}{a_n},$$

et en déduire une relation de récurrence sur les  $(P_n)$ .

- d. Quelle relation de récurrence obtient-on dans le cas d'une famille orthogonale, mais non nécessairement orthonormale ?
- e. Pour les familles habituelles de polynômes orthogonaux, rappeler  $I$  et  $w$ . En déduire les différentes relation de récurrence.
- 10) Soit le poids  $w_{\alpha,\beta}(x) = (1+x)^\alpha(1-x)^\beta$  dans l'intervalle  $I = ]-1; 1[$ .
- a. Donner les conditions sur  $\alpha$  et  $\beta$  pour que  $w \in L^1(dx)$ .
- b. Vérifier que le poids  $w_{0,0}$  est associé aux polynômes de Legendre  $(\mathcal{L}_n)$  et vérifier que la norme  $L^2(w)$  de  $\mathcal{L}_n$  est  $\sqrt{2}/\sqrt{2n+1}$ .
- c. Vérifier que  $w_{-1/2,-1/2}$  est associé aux polynômes de Tchebitcheff  $(T_n)$  et calculer la norme de  $T_n$  pour ce poids.
- 11) Soit  $w$  une fonction définie sur  $[a, b]$  et  $U_n$  une suite de fonctions vérifiant pour tout  $n \geq 0$  :

$$\begin{aligned} U_n^{(n-1)}(a) &= U_n^{(n-2)}(a) = \dots = U_n(a) = 0 \\ U_n^{(n-1)}(b) &= U_n^{(n-2)}(b) = \dots = U_n(b) = 0 \end{aligned}$$

Montrer que les fonctions définies par

$$P_n(x) = \frac{1}{w_n(x)} \frac{d^n}{dx^n} U_n(x)$$

forment une famille orthogonale pour le poids  $w$ .

- 12) On considère une fonction  $r$  de classe  $C^1$  sur  $[a, b]$  avec  $r(a) = r(b) = 0$ . On suppose qu'il existe une suite de réels tous distincts  $\lambda_n$  et des fonctions  $w > 0$  et  $P_n$  telles que

$$\frac{d}{dx} \left( r(x) \frac{d}{dx} P_n(x) \right) = \lambda_n P_n(x) w(x) \text{ dans } [a, b].$$

Calculer  $\frac{d}{dx} \left( r(x) \{P_m(x)P'_n(x) - P_n(x)P'_m(x)\} \right)$ , et en déduire que la famille  $(P_n)$  forme une famille orthogonale pour le poids  $w$ .

## Chapitre 2

# Intégration et dérivation numérique

### 2.1 La méthode classique d'intégration numérique : Newton-Cotes

#### 2.1.1 Présentation de la méthode

Le but est de donner une valeur approchée de  $\int_a^b f(t)dt$  où  $f \in C([a, b])$ , et bien évidemment dans le cas où  $f$  n'a pas de primitive explicite comme par exemple  $f(t) = \exp(-t^2)$ . Ceci se fait en trois étapes :

1. On subdivise l'intervalle  $[a, b]$  en choisissant des points  $x_0 = a < x_1 < x_2 < \dots < x_n = b$ .
2. On estime  $\int_{x_i}^{x_{i+1}} f(t)dt$  en utilisant une formule de quadrature.
3. On reconstitue l'approximation de  $\int_a^b f(t)dt$  via la formule de Chasles.

Il est bien clair que, dans cette stratégie, c'est la deuxième étape qui est la plus importante et c'est par elle que nous allons commencer. En général, on utilise la même formule de quadrature sur chaque intervalle (bien que toute fantaisie soit permise...) et donc on ne fait que "transporter" une seule formule de quadrature définie sur un intervalle de référence par la méthode suivante que l'on décrit (pour simplifier) dans le cas où la subdivision utilisée a des points équidistants. On note  $h = x_{i+1} - x_i$  et on fait le changement de variable :

$$t = \frac{x_{i+1} + x_i}{2} + s \frac{h}{2},$$
$$g(s) = f\left(\frac{x_{i+1} + x_i}{2} + s \frac{h}{2}\right).$$

La fonction  $g$  dépend de l'intervalle considéré donc de “ $i$ ” mais on omet cet indice pour avoir des notations plus simple. On a :

$$\int_{x_i}^{x_{i+1}} f(t)dt = \frac{h}{2} \int_{-1}^1 g(s)ds ,$$

et il suffit de donner une formule de quadrature pour l'intégrale sur  $[-1, 1]$  du type :

$$(2.1) \quad \int_{-1}^1 g(s)ds \simeq L(g) := \sum_{j=0}^K \omega_j g(y_j) .$$

où  $-1 \leq y_0 < y_1 < \dots < y_K \leq 1$  et les  $\omega_j$  sont des coefficients que l'on peut “bien choisir” en utilisant le :

**Lemme 2.1.** *Pour tout choix de points  $y_0 < y_1 < \dots < y_K$ , il existe un unique  $(K+1)$ -uplets  $(\omega_0, \omega_1, \dots, \omega_K)$  tel que la formule (2.1) soit exacte pour tout polynôme de  $\mathcal{P}_K$ .*

**Preuve :** Deux méthodes à connaître :

(i) On considère les polynômes  $(L_i)_i$  d'interpolation de Lagrange associés aux points  $y_0 < y_1 < \dots < y_K$ . On sait que ces polynômes sont dans  $\mathcal{P}_K$  et que :

$$L_i(y_j) = \delta_{i,j} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{sinon} \end{cases}$$

Si on veut que la formule de quadrature soit exacte sur  $\mathcal{P}_K$  alors on doit avoir, pour tout  $j = 0, 1, \dots, K$  :

$$\int_{-1}^1 L_j(s)ds = L(L_j) = \omega_j .$$

et comme les  $(L_i)_i$  forment une base de  $\mathcal{P}_K$  (exo!), on vérifie facilement que cette condition nécessaire est aussi suffisante.

**NB :** il est important de noter ce (premier) lien très étroit entre l'interpolation et le calcul approché d'intégrale.

(ii) On écrit simplement que, pour que la formule de quadrature soit exacte sur  $\mathcal{P}_K$ , il suffit qu'elle le soit sur la base canonique de  $\mathcal{P}_K$ , c'est-à-dire  $(1, x, \dots, x^K)$ . Ceci donne, pour  $0 \leq k \leq K$  :

$$\int_{-1}^1 s^k ds = L(x^k) = \sum_{j=0}^K \omega_j y_j^k .$$

Il s'agit donc d'un système linéaire de matrice :

$$\begin{pmatrix} 1 & 1 & \cdots & 1 & \cdots & 1 \\ y_0 & y_1 & \cdots & y_j & \cdots & y_K \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_0^k & y_1^k & \cdots & y_j^k & \cdots & y_K^k \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_0^K & y_1^K & \cdots & y_j^K & \cdots & y_K^K \end{pmatrix}$$

On reconnaît une *matrice de Vandermonde* qui est inversible dès que les  $y_j$  sont distincts. D'où l'existence et l'unicité des  $\omega_j$ .  $\square$

**Exemple 2.1.** *Les exemples les plus importants de formules de quadrature qui conduisent à des méthodes (simples) d'intégration numérique sont les suivants :*

- *Méthode des rectangles* :  $K = 0$ ,  $y_0 = -1, 0$  ou  $1$ . On a  $\omega_0 = 2$  et :

$$\int_{-1}^1 g(s) ds \simeq 2g(y_0) .$$

- *Méthode des trapèzes* :  $K = 1$  avec, par exemple,  $y_0 = -1$ ,  $y_1 = 1$ . On a  $\omega_0 = 1$ ,  $\omega_1 = 1$  et :

$$\int_{-1}^1 g(s) ds \simeq g(-1) + g(1) .$$

- *Méthode de Simpson* :  $K = 1$ ,  $y_0 = -1$ ,  $y_1 = 0$ ,  $y_2 = 1$ . On a  $\omega_0 = \frac{1}{3}$ ,  $\omega_1 = \frac{4}{3}$ ,  $\omega_2 = \frac{1}{3}$  et :

$$\int_{-1}^1 g(s) ds \simeq \frac{1}{3}g(-1) + \frac{4}{3}g(0) + \frac{1}{3}g(1) .$$

**Remarque :** Si les  $y_j$  sont symétriques par rapport à 0 (c'est-à-dire si  $-y_j$  est l'un des  $y_k$  pour tout  $j$ ), alors, par unicité des  $\omega_j$ , la formule est symétrique :  $\omega_j = \omega_k$  si  $y_k = -y_j$ . De plus, si  $K$  est pair, la formule n'est pas seulement exacte sur  $\mathcal{P}_K$  mais aussi sur  $\mathcal{P}_{K+1}$  car l'intégrale et la formule de quadrature sont toutes les deux nulles pour  $x^{K+1}$ .

Revenons maintenant à l'effet de la formule de quadrature et à l'approximation des intégrales sur chaque intervalles  $[x_i, x_{i+1}]$  et sur  $[a, b]$ . Sur

$[x_i, x_{i+1}] :$

$$\begin{aligned} \int_{x_i}^{x_{i+1}} f(t) dt &= \frac{h}{2} \int_{-1}^1 g(s) ds \\ &\simeq \frac{h}{2} \sum_{j=0}^K \omega_j g(y_j) , \\ &\simeq \frac{h}{2} \sum_{j=0}^K \omega_j f \left( \frac{x_{i+1} + x_i}{2} + y_j \frac{h}{2} \right) . \end{aligned}$$

Puis on somme sur  $i$  pour avoir l'approximation sur  $[a, b]$  :

$$\int_a^b f(t) dt \simeq \sum_{i=0}^{n-1} \frac{h}{2} \sum_{j=0}^K \omega_j f \left( \frac{x_{i+1} + x_i}{2} + y_j \frac{h}{2} \right) .$$

**Exemple 2.2.** Nous reprenons les trois méthodes évoquées dans l'exemple précédent :

– Méthode des rectangles :

$$\int_a^b f(t) dt \simeq h \sum_{i=0}^{n-1} f \left( \frac{x_{i+1} + x_i}{2} \right) .$$

– Méthode des trapèzes :

$$\int_a^b f(t) dt \simeq \frac{h}{2} \sum_{i=0}^{n-1} (f(x_i) + f(x_{i+1}))$$

– Méthode de Simpson :

$$\int_a^b f(t) dt \simeq \frac{h}{2} \sum_{i=0}^{n-1} \left( \frac{1}{3} f(x_i) + \frac{4}{3} f \left( \frac{x_{i+1} + x_i}{2} \right) + \frac{1}{3} f(x_{i+1}) \right)$$

### 2.1.2 Stabilité

Comme nous l'avons déjà vu dans le cas des polynômes d'interpolation, on doit toujours s'intéresser à l'influence des perturbations sur les données : comment sont propagées les erreurs sur les valeurs de  $f$  dans la formule d'approximation de l'intégrale

$$I_n := \sum_{i=0}^{n-1} \frac{h}{2} \sum_{j=0}^K \omega_j f \left( \frac{x_{i+1} + x_i}{2} + y_j \frac{h}{2} \right) ?$$

## 2.1. LA MÉTHODE CLASSIQUE D'INTÉGRATION NUMÉRIQUE : NEWTON-COTES31

Si les valeurs des  $f\left(\frac{x_{i+1}+x_i}{2} + y_j \frac{h}{2}\right)$  sont remplacées par des  $f_{i,j}$ ,  $I_n$  est changé en :

$$\tilde{I}_n := \sum_{i=0}^{n-1} \frac{h}{2} \sum_{j=0}^K \omega_j f_{i,j} ,$$

et on a :

$$\begin{aligned} |I_n - \tilde{I}_n| &= \left| \sum_{i=0}^{n-1} \frac{h}{2} \sum_{j=0}^K \omega_j f\left(\frac{x_{i+1}+x_i}{2} + y_j \frac{h}{2}\right) - \sum_{i=0}^{n-1} \frac{h}{2} \sum_{j=0}^K \omega_j f_{i,j} \right| \\ &= \left| \sum_{i=0}^{n-1} \frac{h}{2} \sum_{j=0}^K \omega_j \left( f\left(\frac{x_{i+1}+x_i}{2} + y_j \frac{h}{2}\right) - f_{i,j} \right) \right| \\ &\leq \sum_{i=0}^{n-1} \frac{h}{2} \sum_{j=0}^K |\omega_j| \left| f\left(\frac{x_{i+1}+x_i}{2} + y_j \frac{h}{2}\right) - f_{i,j} \right|. \end{aligned}$$

Si on note  $M = \max_{i,j} \left| f\left(\frac{x_{i+1}+x_i}{2} + y_j \frac{h}{2}\right) - f_{i,j} \right|$  alors :

$$|I_n - \tilde{I}_n| \leq \sum_{i=0}^{n-1} \frac{h}{2} \sum_{j=0}^K |\omega_j| M = n \frac{h}{2} \left( \sum_{j=0}^K |\omega_j| \right) M = \frac{b-a}{2} \left( \sum_{j=0}^K |\omega_j| \right) M .$$

car  $nh = b - a$ . On voit donc que l'erreur sur les valeurs de  $f$  est amplifiée/atténuée via le coefficient  $\frac{b-a}{2} \left( \sum_{j=0}^K |\omega_j| \right)$ .

### 2.1.3 Estimations d'erreur

#### Estimations d'erreur pour des fonctions continues générales

Nous commençons par un résultat basique valable pour toute fonction continue. Pour cela, on introduit la notion de module de continuité (uniforme).

**Définition 2.1.** On appelle module de continuité (uniforme) d'une fonction continue  $f : [a, b] \rightarrow \mathbb{R}$  toute fonction  $m : [0, +\infty[ \rightarrow \mathbb{R}$  vérifiant les propriétés suivantes :

- Pour tout  $x, y \in [a, b]$ ,  $|f(x) - f(y)| \leq m(|x - y|)$ .
- $m(0) = 0$  et  $m(t) \rightarrow 0$  quand  $t \rightarrow 0^+$ .
- $m$  est une fonction croissante.
- $m(t + s) \leq m(t) + m(s)$  pour tous  $s, t \geq 0$ .

Comme  $f$  est uniformément continue sur  $[a, b]$  (Théorème de Heine), la fonction suivante est un module de continuité pour  $f$  :

$$m(t) = \sup\{|f(x) - f(y)| ; |x - y| \leq t\} ,$$

mais ce n'est pas la seule car, par exemple,  $2m$  convient aussi. Le module de continuité, comme son nom l'indique, est une certaine mesure de la continuité d'une fonction (écart entre  $f(x)$  et  $f(y)$  quand  $x$  et  $y$  sont proches).

**Remarque :** Certaines classes de fonctions (bien connues ?) sont définies via leur module de continuité : les fonctions lipschitziennes où  $m(t) = Ct$  pour une certaine constante  $C$  (= constante de Lipschitz) ou höldériennes où  $m(t) = Ct^\alpha$  pour  $0 < \alpha < 1$  et pour une certaine constante  $C$ .

Le premier résultat est le :

**Théorème 2.1.** *Soit  $f \in C([a, b])$  et  $m$  un module de continuité de  $f$ . Alors :*

$$\left| \int_a^b f(t)dt - \sum_{i=0}^{n-1} \frac{h}{2} \sum_{j=0}^K \omega_j f\left(\frac{x_{i+1} + x_i}{2} + y_j \frac{h}{2}\right) \right| \leq \frac{b-a}{2} \left( \sum_{j=0}^K |\omega_j| \right) m(h) .$$

Le double avantage de ce résultat est sa simplicité et sa généralité mais, par contre, la vitesse de convergence (en  $h$  pour les fonctions lipschitziennes ou  $h^\alpha$  pour les fonctions höldériennes) n'est pas très bonne. On verra, par la suite, comment obtenir de meilleures vitesses de convergence en augmentant la régularité de  $f$ .

**Preuve :** On estime d'abord :

$$Q_i = \left| \int_{x_i}^{x_{i+1}} f(t)dt - \frac{h}{2} \sum_{j=0}^K \omega_j f\left(\frac{x_{i+1} + x_i}{2} + y_j \frac{h}{2}\right) \right| ,$$

qui est l'erreur sur un sous-intervalle  $[x_i, x_{i+1}]$ . On cumulera ensuite ces erreurs pour obtenir l'erreur sur  $[a, b]$ .

On remarque d'abord que, la formule de quadrature étant exacte sur  $\mathcal{P}_0$  (au moins), on a  $\sum_{j=0}^K \omega_j = 2$ . On a donc :

$$\int_{x_i}^{x_{i+1}} f(t)dt = \frac{1}{2} \sum_{j=0}^K \omega_j \int_{x_i}^{x_{i+1}} f(t)dt ,$$

et :

$$Q_i = \left| \frac{1}{2} \sum_{j=0}^K \omega_j \left( \int_{x_i}^{x_{i+1}} f(t)dt - hf\left(\frac{x_{i+1} + x_i}{2} + y_j \frac{h}{2}\right) \right) \right| .$$

Mais  $hf\left(\frac{x_{i+1} + x_i}{2} + y_j \frac{h}{2}\right) = \int_{x_i}^{x_{i+1}} f\left(\frac{x_{i+1} + x_i}{2} + y_j \frac{h}{2}\right) dt$ , d'où, en utilisant plusieurs fois l'inégalité triangulaire :

$$\begin{aligned} Q_i &= \left| \frac{1}{2} \sum_{j=0}^K \omega_j \int_{x_i}^{x_{i+1}} \left( f(t) - f\left(\frac{x_{i+1} + x_i}{2} + y_j \frac{h}{2}\right) \right) dt \right| \\ &\leq \frac{1}{2} \sum_{j=0}^K |\omega_j| \int_{x_i}^{x_{i+1}} \left| f(t) - f\left(\frac{x_{i+1} + x_i}{2} + y_j \frac{h}{2}\right) \right| dt . \end{aligned}$$



## 2.1. LA MÉTHODE CLASSIQUE D'INTÉGRATION NUMÉRIQUE : NEWTON-COTES33

Mais, dans chacune des intégrales, on peut estimer  $|f(t) - f\left(\frac{x_{i+1}+x_i}{2} + y_j \frac{h}{2}\right)|$  par  $m(|t - (\frac{x_{i+1}+x_i}{2} + y_j \frac{h}{2})|) \leq m(h)$  car les deux points sont dans le même intervalle de longueur  $h$  et on utilise ensuite la croissance de  $m$ .

Finalement :

$$Q_i \leq \frac{1}{2} \left( \sum_{j=0}^K |\omega_j| \right) \int_{x_i}^{x_{i+1}} m(h) dt = \frac{1}{2} \left( \sum_{j=0}^K |\omega_j| \right) h m(h) .$$

Si on pose :

$$Q = \left| \int_a^b f(t) dt - \sum_{i=0}^{n-1} \frac{h}{2} \sum_{j=0}^K \omega_j f\left(\frac{x_{i+1}+x_i}{2} + y_j \frac{h}{2}\right) \right| ,$$

on voit facilement que :

$$Q \leq \sum_{i=0}^{n-1} Q_i ,$$

(il suffit d'utiliser la relation de Chasles et de se ramener aux intervalles  $[x_i, x_{i+1}]$ ) il en résulte donc :

$$\begin{aligned} Q &\leq \sum_{i=0}^{n-1} \frac{1}{2} \left( \sum_{j=0}^K |\omega_j| \right) h m(h) \\ &= n \frac{1}{2} \left( \sum_{j=0}^K |\omega_j| \right) h m(h) . \end{aligned}$$

Mais  $nh = b - a$  et on conclut :

$$Q \leq \frac{b-a}{2} \left( \sum_{j=0}^K |\omega_j| \right) m(h) .$$

□

### Estimations d'erreur pour des fonctions régulières

On procède en deux temps : d'abord on obtient une estimation d'erreur sur l'intervalle de référence  $[-1, 1]$  puis on "transporte" cette estimation sur chaque intervalle  $[x_i, x_{i+1}]$  avant de procéder au cumul des erreurs sur l'intervalle  $[a, b]$  tout entier comme dans la section précédente.

On commence par le :

**Théorème 2.2.** Soit  $g$  une fonction de classe  $C^{K+1}$  sur  $[-1, 1]$ . Alors :

$$\left| \int_{-1}^1 g(t) dt - L(g) \right| \leq \frac{2M_K}{(K+1)!} \|g^{(K+1)}\|_{\infty},$$

où l'on rappelle que  $L(g) := \sum_{j=0}^K \omega_j g(y_j)$  et :

$$M_K = \|(y - y_0) \cdots (y - y_K)\|_{\infty}.$$

**Preuve :** Si  $\Pi g \in \mathcal{P}_K$  est l'unique polynôme d'interpolation de  $g$  aux points  $y_0, \dots, y_K$ , on a :

$$L(g) = L(\Pi g) = \int_{-1}^1 \Pi g(t) dt,$$

puisque la formule de quadrature est exacte sur  $\mathcal{P}_K$ .

On a alors :

$$\left| \int_{-1}^1 g(t) dt - L(g) \right| = \left| \int_{-1}^1 (g(t) - \Pi g(t)) dt \right| \leq 2 \|g - \Pi g\|_{\infty}.$$

Mais le Théorème 1.3 implique que :

$$g(t) - \Pi g(t) = \frac{1}{(K+1)!} (t - y_0) \cdots (t - y_K) g^{(K+1)}(\xi_t),$$

et donc :

$$|g(t) - \Pi g(t)| \leq \frac{1}{(K+1)!} \|(t - y_0) \cdots (t - y_K)\|_{\infty} \|g^{(K+1)}\|_{\infty},$$

Le résultat découle des deux inégalités précédentes.  $\square$

On suppose maintenant que  $f$  est de classe  $C^{K+1}$  et on passe maintenant à l'estimation de :

$$Q_i = \left| \int_{x_i}^{x_{i+1}} f(t) dt - \frac{h}{2} \sum_{j=0}^K \omega_j f\left(\frac{x_{i+1} + x_i}{2} + y_j \frac{h}{2}\right) \right|.$$

On fait le changement de variable :

$$t = \frac{x_{i+1} + x_i}{2} + s \frac{h}{2},$$

$$g(s) = f\left(\frac{x_{i+1} + x_i}{2} + s \frac{h}{2}\right),$$

qui nous conduit à :

$$Q_i = \frac{h}{2} \left| \int_{-1}^1 g(t) dt - L(g) \right|$$

## 2.1. LA MÉTHODE CLASSIQUE D'INTÉGRATION NUMÉRIQUE : NEWTON-COTES35

et le résultat sur l'intervalle de référence nous amène à :

$$Q_i \leq \frac{h}{2} \frac{2M_K}{(K+1)!} \|g^{(K+1)}\|_\infty \leq \left(\frac{h}{2}\right)^{K+2} \frac{2M_K}{(K+1)!} \|f^{(K+1)}\|_\infty ,$$

$$\text{car } g^{(K+1)}(s) = \left(\frac{h}{2}\right)^{K+1} f^{(K+1)}\left(\frac{x_{i+1} + x_i}{2} + s\frac{h}{2}\right).$$

On en déduit le :

**Théorème 2.3.** Soit  $f$  une fonction de classe  $C^{K+1}$  sur  $[a, b]$ . Alors :

$$\left| \int_a^b f(t)dt - \sum_{i=0}^{n-1} \frac{h}{2} \sum_{j=0}^K \omega_j f\left(\frac{x_{i+1} + x_i}{2} + y_j \frac{h}{2}\right) \right| \leq (b-a) \left(\frac{h}{2}\right)^{K+1} \frac{M_K}{(K+1)!} \|f^{(K+1)}\|_\infty .$$

**Preuve :** Il suffit d'utiliser le fait que :

$$Q \leq \sum_{i=0}^{n-1} Q_i ,$$

et la relation  $nh = b - a$ . □

**Remarque :** On peut raffiner les preuves ci-dessus pour avoir, non plus une majoration, mais un vrai développement limité, de  $I_n$ . On écrit d'abord :

$$g(t) - \Pi g(t) = \frac{1}{(K+1)!} (t-y_0) \cdots (t-y_K) g^{(K+1)}(0) + O(\|g^{(K+1)} - g^{(K+1)}(0)\|_\infty) .$$

Ce qui se traduit par :

$$\begin{aligned} \int_{x_i}^{x_{i+1}} f(t)dt - \frac{h}{2} \sum_{j=0}^K \omega_j f\left(\frac{x_{i+1} + x_i}{2} + y_j \frac{h}{2}\right) &= \left(\frac{h}{2}\right)^{K+2} \frac{l_K}{(K+1)!} f^{(K+1)}\left(\frac{x_{i+1} + x_i}{2}\right) \\ &\quad + o\left(\left(\frac{h}{2}\right)^{K+2}\right) , \end{aligned}$$

où  $l_K = \int_{-1}^1 (t-y_0) \cdots (t-y_K) dy$  et le "o" provient du module de continuité de  $f^{(K+1)}$  sur  $[a, b]$ .

En sommant sur  $i$ , on remarque que le premier terme donne une somme de Riemann pour  $f^{(K+1)}$  et donc :

$$\int_a^b f(t)dt - I_n = \left(\frac{h}{2}\right)^{K+1} \frac{l_K}{2(K+1)!} \int_a^b f^{(K+1)}(t)dt + o(h^{K+1}) .$$

On peut lire cette égalité comme :

$$I_n = \int_a^b f(t)dt + Ah^{K+1} + o(h^{K+1}) ,$$

ce qui donne bien un développement de l'erreur. De plus, le  $o(h^{K+1})$  est un  $O(h^{K+2})$  si  $f$  est de classe  $C^{K+2}$ .

### 2.1.4 Méthode d'accélération de Romberg

Nous n'entrerons pas dans trop de détails et nous indiquerons simplement le principe.

On pose  $I = \int_a^b f(t)dt$  et on considère la méthode des trapèzes où :

$$I_n = \frac{h}{2}[f(a) + f(b)] + h \sum_{i=1}^{n-1} f(x_i) .$$

La section précédente nous donne, pour des fonctions assez régulières :

$$I = I_n + Ah^2 + O(h^3) ,$$

et en raffinant la subdivision avec un pas  $h$  deux fois plus petit (donc avec deux fois plus de points) :

$$I = I_{2n} + \frac{Ah^2}{4} + O(h^3) .$$

Un simple calcul conduit à :

$$3I = 4I_{2n} - I_n + O(h^3) ,$$

et on gagne une ordre puisque la convergence est en  $h^3$  au lieu de  $h^2$ .

**Remarque : Extrapolation de Richardson :** en fait on peut aller plus loin car si on a :

$$I = I_n + a_1h + a_2h^2 + \dots + a_nh^n + O(h^{n+1}) ,$$

on peut théoriquement éliminer  $a_1h, a_2h^2, \dots, a_nh^n$ , en considérant plusieurs niveaux de subdivisions. Mais attention la complexité peut devenir coûteuse...

## 2.2 Un choix optimal de points : les points de Gauss-Legendre

Au lieu de choisir n'importe quels points  $y_j$  dans la formule de quadrature, on peut essayer de les prendre de manière optimale, c'est-à-dire de telle sorte que la formule de quadrature soit exacte sur l'espace de polynômes le plus grand possible car, comme on l'a vu précédemment, l'ordre de convergence de la méthode de Newton-Cotes associée sera également la plus grande possible. Cet objectif est réalisé via le

**Théorème 2.4.** *Pour tout  $l \in \mathbb{N}$ , il existe  $l+1$  points  $a_j$  et  $l+1$  constantes  $\omega_j$  tels que :*

$$\int_{-1}^1 p(s)ds = \sum_{j=0}^l \omega_j p(a_j) \quad \text{pour tout } p \in \mathcal{P}_{2l+1} .$$

Ce résultat semble effectivement optimal car nous avons  $(2l + 2)$  réels à fixer (les  $a_j$  et les  $\omega_j$ ) avec le but de réaliser  $(2l + 2)$  égalités car  $\mathcal{P}_{2l+1}$  est de dimension  $2l + 2$  et il suffit que l'égalité ait lieu sur une base de  $\mathcal{P}_{2l+1}$ .

**Preuve :** On considère dans  $\mathcal{P}_{l+1}$  le produit scalaire de  $L^2$ , i.e.

$$(p, q) = \int_{-1}^1 p(s)q(s)ds .$$

On note  $S$  l'orthogonal de  $\mathcal{P}_l$  dans  $\mathcal{P}_{l+1}$  pour ce produit scalaire. La dimension de  $S$  est 1 et on considère  $\bar{P} \in S$  un élément non nul (donc un vecteur directeur) de  $S$ . On va étudier les propriétés d'un tel polynôme  $\bar{P}$ .

- $\bar{P}$  est de degré exactement  $l + 1$  : en effet, sinon  $\bar{P} \in \mathcal{P}_l$  et donc  $(\bar{P}, \bar{P}) = 0$  car  $\bar{P}$  est orthogonal à  $\mathcal{P}_l$ , d'où  $\bar{P} \equiv 0$ .

- $\bar{P}$  a toutes ses racines dans  $[-1, 1]$  : en effet, sinon on pourrait écrire sous la forme  $\bar{P}(x) = \tilde{p}(x)q(x)$  avec, d'une part :

$$\tilde{p}(x) = (x - b_1)(x - b_2) \cdots (x - b_k) ,$$

où les  $b_i$  sont les racines de  $\bar{P}$  dans  $[-1, 1]$  et où, d'autre part,  $q$  est sans racine dans  $[-1, 1]$  (donc de signe constant dans  $[-1, 1]$ ). On remarque que  $k \leq l$  sinon  $\bar{P}$  aurait toutes ses racines dans  $[-1, 1]$ . On écrit alors que  $(\bar{P}, \tilde{p}) = 0$  puisque  $\tilde{p} \in \mathcal{P}_l$ , ce qui donne :

$$\int_{-1}^1 |\tilde{p}(s)|^2 q(s) ds = 0,$$

une contradiction car  $q$  est de signe constant et non nul sur  $[-1, 1]$ .

- Si les  $a_j$  ( $0 \leq j \leq l$ ) sont les racines de  $\bar{P}$  dans  $[-1, 1]$  alors les  $a_j$  sont distincts : en effet, si on a une racine double (disons  $a_0 = a_1$ ), alors :

$$\bar{P}(x) = (x - a_0)^2(x - a_2) \cdots (x - a_l) .$$

On note cette fois  $\tilde{p}(x) = (x - a_2) \cdots (x - a_l)$  et on remarque que  $\tilde{p} \in \mathcal{P}_l$ . En écrivant alors que  $(\bar{P}, \tilde{p}) = 0$  car  $\tilde{p} \in \mathcal{P}_l$ , on obtient :

$$\int_{-1}^1 (s - a_0)^2 |\tilde{p}(s)|^2 ds = 0,$$

ce qui donne la contradiction.

On tire alors avantage des propriétés de  $\bar{P}$  : si  $p$  est un polynôme quelconque de  $\mathcal{P}_{2l+1}$ , on fait la division euclidienne de  $p$  par  $\bar{P}$  :

$$p = \bar{P}Q + R ,$$

où le degré de  $R$  est strictement inférieur au degré de  $\bar{P}$  (donc  $R \in \mathcal{P}_l$ ) et un simple argument de degré montre que  $Q$  est forcément dans  $\mathcal{P}_l$  puisque  $p \in \mathcal{P}_{2l+1}$  et que le degré de  $\bar{P}$  est exactement  $l + 1$ .

On écrit alors :

$$\int_{-1}^1 p(t)dt = \int_{-1}^1 \bar{P}(t)Q(t)dt + \int_{-1}^1 R(t)dt ,$$

et on remarque que  $\int_{-1}^1 \bar{P}(t)Q(t)dt = 0$  puisque  $Q \in \mathcal{P}_l$ . On choisit alors les constantes  $\omega_j$  pour que :

$$\int_{-1}^1 r(s)ds = \sum_{j=0}^l \omega_j r(a_j) \quad \text{pour tout } r \in \mathcal{P}_l ,$$

ce qui est possible d'après la partie sur Newton-Cotes. Il en résulte :

$$\begin{aligned} \int_{-1}^1 p(t)dt &= \int_{-1}^1 \bar{P}(t)Q(t)dt + \int_{-1}^1 R(t)dt \\ &= \int_{-1}^1 R(t)dt = \sum_{j=0}^l \omega_j R(a_j) \\ &= \sum_{j=0}^l \omega_j p(a_j) \end{aligned}$$

la dernière ligne provenant du fait que les  $a_j$  sont les racines de  $\bar{P}$ , donc

$$R(a_j) = p(a_j) - \bar{P}(a_j)Q(a_j) = p(a_j) .$$

□

## 2.3 Dérivation numérique

L'objectif est ici de donner une approximation des dérivées d'une fonction à partir de la connaissance de ses valeurs en un certain nombre de points (un grand nombre de points, en général).

La première idée pourrait consister à utiliser l'interpolation de Lagrange : si  $p_n$  est l'interpolée d'une fonction régulière  $f$  sur  $[a, b]$  et si  $x \in ]a, b[$ , on peut penser que  $p'_n(x)$  peut approcher  $f'(x)$ . Mais cette approche se révèle assez vite catastrophique,  $p'_n(x)$  étant le plus souvent très éloigné de  $f'(x)$ .

Une meilleure idée est la *méthode des différences finies* : si  $x \in ]a, b[$  et si  $|h| \ll 1$  alors :

$$f'(x) \simeq \frac{f(x+h) - f(x)}{h} ,$$

puisque  $f'(x)$  est la limite du quotient différentiel. On peut même aller plus loin en utilisant la formule de Taylor pour approcher  $f'(x), f''(x), f^{(3)}(x), \dots, f^{(n)}(x)$  ce qui nécessite l'utilisation de plus en plus de points.

On note  $S(f, h, x)$  une méthode d'approximation de la dérivée de  $f$  en  $x$  (on verra des exemples autres que celui ci-dessus plus loin) : on juge de la qualité (supposée) de cette approximation de la dérivée grâce à la notion d'ordre présentée maintenant.

**Définition 2.2.** On dit que  $S(f, h, x)$  est une méthode d'ordre  $k$ , s'il existe une constante  $K$  dépendant de  $f$  telle que :

$$|S(f, h, x) - f'(x)| \leq Kh^k .$$

Pour connaître l'ordre d'une approximation [définition qui s'étend sans problème au cas de l'approximation des dérivées supérieures], on utilise la formule de Taylor en supposant que  $f$  est régulière. Si :

$$S(f, h, x) = \frac{f(x+h) - f(x)}{h} ,$$

on écrit :

$$f(x+h) = f(x) + f'(x)h + \frac{1}{2}f''(x)h^2 + o(h^2) .$$

D'où :

$$S(f, h, x) = f'(x) + \frac{1}{2}f''(x)h + o(h) .$$

On a une approximation d'ordre 1 car le terme d'erreur est en  $h$ .

Par contre, si :

$$S_1(f, h, x) = \frac{f(x+h) - f(x-h)}{2h} ,$$

alors :

$$f(x+h) = f(x) + f'(x)h + \frac{1}{2}f''(x)h^2 + \frac{1}{3!}f^{(3)}(x)h^3 + o(h^3) ,$$

$$f(x-h) = f(x) - f'(x)h + \frac{1}{2}f''(x)h^2 - \frac{1}{3!}f^{(3)}(x)h^3 + o(h^3) ,$$

et :

$$S_1(f, h, x) = f'(x) + \frac{1}{3!}f^{(3)}(x)h^2 + o(h^2) .$$

Cette fois, l'approximation est d'ordre 2.

## 2.4 Exercices

- 1) Reprendre les formules d'interpolation de Lagrange à 1, 2 et 3 points et en déduire les formules d'intégration élémentaire associées, puis les formules composées.

2) On considère les quatres formes linéaires suivantes :

$$f_1 : P \mapsto P(-1), \quad f_2 : P \mapsto P(1), \quad f_3 : P \mapsto P'(-1), \quad f_4 : P \mapsto P'(1).$$

A partir de ces formes et de la base duale associée, construire une formule d'intégration du type  $\int_0^1 f \sim af(-1)+bf'(1)+cf(1)+df'(1)$ .  
On donnera les réels  $a, b, c, d$  et une estimation de l'erreur commise.

3) On veut estimer l'erreur d'une formule d'intégration approchée :

$$E(f) = \sum \lambda_i f(x_i) - \int_a^b f(x) dx,$$

où les  $\lambda_i$  et les  $x_i$  sont donnés,  $x_i \in [a, b]$ . On suppose  $f$  de classe  $C^{n+1}$  sur  $[a, b]$  et la formule d'intégration exacte pour les polynômes de degré inférieur ou égal à  $n$ .

On note  $g^+ = \max\{g, 0\}$  la partie positive et  $\varphi_n(x, t) = [(x-t)^+]^n$ .

a. On écrit la formule de Lagrange avec reste intégral

$$f(x) = f(a) + (x-a)f'(a) + \dots + (x-a)^n \frac{f^{(n)}(a)}{n!} + r_n(x) = P_n(x) + r_n(x).$$

Montrer que

$$r_n(x) = \frac{1}{n!} \int_a^x f^{(n+1)}(t)(x-t)^n dt = \frac{1}{n!} \int_a^b f^{(n+1)}(t)\varphi_n(x, t) dt.$$

b. En déduire que  $E(f) = E(r_n)$ , puis que

$$E(f) = \frac{1}{n!} \int_a^b f^{(n+1)}(t)k_n(t) dt \text{ avec } k_n(t) = E(x \mapsto \varphi_n(x, t)).$$

c. On suppose  $k_n$  positif sur  $[a, b]$  ; montrer qu'il existe  $\xi \in ]a, b[$  tel que

$$E(f) = \frac{1}{n!} f^{(n+1)}(\xi) \int_a^b k_n(t) dt,$$

et montrer que  $E(f) = E(x^{n+1})f^{(n+1)}(\xi)/(n+1)!$ .

d. Application : On se place sur  $[-1, 1]$  ; déterminer le noyau de Péano  $k_n$  correspondant à la formule des trapèzes, et à la formule de Simpson. Vérifier que ce sont des fonctions positives et conclure.



## Chapitre 3

# Analyse numérique des équations différentielles

### 3.1 Rappels théoriques

#### 3.1.1 Théorie générale

Une équation différentielle ordinaire (EDO) est une équation de la forme :

$$\dot{y}(t) = f(t, y(t)) \quad \text{dans } I ,$$

auquel il faut adjoindre une “condition initiale” :

$$y(t_0) = y_0 ,$$

où  $t_0$  est un point de l'intervalle  $I$ . L'inconnue est la fonction  $y : I \rightarrow \mathbb{R}^n$  alors que la fonction  $f : I \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  et la donnée initiale  $y_0 \in \mathbb{R}^n$  sont des données du problème. En fait, en pratique, l'intervalle  $I$  sera souvent à déterminer.

Pour simplifier, nous allons supposer que  $t_0 = 0$ , ce qui n'est pas une perte de généralité car on peut toujours remplacer la fonction  $y(\cdot)$  par  $\tilde{y}(\cdot) = y(t_0 + \cdot)$  et  $I$  par  $I - t_0$ . L'étude de l'EDO se fait (en général) en deux temps :

- (i) on prouve l'existence et l'unicité “locale” de la solution (c'est-à-dire sur un intervalle de la forme  $[0, \tau]$  ou  $[-\tau, \tau]$  où  $\tau > 0$  est a priori un temps “petit”) ; le résultat de base pour obtenir ces propriétés est le *Théorème de Cauchy-Lipschitz*.
- (ii) On entend la solution sur un intervalle  $I$  le plus grand possible :  $[0, T]$  ou  $[-T, T]$ .

Dans toute la suite, nous ne considérerons que le problème de Cauchy de manière “progressive”, c'est-à-dire sur des temps positifs :  $y(0)$  est donné et on veut calculer  $y(t)$  pour  $t \in [0, T]$ . Mais on peut résoudre aussi pour  $t < 0$  (de manière “rétrograde”).

Nous commençons par rappeler le Théorème de Cauchy-Lipschitz qui utilise l'hypothèse suivante (notée **(LL)** pour “localement Lipschitz”) :

**(LL)** Pour tout  $0 < T' < T$  et pour tout  $R > 0$ , il existe une constante  $L(T', R)$  telle que :

$$|f(t, y_1) - f(t, y_2)| \leq L(T', R)|y_1 - y_2| ,$$

pour tous  $t \in [0, T']$  et  $|y_1|, |y_2| \leq R$ .

Le résultat est le :

**Théorème 3.1.** *Sous l'hypothèse **(LL)**, pour toute donnée initiale  $y_0 \in \mathbb{R}^n$ , il y a existence et unicité locale de la solution de l'EDO.*

**Remarque :** Le théorème est évidemment admis dans ce cours : nous rappelons néanmoins qu'il s'obtient via un argument de point fixe pour les applications contractantes dans l'espace de Banach  $C([0, \tau])$  pour  $\tau > 0$  assez petit. Pour mettre en place cet argument de point fixe, on intègre l'équation :

$$y(t) = y_0 + \int_0^t f(s, y(s))ds ,$$

et l'application contractante est  $T : C([0, \tau]) \rightarrow C([0, \tau])$  définie par :

$$Ty(t) = y_0 + \int_0^t f(s, y(s))ds .$$

Nous considérons maintenant trois exemples typiques d'EDO dans  $\mathbb{R}$ .

- $\dot{y}(t) = y(t)$ . Il est plus que clair que le Théorème de Cauchy-Lipschitz s'applique car  $f(y) = y$  est même globalement lipschitzienne. En fait, on sait même calculer la solution :  $y(t) = y_0 \exp(t)$ . On a donc existence et unicité, non seulement localement mais “globalement” (i.e. pour tous temps, positifs et négatifs).

- $\dot{y}(t) = [y(t)]^2$ . Il est un peu moins clair que le Théorème de Cauchy-Lipschitz s'applique mais on vérifie tout de même facilement que  $f(y) = y^2$  est localement lipschitzienne car :

$$|y_1^2 - y_2^2| \leq 2R|y_1 - y_2| ,$$

pour tous  $|y_1|, |y_2| \leq R$  (utiliser ou bien le Théorème des accroissements finis qui dit, au passage, que toute fonction  $C^1$  est localement lipschitzienne, ou bien une identité remarquable). Là encore on sait calculer la solution :  $y(t) = 0$  si  $y_0 = 0$  et :

$$y(t) = \frac{y_0}{1 - ty_0} ,$$

si  $y_0 \neq 0$ . On a donc existence et unicité locale mais pas globale car si  $y_0 > 0$ , la solution tend vers  $+\infty$  quand  $t$  tend vers  $y_0^{-1}$ . On a donc un

exemple où la solution ne peut pas être prolongée à  $\mathbb{R}$  tout entier. Comme on le reverra plus loin, la possibilité de pouvoir étendre la solution  $y$  pour tout temps dépend d'hypothèses sur la croissance de  $f$  en  $y$  à l'infini. Et on peut toujours le faire si  $f$  est sous-linéaire, i.e. s'il existe  $K > 0$  tel que :

$$|f(t, y)| \leq K(1 + |y|) ,$$

pour tout  $t \in [0, T]$  (ou  $]0, +\infty[$ ) et  $y \in \mathbb{R}$ , ce qui n'est pas le cas pour  $f(y) = y^2$ .

•  $\dot{y}(t) = [y(t)]^{1/3}$  avec  $y_0 = 0$ . Dans ce dernier cas, le Théorème de Cauchy-Lipschitz ne s'applique pas car la fonction  $f(y) = y^{1/3}$  n'est pas localement lipschitzienne à cause du point 0 où la pente est infinie. On a pourtant existence locale car  $y(t) = 0$  est solution mais pas unicité car on a une autre solution de la forme  $y(t) = ct^{3/2}$  si  $t > 0$  et  $y(t) = 0$  si  $t < 0$ . Ce cas relève du *Théorème de Peano* qui donne l'existence locale (mais pas l'unicité) quand  $f$  est seulement continue. On voit donc que le caractère lipschitz de  $f$  est surtout utile pour l'unicité.

Pour simplifier encore plus, nous allons supposer que  $f$  est globalement lipschitzienne, c'est-à-dire que  $f$  satisfait **(LL)** avec une constante  $L$  qui ne dépend ni de  $T'$  ni de  $R$ . Nous supposons, de plus que  $f$  est continue en temps et nous noterons cette hypothèse (GL).

### 3.1.2 Effets des perturbations

Nous avons déjà vu plusieurs fois qu'il est important de mesurer l'effet des perturbations diverses : erreurs sur les données ou erreurs d'arrondies, incertitudes sur le modèle...etc. Cette section va montrer comment ces perturbations se transmettent au cours du temps, soit par une estimation "grossière" qui utilisera un outil fondamental de l'étude des EDO, le *lemme de Gronwall*, soit par une approche un peu plus précise, la linéarisation.

Nous commençons par montrer que la solution reste bornée sur  $[0, T]$  si  $f$  est sous-linéaire :

**Proposition 3.1.** *Sous l'hypothèse (GL), on a :*

$$|y(t)| \leq |y_0| \exp(Lt) + \frac{M}{L}(\exp(Lt) - 1) \quad \text{pour tout } t \in [0, T] ,$$

où  $M := \|f(t, 0)\|_\infty$ , la norme infinie étant prise sur l'intervalle  $[0, T]$ .

**Preuve :** Pour  $\alpha > 0$  petit, on introduit les fonctions  $\varphi_\alpha$  définie pour  $t \in [0, T]$  par  $\varphi_\alpha(t) = (|y(t)|^2 + \alpha)^{1/2}$ . Cette fonction est de classe  $C^1$  et :

$$\dot{\varphi}_\alpha(t) = \frac{1}{\varphi_\alpha(t)}(y(t), \dot{y}(t)) = \frac{1}{\varphi_\alpha(t)}(y(t), f(t, y(t))) .$$

Par (GL), on a :

$$\begin{aligned} |f(t, y(t))| &= |[f(t, y(t)) - f(t, 0)] + f(t, 0)| \\ &\leq |f(t, y(t)) - f(t, 0)| + |f(t, 0)| \\ &\leq L|y(t)| + M . \end{aligned}$$

En appliquant l'inégalité de Cauchy-Schwarz au produit scalaire  $(y(t), f(t, y(t)))$ , on en déduit :

$$\dot{\varphi}_\alpha(t) \leq \frac{1}{\varphi_\alpha(t)} |y(t)| (L|y(t)| + M) .$$

Mais on a clairement :

$$|y(t)| \leq \varphi_\alpha(t) \quad \text{pour tout } t ,$$

donc :

$$\dot{\varphi}_\alpha(t) \leq L\varphi_\alpha(t) + M .$$

On réécrit cette inégalité sous la forme :

$$\dot{\varphi}_\alpha(t) - L\varphi_\alpha(t) \leq M ,$$

puis on multiplie par  $\exp(-Lt)$  :

$$\exp(-Lt)\dot{\varphi}_\alpha(t) - L\exp(-Lt)\varphi_\alpha(t) \leq M\exp(-Lt) .$$

Le premier membre est une dérivée exacte [de  $\exp(-Lt)\varphi_\alpha(t)$ ] et, en intégrant de 0 à  $t$ , on obtient :

$$\exp(-Lt)\varphi_\alpha(t) - \varphi_\alpha(0) \leq \int_0^t M\exp(-Ls)ds = \frac{M}{L}(1 - \exp(-Lt)) .$$

D'où :

$$\varphi_\alpha(t) \leq \varphi_\alpha(0)\exp(Lt) + \frac{M}{L}(\exp(Lt) - 1) .$$

Il suffit alors de faire tendre  $\alpha$  vers 0 dans cette inégalité puisque  $\varphi_\alpha(t) \rightarrow |y(t)|$ .  $\square$

L'argument de la preuve précédente est un cas particulier d'un résultat plus général que l'on peut décliner de plusieurs manières, par exemple :

**Lemme 3.1. (Lemme de Gronwall)**

Si  $\chi : [0, T] \rightarrow \mathbb{R}$  est une fonction continue qui satisfait :

$$\chi(t) \leq \int_0^t \chi(s)\psi(s)ds + r(t) ,$$

où  $\psi \geq 0$  et  $r$  sont aussi des fonctions continues alors :

$$\chi(t) \leq \int_0^t r(s)\psi(s) \exp\left(\int_s^t \psi(\tau)d\tau\right) ds + r(t) .$$

**Idée de preuve :** on pose  $f(t) = \int_0^t \chi(s)\psi(s)ds$ . En multipliant l'inégalité satisfaite par  $\chi$  par  $\psi(t)$ , on aboutit à :

$$f'(t) \leq \psi(t)f(t) + r(t)\psi(t) ,$$

et on laisse la suite à la libre imagination du lecteur...

**NB :** une estimation de  $f$  donne une estimation de  $\chi$  puisque  $\chi(t) \leq f(t) + r(t)$ .

On considère maintenant la solution  $y_\varepsilon$  de l'EDO perturbée :

$$\dot{y}_\varepsilon(t) = f(t, y_\varepsilon(t)) + \varepsilon_1 g(t) \quad \text{dans } ]0, T[ ,$$

auquel il faut adjoindre une “condition initiale perturbée” :

$$y_\varepsilon(0) = y_0 + \varepsilon_0 \alpha ,$$

où  $\varepsilon = (\varepsilon_0, \varepsilon_1)$ ,  $\varepsilon_0, \varepsilon_1$  étant des paramètres petits,  $g$  est une fonction continue et  $\alpha \in \mathbb{R}^n$ . Ce problème admet, bien sûr, une solution par le Théorème de Cauchy-Lipschitz.

**Théorème 3.2.** *Pour tout  $t \in [0, T]$ , on a :*

$$|y_\varepsilon(t) - y(t)| \leq |\varepsilon_0 \alpha| \exp(Lt) + \int_0^t \exp(L(t-s)) |\varepsilon_1 g(s)| ds .$$

De plus, si  $f$  est de classe  $C^2$  alors :

$$|y_\varepsilon(t) - y(t) - \varepsilon_0 z_0(t) - \varepsilon_1 z_1(t)| \leq C(\varepsilon_0^2 + \varepsilon_1^2) ,$$

où les **correcteurs**  $z_0, z_1$  satisfont les **équations linéarisées** :

$$\begin{cases} \dot{z}_0(t) = D_y f(t, y(t)) z_0(t) \\ z_0(0) = \alpha \end{cases} \quad \text{et} \quad \begin{cases} \dot{z}_1(t) = D_y f(t, y(t)) z_1(t) + g(t) \\ z_1(0) = 0 \end{cases}$$

où  $D_y f$  désigne la dérivée (partielle) par rapport à la variable  $y$ .

**Preuve :** On ne va prouver en détails que le premier résultat, le second étant laissé en exercice avec quelques indications.

On pose  $\chi(t) = (|y_\varepsilon(t) - y(t)|^2 + \alpha)^{1/2}$  et on s'intéresse aux propriétés de cette fonction. On a :

$$\chi(t) \geq |y_\varepsilon(t) - y(t)| \quad \text{pour tous } t ,$$

et par des arguments analogues à ceux utilisés dans la preuve de la Proposition 3.1 :

$$\begin{aligned} \dot{\chi}(t) &= \frac{1}{\chi(t)} (y_\varepsilon(t) - y(t), f(t, y_\varepsilon(t)) + \varepsilon_1 g(t) - f(t, y(t))) \\ &\leq \frac{1}{\chi(t)} |y_\varepsilon(t) - y(t)| (|f(t, y_\varepsilon(t)) - f(t, y(t))| + \varepsilon_1 |g(t)|) \\ &\leq \frac{1}{\chi(t)} |y_\varepsilon(t) - y(t)| (L |y_\varepsilon(t) - y(t)| + \varepsilon_1 |g(t)|) \\ &\leq L \chi(t) + \varepsilon_1 |g(t)| . \end{aligned}$$

On se retrouve dans une situation quasi-analogue à celle de preuve du la Proposition 3.1 et on conclut de la même manière.

Pour la seconde partie du résultat, il faut introduire  $\chi(t) = (|y_\varepsilon(t) - y(t) - \varepsilon_0 z_0(t) - \varepsilon_1 z_1(t)|^2 + \alpha)^{1/2}$ . Seuls les calculs sur les termes en  $f$  sont différents car il faut utiliser la formule de Taylor avec reste intégral qui donne :

$$f(t, y_2) = f(t, y_1) + D_y f(t, y_1)(y_2 - y_1) + O(|y_2 - y_1|^2),$$

où le terme  $O(|y_2 - y_1|^2)$  est contrôlé de manière uniforme sur tout compact. En utilisant ce nouvel ingrédient et la première estimation de  $|y_\varepsilon(t) - y(t)|$ , la preuve est “straightforward”.  $\square$

### 3.1.3 Régularité de la solution

Nous terminons cette partie théorique par un résultat de régularité sur la solution  $y$  de l'EDO. Nous formulons ce résultat en dimension 1, c'est-à-dire quand  $y$  est à valeurs dans  $\mathbb{R}$ .

**Théorème 3.3.** *Si la fonction  $f$  est de classe  $C^p$  sur  $[0, T] \times \mathbb{R}$  alors  $y$  est de classe  $C^{p+1}$  et :*

$$y^{(k+1)}(t) = f^{[k]}(t, y(t)) \quad \text{dans } ]0, T[ \quad (k = 0, 1, \dots, p),$$

où les fonctions  $f^{[k]}$  sont définies sur  $[0, T]$  par :

$$f^{[0]}(t, z) = f(t, z),$$

$$f^{[k+1]}(t, z) = \frac{\partial}{\partial t} f^{[k]}(t, z) + f(t, z) \frac{\partial}{\partial y} f^{[k]}(t, z),$$

pour  $k = 0, 1, \dots, p-1$ .

**Preuve :** La preuve se fait facilement par récurrence.  $\square$

## 3.2 La méthode d'Euler

### 3.2.1 Présentation de la méthode

Pour résoudre numériquement l'EDO, on va se donner une *grille*, c'est-à-dire des points  $t_0 = 0 < t_1 < t_2 < t_3 < \dots < t_N = T$ , et on va essayer de calculer une “bonne” approximation des valeurs de la solution en tous ces points, c'est-à-dire des valeurs  $y_i \simeq y(t_i)$ .

Numériquement le sens de “bonne approximation” n'est pas absolu car, d'une part, l'ordre de grandeur joue un rôle (une approximation à 100 000

ans près peut être excellente si on est géologue...) et, d'autre part, le temps mis pour calculer la solution peut être un facteur important (quel est l'intérêt d'avoir un résultat très précis s'il faut 10 ans pour l'obtenir?). Il y a toujours, dans les méthodes numériques un "rapport qualité - prix" : précision vs temps de calcul ou complexité.

Essentiellement il y a deux approches pour calculer les  $y_i$  qui, dans le cas de la méthode d'Euler, vont aboutir au même résultat : soit on approche directement l'EDO par "différences finies" en utilisant une approximation de la dérivée  $y'(t)$ , soit on intègre l'EDO, se rapprochant ainsi de la preuve d'existence.

L'approche par "*différences finies*" consiste à approcher  $y'(t_i)$  dans l'esprit de la section 2.3 ; par exemple :

$$y'(t_i) \simeq \frac{y(t_{i+1}) - y(t_i)}{t_{i+1} - t_i} .$$

L'EDO se réécrit alors sous la forme :

$$\frac{y(t_{i+1}) - y(t_i)}{t_{i+1} - t_i} \simeq f(t_i, y(t_i)) ,$$

et donc :

$$y(t_{i+1}) \simeq y(t_i) + (t_{i+1} - t_i)f(t_i, y(t_i)) .$$

Ceci suggère que l'on peut calculer les  $y_k$  via la relation de récurrence :

$$y_{i+1} = y_i + (t_{i+1} - t_i)f(t_i, y_i) ,$$

le terme  $y_0$  étant connu (donnée initiale). C'est la *méthode d'Euler*.

La deuxième approche, qui va nous conduire au même résultat mais avec une philosophie très différente, consiste à intégrer l'équation de  $t_i$  à  $t_{i+1}$  :

$$y(t_{i+1}) = y(t_i) + \int_{t_i}^{t_{i+1}} f(s, y(s))ds .$$

Si on applique la méthode des rectangles à l'intégrale de la manière suivante :

$$\int_{t_i}^{t_{i+1}} f(s, y(s))ds \simeq (t_{i+1} - t_i)f(t_i, y(t_i)) ,$$

on retrouve les calculs ci-dessus et la méthode d'Euler.

**Remarque :** En pratique, il peut être intéressant d'utiliser une subdivision adaptée avec plus de points aux endroits où la fonction  $y$  varie beaucoup et moins de points aux endroits où les variations sont faibles. Mais choisir une telle subdivision (ou faire en sorte que l'ordinateur choisisse automatiquement cette subdivision = schémas adaptatifs) n'est pas toujours simple.

### 3.2.2 Étude de l'erreur (I)

Désormais nous nous plaçons dans le cadre d'une grille uniforme :

$$t_{i+1} - t_i = h = \frac{T}{N} .$$

La méthode d'Euler s'écrit alors :

$$y_{i+1} = y_i + hf(t_i, y_i) .$$

On notera :

$$e_i = y(t_i) - y_i ,$$

l'erreur commise au point  $t_i$  et :

$$\varepsilon_i = y(t_{i+1}) - y(t_i) - hf(t_i, y(t_i)) ,$$

l'erreur de "consistance" ; c'est l'erreur systématique commise sur  $y_{i+1}$  : même si on a calculé exactement la valeur à l'instant  $t_i$  ( $y_i \simeq y(t_i)$ ), on a une erreur sur  $y(t_{i+1})$  qui est  $\varepsilon_i$ .

Pour évaluer l'erreur, on procède comme suit :

$$\begin{aligned} e_{i+1} &= y(t_{i+1}) - y_{i+1} \\ &= [y(t_{i+1}) - y(t_i) - hf(t_i, y(t_i))] + [y(t_i) - y_i] + [y_i + hf(t_i, y_i)] + \\ &\quad h[f(t_i, y(t_i)) - hf(t_i, y_i)] - y_{i+1} \\ &= \varepsilon_{i+1} + e_i + h[f(t_i, y(t_i)) - hf(t_i, y_i)] . \end{aligned}$$

D'où :

$$\begin{aligned} |e_{i+1}| &\leq |\varepsilon_{i+1}| + |e_i| + h|f(t_i, y(t_i)) - hf(t_i, y_i)| \\ &\leq |\varepsilon_{i+1}| + (1 + Lh)|e_i| , \end{aligned}$$

en utilisant le caractère lipschitz de  $f$ .

Il reste à estimer  $|\varepsilon_{i+1}|$ . En utilisant l'équation, on a :

$$\begin{aligned} |\varepsilon_{i+1}| &= |y(t_{i+1}) - y(t_i) - hf(t_i, y(t_i))| \\ &= \left| \int_{t_i}^{t_{i+1}} y'(s) ds - \int_{t_i}^{t_{i+1}} y'(t_i) ds \right| \\ &= \left| \int_{t_i}^{t_{i+1}} [y'(s) - y'(t_i)] ds \right| \\ &\leq h \sup_{s \in [t_i, t_{i+1}]} |y'(s) - y'(t_i)| \\ &\leq h \cdot \omega(h, y') , \end{aligned}$$

où  $\omega(\cdot, y')$  est le module de continuité de la fonction (continue)  $y'$  sur  $[0, T]$ .



Finalement on a l'estimation suivante de l'erreur  $e_{i+1}$  :

$$|e_{i+1}| \leq (1 + Lh)|e_i| + h.\omega(h, y') .$$

Pour conclure on utilise le :

**Lemme 3.2. (Lemme de Gronwall discret)**

Si  $(\theta_i)_i$  est une suite de réels positifs qui satisfait :

$$\theta_{i+1} \leq (1 + A)\theta_i + B ,$$

où  $A, B$  sont des constantes strictement positives alors :

$$\theta_i \leq \exp(iA)\theta_0 + \frac{\exp(iA) - 1}{A} B .$$

On utilise d'abord le lemme avec  $A = Lh$  et  $B = h.\omega(h, y')$  :

$$|e_i| \leq \exp(Lih)|e_0| + \frac{\exp(Lih) - 1}{Lh} h.\omega(h, y') .$$

Mais  $ih = t_i$  et (a priori)  $e_0 = 0$  [ou du moins,  $e_0$  est très petit] donc :

$$|e_i| \leq \frac{\exp(Lt_i) - 1}{L} \omega(h, y') \leq \frac{\exp(LT) - 1}{L} \omega(h, y') .$$

On vient donc de prouver le :

**Théorème 3.4.** *Sous l'hypothèse (GL) :*

$$\max_{0 \leq i \leq N} |e_i| \leq \frac{\exp(LT) - 1}{L} \omega(h, y') .$$

En particulier,  $\max_{0 \leq i \leq N} |e_i| \rightarrow 0$  quand  $N \rightarrow +\infty$ .

On donne maintenant la :

**Preuve du Lemme de Gronwall discret :** On note  $(u_i)_i$  la suite définie par :

$$u_i = \frac{\theta_i}{(1 + A)^i} .$$

En divisant la propriété satisfait par la suite  $(\theta_i)_i$  par  $(1 + A)^{i+1}$ , on voit que :

$$u_{i+1} \leq u_i + \frac{B}{(1 + A)^{i+1}} .$$

Une récurrence immédiate montre que :

$$u_i \leq u_0 + \sum_{k=0}^{i-1} \frac{B}{(1 + A)^{k+1}} = u_0 + \frac{B}{1 + A} \frac{1 - a^i}{1 - a} ,$$

où  $a = 1/(1 + A)$ . Comme :

$$\frac{1}{1-a} = \frac{1+A}{A} ,$$

il en résulte :

$$u_i \leq u_0 + B \frac{1-a^i}{A} ,$$

et donc :

$$\theta_i \leq (1+A)^i u_0 + B \frac{(1+A)^i - 1}{A} .$$

Il reste à remarquer que  $1 + A \leq \exp(A)$ , ce qui est clair puisque :

$$\exp(A) = 1 + A + \frac{A^2}{2!} + \cdots + \frac{A^n}{n!} + \cdots .$$

□

À titre d'exercice, on pourra démontrer la :

**Proposition 3.2.** *Si  $y_h$  est la fonction affine par morceaux telle que  $y_h(t_i) = y_i$  pour tout  $i$ , on a :*

$$\|y_h - y\|_\infty \rightarrow 0 \quad \text{quand } h \rightarrow 0 .$$

### 3.2.3 Étude de l'erreur (II)

Le section précédente donne une estimation de convergence qui dépend du module de continuité de  $y'$ . Mais la fonction  $y$  est inconnue donc ce résultat n'est pas satisfaisant car il n'est pas explicite. Nous allons maintenant donner une autre estimation qui ne dépend que des données, c'est-à-dire de  $f$ .

Pour cela, on étudie le module de continuité de  $y'$  :

$$\begin{aligned} |y'(t) - y'(s)| &= |f(t, y(t)) - f(s, y(s))| \\ &= |f(t, y(t)) - f(s, y(t)) + f(s, y(t)) - f(s, y(s))| \\ &\leq |f(t, y(t)) - f(s, y(t))| + |f(s, y(t)) - f(s, y(s))| \\ &\leq |f(t, y(t)) - f(s, y(t))| + L|y(t) - y(s)| \end{aligned}$$

D'après la Proposition 3.1,  $|y(t)| \leq D$  pour une certaine constante  $D$  sur l'intervalle  $[0, T]$  et le premier terme est estimé par le module de continuité de  $f$  sur  $[0, T] \times \overline{B}(0, D)$ , noté  $\omega_D(\cdot, f)$ .

Quant au second, par le Théorème des Accroissements Finis :

$$|y(t) - y(s)| \leq M_f |t - s| ,$$

où :

$$M_f = \max_{[0,T] \times \bar{B}(0,D)} |f(t,y)| .$$

Finalement :

$$\omega(h, y') \leq \omega_D(h, f) + LM_f h ,$$

ce qui donne le résultat suivant qui était notre objectif :

**Théorème 3.5.** *Sous l'hypothèse (GL) :*

$$\max_{0 \leq i \leq N} |e_i| \leq \frac{\exp(LT) - 1}{L} (\omega_D(h, f) + LM_f h) .$$

### 3.3 Étude générale des méthodes à un pas

On conserve, dans cette section, une grille uniforme de pas  $h = \frac{T}{N}$ . Les méthodes à un pas sont des méthodes de la forme :

$$\begin{cases} y_{i+1} &= y_i + h\Phi(t_i, y_i, h) \\ y_0 &= y_{0,h} \end{cases}$$

où  $\Phi$  est une fonction continue sur  $[0, T] \times \mathbb{R}^n \times [0, H]$ ,  $H$  désignant un pas de discrétisation maximal.

#### 3.3.1 Propriétés importantes d'une méthode à un pas

##### • CONSISTANCE

**Définition 3.1.** *On appelle erreur de consistance de la méthode à un pas, la quantité :*

$$\Sigma_h = \sum_{i=0}^{N-1} |y(t_{i+1}) - y(t_i) - h\Phi(t_i, y(t_i), h)| .$$

La méthode est dite consistante si  $\Sigma_h \rightarrow 0$  quand  $h \rightarrow 0$ .

La quantité  $y(t_{i+1}) - y(t_i) - h\Phi(t_i, y(t_i), h)$  est l'analogie de ce que nous avons noté  $\varepsilon_i$  ci-dessus ; nous conserverons au besoin cette notation.

##### • STABILITÉ

**Définition 3.2.** *La méthode à un pas est dite stable s'il existe deux constantes  $S_1, S_2$  telles que, si  $(\tilde{y}_i)_i$  est défini par :*

$$\begin{cases} \tilde{y}_{i+1} &= \tilde{y}_i + h\Phi(t_i, \tilde{y}_i, h) + \varepsilon_i \\ \tilde{y}_0 &= \tilde{y}_{0,h} \end{cases} ,$$

alors :

$$\max_i |y_i - \tilde{y}_i| \leq S_1 |y_{0,h} - \tilde{y}_{0,h}| + S_2 \sum_{i=0}^{N-1} |\varepsilon_i| .$$

Bien entendu, la méthode est dite **convergente** si  $\max_i |y_i - y(t_i)| \rightarrow 0$  quand  $h \rightarrow 0$ .

**Théorème 3.6.** *Toute méthode stable et consistante converge à condition que  $y_{0,h} \rightarrow y(0)$  quand  $h \rightarrow 0$ .*

La dernière condition étant toujours satisfaite en pratique, l'étude de la convergence des méthodes à un pas se réduit à l'étude de leur consistance et de leur stabilité, ce qui est plus simple comme on va le voir.

**Preuve :** Si on note  $\tilde{y}_i = y(t_i)$ , on a par définition de  $\varepsilon_i$  (juste après la définition de la consistance) :

$$\tilde{y}_{i+1} = \tilde{y}_i + h\Phi(t_i, \tilde{y}_i, h) + \varepsilon_i .$$

Et  $\tilde{y}_0 = y(0)$ . Puisque la méthode est stable :

$$\max_i |y_i - y(t_i)| \leq S_1 |y_{0,h} - y(0)| + S_2 \sum_{i=0}^{N-1} |\varepsilon_i| .$$

Hors, les deux quantités du membre de droite tendent vers 0 quand  $h$  tend vers 0 par consistance, donc le résultat est acquis.  $\square$

### 3.3.2 Condition nécessaire et suffisante de consistance

**Théorème 3.7.** *La méthode à un pas est consistante si et seulement si :*

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

pour tous  $t \in [0, T]$ ,  $z \in \mathbb{R}^n$ .

**Preuve :** On ne va vraiment détailler que la condition suffisante.

$$\begin{aligned} \varepsilon_i &= y(t_{i+1}) - y(t_i) - h\Phi(t_i, y(t_i), h) \\ &= \int_{t_i}^{t_{i+1}} [f(s, y(s)) - \Phi(t_i, y(t_i), h)] ds . \end{aligned}$$

Mais, pour  $s \in [t_i, t_{i+1}]$  :

$$\begin{aligned} |f(s, y(s)) - \Phi(t_i, y(t_i), h)| &\leq |f(s, y(s)) - f(t_i, y(t_i))| + \\ &|f(t_i, y(t_i)) - \Phi(t_i, y(t_i), 0)| + |\Phi(t_i, y(t_i), 0) - \Phi(t_i, y(t_i), h)| \end{aligned}$$

Le premier et le troisième terme sont des termes petits (uniformément en  $i$ ) par l'uniforme continuité de  $f$ ,  $y$  et  $\Phi$  ; on les estime par un  $\delta(h)$  qui tend vers 0 avec  $h$ .

Si le terme du milieu est nul (condition de consistance) alors  $|\varepsilon_i| \leq h\delta(h)$  et  $\Sigma_h \leq Nh\delta(h) = T\delta(h) \rightarrow 0$  quand  $h \rightarrow 0$ . D'où la consistance.

NB : la condition suffisante se prouve en examinant la preuve d'un peu plus près : si si le terme du milieu n'est pas nul...  $\square$

### 3.3.3 Condition suffisante de stabilité

**Théorème 3.8.** *La méthode à un pas est stable si  $\Phi(t, z, h)$  est lipschitzienne en  $z$  pour tous  $t \in [0, T]$ ,  $h \in [0, H]$  avec une constante de lipschitz indépendante de  $t$  et  $h$ .*

**Preuve :** On note  $\theta_i = |y_i - \tilde{y}_i|$ . On a :

$$\theta_{i+1} = |y_i + h\Phi(t_i, y_i, h) - \tilde{y}_i - h\Phi(t_i, \tilde{y}_i, h) - \varepsilon_i|.$$

D'où :

$$\theta_{i+1} \leq \theta_i + h|\Phi(t_i, y_i, h) - \Phi(t_i, \tilde{y}_i, h)| + |\varepsilon_i|.$$

Si  $\Phi$  est lipschitzienne en sa deuxième variable de constante de lipschitz  $\tilde{L}$  :

$$\theta_{i+1} \leq (1 + \tilde{L}h)\theta_i + |\varepsilon_i|.$$

Une petite amélioration du Lemme de Gronwall discret nous donne :

$$\theta_{i+1} \leq \exp(\tilde{L}(i+1)h)\theta_0 + \sum_{k=0}^i \exp(\tilde{L}(i-k)h)|\varepsilon_k|.$$

En majorant  $\exp(\tilde{L}(i+1)h)$  et  $\exp(\tilde{L}(i-k)h)$  par  $\exp(\tilde{L}T)$ , on a le résultat avec  $S_1 = S_2 = \exp(\tilde{L}T)$ .  $\square$

### 3.3.4 Ordre d'un schéma

La question que l'on se pose ici est la suivante : peut-on avoir une meilleure précision que dans le cas de la méthode d'Euler en choisissant bien la méthode à un pas et comment faut-il la choisir ?

**Définition 3.3.** *On dit qu'une méthode à un pas est d'ordre  $p \geq 1$  si, pour toute solution de l'EDO, il existe une constante  $C > 0$  telle que :*

$$|\varepsilon_i| = |y(t_{i+1}) - y(t_i) - h\Phi(t_i, y(t_i), h)| \leq Ch^{p+1}.$$

Pourquoi le " $p+1$ " dans l'ordre  $p$  ? Deux raisons concourantes :

1.  $|\frac{y(t_{i+1}) - y(t_i)}{h} - \Phi(t_i, y(t_i), h)| \leq Ch^p$  et la quantité considérée approche l'équation à l'ordre  $p$ .
2.  $\Sigma_h = \sum_{i=0}^{N-1} |y(t_{i+1}) - y(t_i) - h\Phi(t_i, y(t_i), h)| \leq Ch^p$ , donc ordre  $p =$  erreur en  $h^p$ .

Cette dernière idée est justifiée par le résultat suivant dont la preuve est immédiate à partir du résultat de convergence :

**Corollaire 3.1.** *Si la méthode à un pas est d'ordre  $p \geq 1$  et si  $|y(0) - y_{0,h}| \leq \tilde{C}h^p$  alors  $\max_i |y_i - y(t_i)| \leq \hat{C}h^p$ .*

Nous donnons maintenant une condition nécessaire et suffisante pour qu'une méthode à un pas soit d'ordre  $p$  dans  $\mathbb{R}$ .

**Théorème 3.9.** *On suppose que  $f$  est de classe  $C^p$  et que, pour  $k \leq p$ , les dérivées partielles  $\frac{\partial^k \Phi}{\partial h^k}$  existent et sont continues sur  $[0, T] \times \mathbb{R} \times [0, H]$ . Alors la méthode à un pas est d'ordre  $p$  si, pour tous  $t \in [0, T]$ ,  $z \in \mathbb{R}$  :*

$$\begin{aligned} \Phi(t, z, 0) &= f(t, z) \\ \frac{\partial \Phi}{\partial h}(t, z, 0) &= \frac{1}{2}f^{[1]}(t, z) \\ &\vdots \\ \frac{\partial^k \Phi}{\partial h^k}(t, z, 0) &= \frac{1}{k+1}f^{[k]}(t, z), \quad k \leq p-1. \end{aligned}$$

Dans ce cas :

$$|\varepsilon_i| = \frac{1}{i!}h^{p+1} \left( \frac{1}{p+1}f^{[p]}(t_i, y(t_i)) - \frac{\partial^p \Phi}{\partial h^p}(t_i, y(t_i), 0) \right) + o(h^{p+1}).$$

**Preuve :** Comme  $f$  est de classe  $C^p$ ,  $y$  est de classe  $C^{p+1}$  et  $y^{(k)}(t) = f^{[k-1]}(t, y(t))$  (voir la partie de cours relative à la régularité des solutions).

Par la formule de Taylor, on a :

$$\begin{aligned} y(t_{i+1}) - y(t_i) &= \sum_{k=1}^p \frac{h^k}{k!} y^{(k)}(t_i) + O(h^{p+1}) \\ &= \sum_{k=1}^p \frac{h^k}{k!} f^{[k-1]}(t_i, y(t_i)) + O(h^{p+1}) \end{aligned}$$

et :

$$\Phi(t_i, y(t_i), h) = \sum_{k=0}^p \frac{h^k}{k!} \frac{\partial^k \Phi}{\partial h^k}(t_i, y(t_i), 0) + o(h^p).$$

Il en résulte que :

$$\begin{aligned} h\Phi(t_i, y(t_i), h) &= \sum_{k=0}^p \frac{h^{k+1}}{k!} \frac{\partial^k \Phi}{\partial h^k}(t_i, y(t_i), 0) + o(h^{p+1}) \\ &= \sum_{k=1}^{p+1} \frac{h^k}{(k-1)!} \frac{\partial^{k-1} \Phi}{\partial h^{k-1}}(t_i, y(t_i), 0) + o(h^{p+1}) \end{aligned}$$

et :

$$\varepsilon_i = \sum_{k=1}^p \left[ \frac{1}{k} f^{[k-1]}(t_i, y(t_i)) - \frac{\partial^{k-1} \Phi}{\partial h^{k-1}}(t_i, y(t_i), 0) \right] \frac{h^k}{(k-1)!} + O(h^{p+1}) .$$

Sous les hypothèses du théorème, tous les termes entre crochets sont nuls et le résultat est acquis.

On peut le préciser avec la formule de Taylor avec reste intégral, ce qui donne la deuxième partie du théorème.  $\square$

### 3.3.5 Exemples

On va chercher la meilleure fonction  $\Phi$ , c'est-à-dire celle qui donne le meilleur ordre de convergence, parmi celles qui sont de la forme :

$$\Phi(t, z, h) = a_1 f(t, z) + a_2 f(t + p_1 h, z + p_2 h f(t, z)) ,$$

où  $a_1, a_2, p_1, p_2$  sont des paramètres à fixer "au mieux".

On a une méthode consistante d'ordre 1 si :

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

donc si :

$$a_1 f(t, z) + a_2 f(t, z) = f(t, z) ,$$

d'où la première condition  $a_1 + a_2 = 1$ .

Pour avoir de l'ordre 2, il faut que :

$$\frac{\partial \Phi}{\partial h}(t, z, 0) = \frac{1}{2} f^{[1]}(t, z) = \frac{1}{2} \left( \frac{\partial f}{\partial t}(t, z) + f(t, z) \frac{\partial f}{\partial y}(t, z) \right) .$$

Or :

$$\frac{\partial \Phi}{\partial h}(t, z, h) = a_2 p_1 \frac{\partial f}{\partial t}(t + p_1 h, z + p_2 h f(t, z)) + a_2 p_2 f(t, z) \frac{\partial f}{\partial y}(t + p_1 h, z + p_2 h f(t, z)) .$$

La deuxième condition est donc  $a_2 p_1 = a_2 p_2 = \frac{1}{2}$ .

Par contre, on vérifie facilement (le faire!) que l'on ne peut pas aller plus loin. En prenant,  $a_2 = \alpha$  comme paramètre, on a une famille de méthodes à un pas d'ordre 2 :

$$\Phi(t, z, h) = (1 - \alpha) f(t, z) + \alpha f\left(t + \frac{h}{2\alpha}, z + \frac{h}{2\alpha} f(t, z)\right) .$$

Ces méthodes sont connues sous le nom de :

- $\alpha = 1$ , méthode de la tangente améliorée,
- $\alpha = 1/2$ , méthode d'Euler modifiée,
- $\alpha = 1$ , méthode de Heun.

### 3.4 Quelques éléments sur les méthodes de Runge-Kutta

Ces méthodes sont les plus utilisées : elles sont rentrées “en standard” dans la plupart des logiciels. Comment marchent-elles ?

On repart de l'idée fondamentale qui consiste à écrire :

$$y(t_{i+1}) = y(t_i) + \int_{t_i}^{t_{i+1}} f(s, y(s)) ds .$$

Nous avons vu dans la deuxième partie que pour calculer une intégrale, on utilise une formule de quadrature :

$$\int_0^1 \psi(s) ds \simeq \sum_{j=0}^q b_j \psi(c_j) ,$$

où  $c_0 < c_1 < \dots < c_q$ , ce qui, en prenant  $\psi(s) = f(t_i + sh, y(t_i + sh))$ , nous donne :

$$y(t_{i+1}) \simeq y(t_i) + h \sum_{j=0}^q b_j f(t_{i,j}, y(t_{i,j})) ,$$

où  $t_{i,j} = t_i + c_j h$ . Ceci suggère une méthode que l'on peut écrire :

$$y_{i+1} = y_i + h \sum_{j=0}^q b_j k_{i,j} ,$$

où  $k_{i,j}$  est une approximation de  $f(t_{i,j}, y(t_{i,j}))$

Le problème, c'est qu'il faut encore calculer des approximations  $y_{i,j}$  des  $y(t_{i,j})$  pour avoir celle de  $k_{i,j}$  et ceci est fait via :

$$y_{i,j} = y_i + h \sum_{k=0}^q a_{i,k} f(t_{i,k}, y(t_{i,k})) .$$

Cette procédure a l'air d'induire des équations non-linéaires couplées difficiles à résoudre et pour que ce ne soit pas le cas, on suppose que les  $y_{i,j}$  ne dépendent que des points déjà calculés, c'est-à-dire des  $y_{i,k}$  pour  $k < j$ . On a donc :

$$y_{i,j} = y_i + h \sum_{k=0}^{j-1} a_{i,k} f(t_{i,k}, y(t_{i,k})) ,$$

et les  $y_{i,j}$ , ainsi que les  $k_{i,j}$ , sont calculés de proche en proche.

On résume souvent une méthode de Runge-Kutta grâce à un tableau de la forme :



### 3.4. QUELQUES ÉLÉMENT SUR LES MÉTHODES DE RUNGE-KUTTA 57

$c_1$	$a_{1,1}$	$\cdots$	$a_{1,q}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$
$c_q$	$a_{q,1}$	$\cdots$	$a_{q,q}$
	$b_1$	$\cdots$	$b_q$

#### Exemples :

- $q = 1$  : C'est la méthode d'Euler basée sur la méthode des rectangles.
- $q = 2$  C'est l'exemple de la section précédente, basée sur la méthode des trapèzes, avec le tableau :

0	0	0
$\beta$	$\beta$	0
	$1 - \frac{1}{2\beta}$	$\frac{1}{2\beta}$

NB :  $\beta = (2\alpha)^{-1}$ .

- $q = 4$  : la méthode de Runge-Kutta "classique" (la plus utilisée) basée sur la formule de Simpson :

0	0	0	0	0
1/2	1/2	0	0	0
1/2	0	1/2	0	0
1	0	0	1	0
	1/6	2/6	2/6	1/6

La méthode s'écrit aussi :

$$\Phi(t, y, h) = \frac{1}{6} [k_1 + 2k_2 + 2k_3 + k_4]$$

avec :

$$\begin{aligned} k_1 &= f(t, y) \\ k_2 &= f\left(t + \frac{h}{2}, y + \frac{h}{2}k_1\right) \\ k_3 &= f\left(t + \frac{h}{2}, y + \frac{h}{2}k_2\right) \\ k_4 &= f(t + h, y + hk_3) \end{aligned}$$

La méthode est d'ordre 4 mais il vaut mieux avoir Maple pour le vérifier !

### 3.5 Exercices

- 1) On considère les trois méthodes de résolution de  $y'(t) = f(t, f(t))$  :

$$\begin{array}{ll} \text{Euler} & y_{n+1} = y_n + h_n f(t_n, y_n) \\ \text{Euler Rétrograde} & y_{n+1} = y_n + h_n f(t_{n+1}, y_{n+1}) \\ \text{Point Milieu} & \begin{cases} y_{n+1/2} = y_n + (h_n/2) f(t_n, y_n) \\ p_n = f(t_n + h_n/2, y_{n+1/2}) \\ y_{n+1} = y_n + h_n p_n \end{cases} \end{array}$$

- Expliquer géométriquement à quoi correspondent ces 3 méthodes. Pourquoi la méthode d'Euler rétrograde est-elle aussi appelée méthode d'Euler implicite ?
  - Appliquer les trois méthodes à l'équation  $y'(t) = y(t)$ ,  $y(0) = 1$  et donner pour chacune l'expression de  $y_n$  en fonction de  $n$  en supposant le pas  $h$  constant.
  - Essayer de deviner l'ordre de chaque méthode.
- 2) On considère l'équation  $y'(t) = y^2(t)$ ,  $y(0) = 1$ . Déterminer explicitement la solution  $y$  ; que pensez-vous de la suite  $y_n$  correspondante à la méthode d'Euler ?
- 3) On considère l'équation différentielle suivante :

$$(E_1) \quad y'(t) = 150y(t) - 30, \quad y(0) = 1/5.$$

- Déterminer explicitement la solution de  $(E_1)$ . Quel comportement obtient-on si on remplace la condition initiale par  $y(0) = 1/5 + \varepsilon$  (représentant l'erreur d'arrondi dans le calcul de la donnée initiale) ?
- On s'intéresse maintenant à l'équation

$$(E_2) \quad y'(t) = -150y(t) + 30, \quad y(0) = 1/5.$$

Reprendre les mêmes questions qu'au a. pour  $(E_2)$ .

- On applique la méthode d'Euler à l'équation  $(E_2)$  ; donner une expression explicite de  $y_n$ . A quelle condition  $y_n$  est-elle une "bonne" approximation de la solution  $y$  ?
  - Même questions avec la méthode d'Euler rétrograde (ou implicite).
- 4) On veut justifier que la méthode du point milieu est d'ordre 2. Pour cela on calcule l'erreur de consistance

$$e_n = z(t_{n+1}) - y_{n+1},$$

où  $z$  est la solution exacte sur  $[t_n, t_{n+1}]$  de  $y' = f(t, y)$  avec  $z(t_n) = y_n$ .

- Montrer que  $e_n$  s'écrit  $e_n = \varepsilon_n + \varepsilon'_n$  avec

$$\begin{aligned} \varepsilon_n &= z(t_{n+1}) - z(t_n) - h_n z'(t_n + h_n/2), \\ \varepsilon'_n &= h_n \left( f(t_n + h_n/2, z(t_n + h_n/2)) - f(t_n + h_n/2, y_{n+1/2}) \right). \end{aligned}$$

- b. Montrer que  $\varepsilon_n = \frac{h_n^3}{24} + o(h_n^3)$  puis calculer de façon analogue  $\varepsilon'_n$  et conclure.

5) On considère l'équation différentielle :

$$y'(t) = f(t, y(t)) \quad y(0) = y_0 ,$$

où  $f : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  est une fonction de classe  $C^2$  et  $y_0 \in \mathbb{R}$ . On suppose qu'il existe une constante  $C > 0$  telle que, pour tout  $t \in \mathbb{R}$  et  $y \in \mathbb{R}$  :

$$\left| \frac{\partial f}{\partial y}(t, y) \right| \leq C .$$

- a. Rappelez très brièvement pourquoi cette condition implique l'existence et l'unicité de la solution  $y(\cdot)$  de l'équation différentielle qui est définie sur  $[0, +\infty[$ .
- b. Pour résoudre numériquement cette équation différentielle, on considère une famille de méthodes à un pas définie de la manière suivante : pour  $n \in \mathbb{N}$ ,  $t_n = nh$  où  $h$  est le pas de temps ; on considère des approximations  $y_n$  de  $y(t_n)$  que l'on calcule par la formule de récurrence :

$$y_{n+1} = y_n + h\phi(t_n, y_n, h) ,$$

où  $\phi$  est une fonction de la forme :

$$\phi(t, z, h) = a_1 f(t, z) + a_2 f(t + p_1 h, z + p_2 h f(t, z)) ,$$

avec  $a_1, a_2, p_1, p_2 \in \mathbb{R}$ .

- b. Donner des conditions sur les paramètres  $a_1, a_2, p_1, p_2$  pour que cette méthode soit stable et consistante.
- c. En déduire pour quelles valeurs de  $a_1, a_2, p_1, p_2$ , cette méthode est convergente.
- d. Donner des conditions sur les paramètres  $a_1, a_2, p_1, p_2$  pour que cette méthode soit d'ordre 2.
- e. En déduire la forme générale des fonctions  $\phi$  (définies comme ci-dessus) qui donnent une méthode d'ordre 2. (On pourra montrer qu'il s'agit d'une famille à un paramètre et utiliser par exemple le paramètre  $\alpha = a_1$ .)
- 6) Vérifier que parmi les méthodes de Runge-Kutta à deux points intermédiaires, les seules méthodes d'ordre 2 sont celles du type :

0	0	0
$\alpha$	$\alpha$	0
$1 - \frac{1}{2\alpha}$	$1 - \frac{1}{2\alpha}$	$\frac{1}{2\alpha}$

7) On considère la méthode de Runge-Kutta associée au tableau suivant :

0	0	0	0
$\frac{1}{3}$	$\frac{1}{3}$	0	0
$\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0
	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

- Décrire cette méthode (on donnera la récurrence qui permet de calculer  $y_{n+1}$  en fonction de  $y_n$  et des points intermédiaires  $(t_{n,i}, y_{n,i})$ ,  $i = 1..3$ ).
- A quelles méthodes d'intégration correspondent les différentes lignes ?
- Donner l'ordre de la méthode obtenue.