COURS DE L3 : ANALYSE NUMÉRIQUE

Laurent BRUNEAU
Université de Cergy-Pontoise

Table des matières

Chapitre 1

Calcul approché d'intégrales

L'objectif de ce chapitre est de voir quelques méthodes (simples) de calcul approché d'intégrales. D'autres méthodes un peu plus élaborées sont présentées en compléments dans le Chapitre 5. Avant de voir ces différentes méthodes, il y a deux points importants qu'il s'agit de toujours avoir en tête dès que l'on parle de calcul (ou résolution) approchée : la précision de la valeur obtenue (ou l'erreur commise) et la vitesse de convergence.

Erreur

Tout comme dans les chapitres suivants (résolution approché de $f(x) = 0$, résolution d'équations différentielles), il ne s'agira pas uniquement de *donner une valeur*. Comme on parle ici de valeur *approchée*, une valeur seule ne veut rien dire ! Dire que $3,14$ est une valeur approchée de π, sans autre précision, n'a pas plus d'intérêt et n'est pas plus correct que de dire que 5 est une valeur approchée de π. Quand on parlera de valeur approchée il faudra toujours préciser quelle est l'erreur (maximum) commise. 5 est en effet une valeur approchée de π, à 2 près, ce qui signifie que $|5-\pi| < 2$. Le nombre $3,14$ est aussi une valeur approchée de π à 2 près puisque $|3,14 - \pi| < 2$. Il est vrai que $3,14$ est une meilleur approximation puisqu'en fait c'est une valeur approchée à 10^{-2} de π : $|3,14 - \pi| < 10^{-2}$, alors que ce n'est pas le cas pour 5. Mais si vous savez a priori que $3,14$ est meilleur que 5 c'est parce que vous avez déjà rencontré π et que vous avez déjà une idée de sa valeur. Dans le cas contraire rien ne vous permet de vous faire une idée. Par exemple, on peut montrer que la suite u définie par $u_n = \sum_{k=1}^{n} \frac{1}{k} - \ln(n)$ converge. Sa limite est appelée la constante d'Euler et notée γ. Si je vous dis que 3 et $6,73$ sont des valeurs approchées de γ, qu'est ce que ça vous apprend sur cette dernière ? Laquelle de ces deux valeurs vous donne le plus d'information sur ce que vaut réellement γ ? Le fait qu'il y ait deux chiffres après la virgule à $6,73$ ne signifie rien, l'ordinateur (ou la calculatrice) a très bien pu me donner $3,00345$ que j'ai arrondi à $3,00$ et donc à 3. (En fait γ vaut $0,577$ à 10^{-3} près.) Cet exemple simple est là pour attirer votre attention sur le fait qu'une valeur approchée seule, sans dire quelle est l'erreur maximale possible, ne veut rien dire : n'importe quel nombre est une valeur approchée d'un autre, toute la question est de savoir "à combien près".

Vitesse de convergence

La notion de vitesse de convergence est liée à celle de l'erreur. La plupart des méthodes de calcul approché et de résolution approchée d'une équation reposent sur un algorithme qui permet d'obtenir ces valeurs de "proche en proche". Une des questions est alors de savoir quelle est la précision obtenue a priori après n étapes de calculs, l'idée étant surtout de savoir : quand dois-je arrêter l'algorithme pour être sûr d'avoir la précision souhaitée.

1.1 Méthodes des rectangles

1.1.1 Les sommes de Riemann

Les méthodes de calcul approché dites des rectangles sont naturelles et reposent sur la notion de sommes de Riemann que vous avez vues en L2. On rappelle brièvement celle-ci.

Soit $f : [a, b] \to \mathbb{R}$ une fonction continue par morceaux. Une subdivision σ de $[a, b]$ est une partie finie $\{a_0, \ldots, a_k\}$ telle que

$$a = a_0 < a_1 < a_2 < \cdots < a_k = b.$$

On note $|\sigma| = \max_{i=0,\ldots,k-1}(a_{i+1} - a_i)$ la longueur de la subdivision. Si pour tout $i = 0, \ldots, k-1$ on se donne $x_i \in [a_i, a_{i+1}]$ on dit que $\sigma^p = (\sigma, (x_i)_i)$ est une subdivision pointée de $[a, b]$. On note

$$S_{\sigma^p}(f) := \sum_{i=0}^{k-1} f(x_i)(a_{i+1} - a_i).$$

Cette somme est appelée somme de Riemann associée à la subdivision pointée σ^p, voir la Figure 2.1.

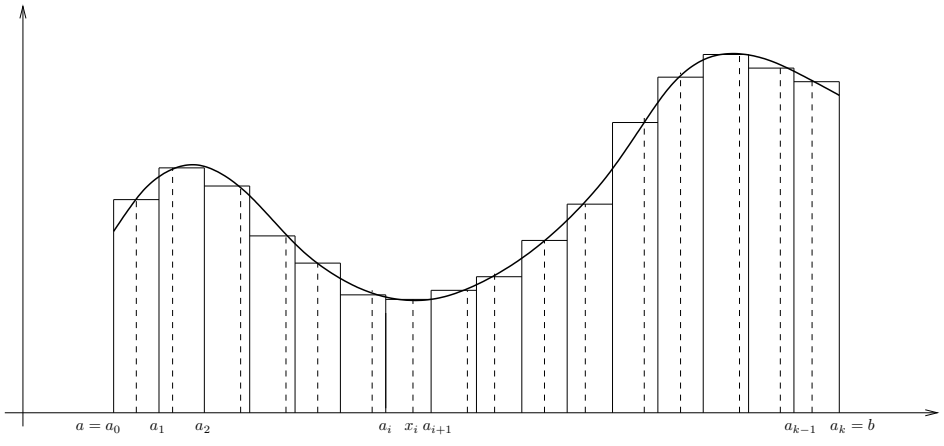

FIGURE 2.1 - Somme de Riemann associée à la subdivision pointée σ^p

La base de la théorie de l'intégration au sens de Riemann affirme que

Théorème 1.1. *Si f est continue par morceaux, pour toute suite $(\sigma_n^p)_n$ de subdivisions pointées de $[a,b]$ telle que $|\sigma_n^p| \to 0$ lorsque $n \to \infty$ on a*

$$\lim_{n\to\infty} S_{\sigma_n^p}(f) = \int_a^b f(x)\mathrm{d}x.$$

1.1.2 Rectangles à gauche / à droite

Les méthodes des rectangles à gauche et à droite reposent sur le Théorème 1.1 pour les choix de subdivisions pointées suivants :

- rectangles à gauche : $\sigma_n^p = \{a_0, \dots, a_n\}$ où $a_k = a + \frac{k(b-a)}{n}$ et $x_k = a_k$,
- rectangles à droite : $\sigma_n^p = \{a_0, \dots, a_n\}$ où $a_k = a + \frac{k(b-a)}{n}$ et $x_k = a_{k+1}$.

Autrement dit, on coupe l'intervalle $[a,b]$ en n segments de longueur égale, chacune à $\frac{b-a}{n}$, et on choisit comme points x_k les extrémités gauches, respectivement droites, de chacun de ces intervalles. On notera simplement S_n^g et S_n^d les sommes de Riemann correspondantes, i.e.

$$S_n^g = \frac{b-a}{n}\sum_{k=0}^{n-1} f\left(a + \frac{k}{n}(b-a)\right) \quad \text{et} \quad S_n^d = \frac{b-a}{n}\sum_{k=1}^{n} f\left(a + \frac{k}{n}(b-a)\right).$$

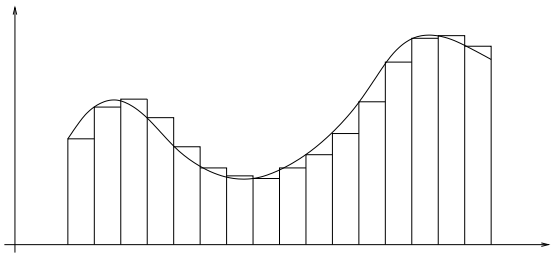

FIGURE 2.2 - Rectangles à gauche

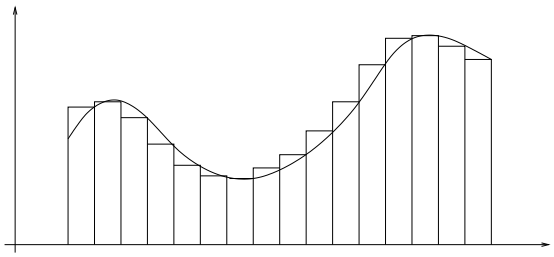

FIGURE 2.3 - Rectangles à droite

On sait que les deux suites $(S_n^g)_n$ et $(S_n^d)_n$ convergent vers $\int_a^b f(x)\mathrm{d}x$. La question est donc d'estimer l'erreur si on prend une de ces valeurs comme valeur approchée.

Proposition 1.2. *Soit $f : [a,b] \to \mathbb{R}$ de classe C^1 et $M_1 = \sup_{x\in[a,b]} |f'(x)|$. On a alors, pour tout n,*

$$\left|\int_a^b f(x)\mathrm{d}x - S_n^g\right| \leq \frac{M_1(b-a)^2}{2n} \quad \textit{et} \quad \left|\int_a^b f(x)\mathrm{d}x - S_n^d\right| \leq \frac{M_1(b-a)^2}{2n}. \tag{1.1}$$

Démonstration. On montre le résultat pour S_n^g, la preuve est la même pour S_n^d. On note $a_k = a + \frac{k}{n}(b-a)$. On peut commencer par noter que M_1 est bien un nombre fini. En effet f est C^1 donc f' est continue sur le segment $[a,b]$, elle y est donc bornée.

La relation de Chasles et l'inégalité triangulaire donnent

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

L'inégalité des accroissements finis permet alors d'écrire, pour tout k et pour tout $x \in [a_k, a_{k+1}]$,

$$|f(x) - f(a_k)| \leq M_1(x - a_k).$$

On a donc

$$\begin{aligned}\left|\int_a^b f(x)\mathrm{d}x - S_n^g\right| &\leq M_1\sum_{k=0}^{n-1}\int_{a_k}^{a_{k+1}} (x-a_k)\,\mathrm{d}x \\ &\leq \frac{M_1}{2}\sum_{k=0}^{n-1}\underbrace{(a_{k+1}-a_k)}_{=\frac{b-a}{n}}{}^2 \\ &\leq \frac{M_1(b-a)^2}{2n}.\end{aligned}$$

□

Remarque 1.1. *On peut montrer que la majoration de l'erreur est optimale dans le sens où il existe des fonctions f pour lesquelles on a égalité dans (1.1). Il faut pour cela que toutes les inégalités dans la preuve soient des égalités. C'est le cas si f est affine. Par exemple, si $f(x) = x$ sur $[0,1]$ on a*

$$S_n^g = \frac{n-1}{2n} \quad \textit{et} \quad S_n^d = \frac{n+1}{2n},$$

tandis que $\displaystyle\int_0^1 x\,\mathrm{d}x = \frac{1}{2}$. On trouve bien que

$$\left|\int_0^1 x\,\mathrm{d}x - S_n^g\right| = \left|\int_0^1 x\,\mathrm{d}x - S_n^d\right| = \frac{1}{2n} = \frac{M_1(b-a)^2}{2n}.$$

1.1.3 Méthode du point milieu

Dans les méthodes des rectangles à gauche ou à droite la convergence semble lente. Cependant l'erreur réelle peut a priori être beaucoup plus petite que celle donnée par (1.1). En effet, pour obtenir cette dernière on a estimé l'erreur commise sur chacun des sous-intervalles $[a_k, a_{k+1}]$ et on

a additionné ces erreurs. Cependant celles-ci pourraient se "compenser", certaines étant positives et d'autres négatives. Dans la méthode des rectangles à gauche, l'erreur commise est négative dans les intervalles où f est croissante (dans le sens où $\dfrac{b-a}{n} f(a_k) \leq \displaystyle\int_{a_k}^{a_{k+1}} f(x)\,\mathrm{d}x$) et elle est positive quand f est décroissante. Inversement, dans la méthode des rectangles à droite, l'erreur commise est positive dans les intervalles où f est croissante et elle est négative quand f est décroissante. En particulier, si la fonction f est croissante sur $[a,b]$ ou bien décroissante sur $[a,b]$ il n'y a pas de compensations d'erreurs entre les intervalles, ce qui explique que l'erreur (1.1) peut être atteinte.

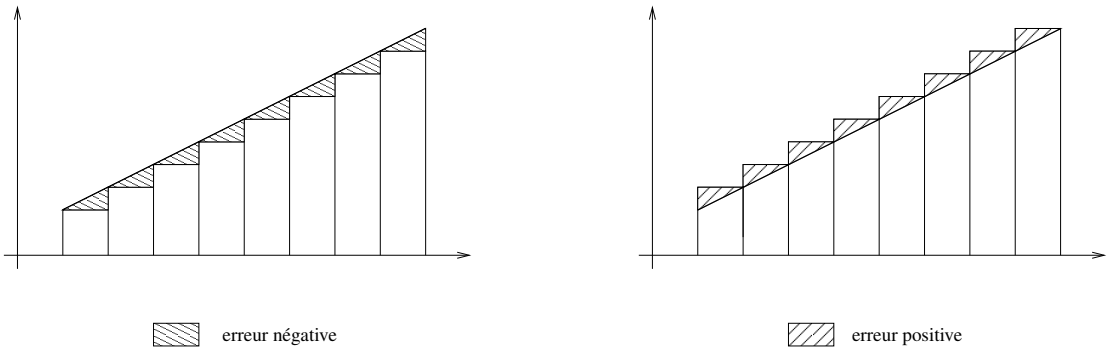

FIGURE 2.4 - Erreurs dans les méthodes des rectangles à gauche et à droite

La méthode du point milieu permet d'améliorer sensiblement la vitesse de convergence par rapport aux rectangles à gauche ou à droite. L'idée est ici de couper à nouveau $[a,b]$ en intervalles de même longeur, i.e. on prend toujours $a_k = a + \frac{k}{n}(b-a)$, mais on prend les points x_k au milieu des intervalles, i.e. $x_k = \frac{a_k + a_{k+1}}{2}$. Que sur un intervalle $[a_k, a_{k+1}]$ la fonction f soit croissante ou décroissante, il y aura une "compensation partielle" au sein même de l'intervalle.

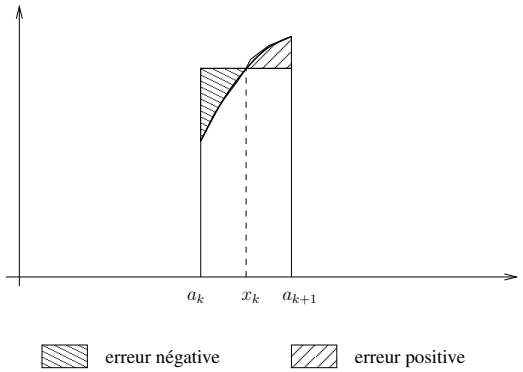

FIGURE 2.5 - Erreur dans la méthode du point milieu

On note $m_k = \dfrac{a_k + a_{k+1}}{2}$ le milieu de l'intervalle $[a_k, a_{k+1}]$ et

$$S_n^m = \frac{b-a}{n} \sum_{k=0}^{n-1} f\left(m_k\right),$$

la somme de Riemann associée. A nouveau le Théorème 1.1 assure que la suite $(S_n^m)_n$ converge vers $\int_a^b f(x)\,\mathrm{d}x$ et on souhaite estimer l'erreur $|\int_a^b f(x)\,\mathrm{d}x - S_n^m|$.

Proposition 1.3. *Soit $f : [a,b] \to \mathbb{R}$ de classe C^2 et $M_2 = \sup_{x\in[a,b]} |f''(x)|$. On a alors, pour tout n,*

$$\left|\int_a^b f(x)\mathrm{d}x - S_n^m\right| \leq \frac{M_2(b-a)^3}{24n^2}. \tag{1.2}$$

Remarque 1.2. *Si f est une fonction affine sa dérivée seconde est nulle et on a $M_2 = 0$. Cela montre que la méthode du point milieu est exacte pour des fonctions affines alors qu'on approche f par une fonction constante. C'est un autre reflet des compensations partielles d'erreurs au sein des différents intervalles.*

Démonstration. Le même calcul que dans la preuve de (1.1) donne

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

Pour tout k, on écrit la formule de Taylor-Lagrange entre x et m_k : il existe c_k entre x et m_k tel que

$$f(x) - f(m_k) = f'(m_k)(x-m_k) + \frac{f''(c_k)}{2}(x-m_k)^2,$$

et donc [a]

$$\int_a^b f(x)\,\mathrm{d}x - S_n^m = \sum_{k=0}^{n-1} \left(f'(m_k) \int_{a_k}^{a_{k+1}} (x-m_k)\,\mathrm{d}x + \int_{a_k}^{a_{k+1}} \frac{f''(c_k)}{2}(x-m_k)^2\,\mathrm{d}x \right).$$

Comme m_k est le milieu de $[a_k, a_{k+1}]$ on a

$$\int_{a_k}^{a_{k+1}} (x-m_k)\,\mathrm{d}x = \frac{1}{2}\left[(a_{k+1}-m_k)^2 - (a_k - m_k)^2\right] = 0,$$

d'où on tire

$$\begin{aligned}
\left|\int_a^b f(x)\,\mathrm{d}x - S_n^m\right| &= \left|\sum_{k=0}^{n-1} \int_{a_k}^{a_{k+1}} \frac{f''(c_k)}{2}(x-m_k)^2\,\mathrm{d}x\right| \\
&\leq \sum_{k=0}^{n-1} \frac{M_2}{2} \int_{a_k}^{a_{k+1}} (x-m_k)^2\,\mathrm{d}x \\
&\leq \frac{M_2(b-a)^3}{24n^2}.
\end{aligned}$$

□

a. **Attention !!** Bien que la notation ne l'indique pas, la valeur de c_k dépend a priori de x et on ne peut donc pas sortir $f''(c_k)$ de l'intégrale.

Remarque 1.3. *On pourrait croire que le gain dans l'erreur par rapport à la méthode des rectangles à gauche ou à droite provient de l'utilisation de la formule de Taylor à l'ordre* 2*. Ce n'est pas le cas. Si on utilise la même stratégie pour les rectangles à gauche par exemple, le terme d'ordre* 1 *donnera*

$$\int_{a_k}^{a_{k+1}} (x - a_k)\,\mathrm{d}x = \frac{1}{2}(a_{k+1} - a_k)^2 = \frac{(b-a)^2}{2n^2},$$

On aura ainsi

$$\int_a^b f(x)\,\mathrm{d}x - S_n^g = \sum_{k=0}^{n-1} f'(a_k)\frac{(b-a)^2}{2n^2} + \sum_{k=0}^{n-1}\int_{a_k}^{a_{k+1}} \frac{f''(c_k)}{2}(x-a_k)^2\,\mathrm{d}x.$$

Le second terme du membre de droite peut être borné par $\dfrac{(b-a)^3}{6n^2}$*, et est donc aussi de l'ordre de* $\dfrac{1}{n^2}$*. Cependant le premier terme est a priori borné lui par*

$$\left|\sum_{k=0}^{n-1} f'(a_k)\frac{(b-a)^2}{2n^2}\right| \leq \frac{M_1(b-a)^2}{2n}.$$

On retrouve précisément l'erreur obtenue dans (1.1).

Remarque 1.4. *Même si l'erreur est en* $O\left(\frac{1}{n^2}\right)$ *au lieu de* $O\left(\frac{1}{n}\right)$ *pour la méthode des rectangles à gauche ou à droite, cela ne signifie pas que pour un* n *donné l'erreur sera plus petite. D'une part on obtient ici une majoration de l'erreur (et l'erreur réelle peut être inférieure), d'autre part la constante* M_2 *peut être beaucoup plus grande que la constante* M_1 *rendant l'estimation meilleure uniquement lorsque* n *devient très grand.*

1.2 Méthode des trapèzes

Les méthodes précédentes reposent sur la notion de somme de Riemann, et donc sur l'idée d'approcher sur chacun des sous-intervalles $[a_k, a_{k+1}]$ la fonction f par une fonction constante (égale à la valeur de f au point choisi). Il est naturel d'essayer d'approcher f non plus par une constante mais par une fonction affine (ou même un polynôme d'ordre plus élevé, voir la Section 1.3 et le Chapitre 3).

À nouveau on ne considère que le cas où les intervalles $[a_k, a_{k+1}]$ sont de longueur égale, i.e. $a_k = \dfrac{k}{n}(b-a)$. On note f_k la fonction affine qui interpole f aux points $(a_k, f(a_k))$ et $(a_{k+1}, f(a_{k+1}))$. Autrement dit

$$f_k(x) = \frac{f(a_{k+1}) - f(a_k)}{a_{k+1} - a_k}(x - a_k) + f(a_k).$$

On va prendre f_k comme approximation de f sur l'intervalle $[a_k, a_{k+1}]$. La valeur approchée de l'intégrale de f sera donc

$$T_n := \sum_{k=0}^{n-1}\int_{a_k}^{a_{k+1}} f_k(x)\,\mathrm{d}x. \tag{1.3}$$

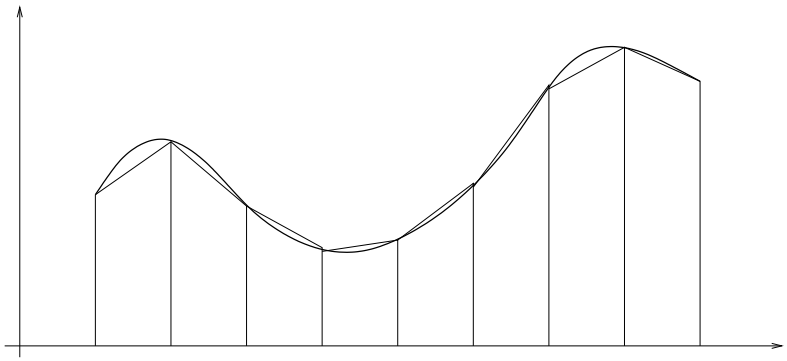

FIGURE 2.6 - Méthode des trapèzes

On calcule par ailleurs facilement (c'est l'aire d'un trapèze, voir la Figure 2.6)

$$\int_{a_k}^{a_{k+1}} f_k(x)\,\mathrm{d}x = \frac{1}{2}\left(f(a_{k+1}) + f(a_k)\right)\underbrace{(a_{k+1} - a_k)}_{=\frac{b-a}{n}}.$$

On obtient donc

$$T_n = \frac{b-a}{2n}\sum_{k=0}^{n-1}\left(f(a_{k+1}) + f(a_k)\right) = \frac{b-a}{2n}\left(f(a) + 2\sum_{k=1}^{n-1} f(a_k) + f(b)\right).$$

On a alors la majoration suivante de l'erreur commise.

Proposition 1.4. *Soit $f : [a,b] \to \mathbb{R}$ de classe C^2 et $M_2 = \sup_{x\in[a,b]} |f''(x)|$. On a, pour tout n,*

$$\left|\int_a^b f(x)\mathrm{d}x - T_n\right| \leq \frac{M_2(b-a)^3}{12n^2}. \tag{1.4}$$

Démonstration. Le début de la preuve est identique aux précédentes. A l'aide de la relation de Chasles et de l'inégalité triangulaire on écrit

$$\left|\int_a^b f(x)\mathrm{d}x - T_n\right| \leq \sum_{k=0}^{n-1}\left|\int_{a_k}^{a_{k+1}} f(x) - f_k(x)\,\mathrm{d}x\right|.$$

L'idée est ensuite d'estimer chacun des termes $\displaystyle\int_{a_k}^{a_{k+1}} g_k(x)\,\mathrm{d}x$ où $g_k(x) = f(x) - f_k(x)$ en effectuant d'abord une intégration par parties. On commence par noter que par construction $g(a_k) = g(a_{k+1}) = 0$, il n'y a donc pas de terme de bord dans l'intégration par parties. On écrit ainsi

$$\int_{a_k}^{a_{k+1}} g_k(x)\,\mathrm{d}x = -\int_{a_k}^{a_{k+1}} g_k'(x)(x-c)\,\mathrm{d}x,$$

où c est une constante que l'on précisera ensuite. L'idée est de faire une seconde intégration par parties toujours sans générer de termes de bords. Les primitives de $v(x) = x - c$ sont de la forme $\frac{1}{2}(x-c)^2 + d$ et on a donc

$$\begin{aligned}\int_{a_k}^{a_{k+1}} g_k(x)\,\mathrm{d}x &= -\int_{a_k}^{a_{k+1}} g_k'(x)(x-c)\,\mathrm{d}x \\ &= \int_{a_k}^{a_{k+1}} g_k''(x)\left(\frac{1}{2}(x-c)^2+d\right)\mathrm{d}x - \left[g_k'(x)\left(\frac{1}{2}(x-c)^2+d\right)\right]_{a_k}^{a_{k+1}}.\end{aligned}$$

Pour que le nouveau terme de bord s'annule il suffit que le polynôme $\frac{1}{2}(x-c)^2 + d$ s'annule en a_k et en a_{k+1}, autrement dit on choisit c et d pour que

$$\frac{1}{2}(x-c)^2 + d = \frac{1}{2}(x-a_k)(x-a_{k+1}).$$

On a alors

$$\begin{aligned}\left|\int_{a_k}^{a_{k+1}} g_k(x)\,\mathrm{d}x\right| &\leq \frac{1}{2}\int_{a_k}^{a_{k+1}} |g_k''(x)(x-a_k)(x-a_{k+1})|\,\mathrm{d}x \\ &\leq \frac{M_2}{2}\int_{a_k}^{a_{k+1}} (x-a_k)(a_{k+1}-x)\,\mathrm{d}x \\ &= \frac{M_2(a_{k+1}-a_k)^3}{12} = \frac{M_2(b-a)^3}{12n^3},\end{aligned}$$

où on a utilisé à la deuxième ligne le fait que f_k est affine et donc $|g_k''(x)| = |f''(x)| \leq M_2$. Finalement on obtient

$$\left|\int_a^b f(x)\mathrm{d}x - T_n\right| \leq \sum_{k=0}^{n-1}\left|\int_{a_k}^{a_{k+1}} g_k(x)\,\mathrm{d}x\right| \leq \frac{M_2(b-a)^3}{12n^2}.$$

□

Remarque 1.5. *On peut facilement voir que l'on a en fait $T_n = \frac{S_n^g + S_n^d}{2}$. Alors que les deux méthodes des rectangles à gauche et à droite conduisent à une erreur a priori en $O\left(\frac{1}{n}\right)$, la méthode des trapèzes donne une erreur en $O\left(\frac{1}{n^2}\right)$. Comme on l'a signalé dans la Section 1.1.3, dans la méthode des rectangles à gauche l'erreur est négative, dans le sens où on sous-estime l'intégrale, lorsque f est croissante tandis que la méthode des rectangles à droite surestime l'intégrale, et vice-versa lorsque f est décroissante. Quand on prend la moyenne entre S_n^g et S_n^d on peut donc raisonnablement s'attendre à une compensation partielle entre les deux erreurs. C'est exactement ce que montre (1.4).*

1.3 Méthode de Simpson

On termine ce chapitre sur l'intégration numérique avec la méthode de Simpson. Dans cette dernière, l'idée est à nouveau d'approcher la fonction f sur chaque intervalle $[a_k, a_{k+1}]$, $a_k = \frac{k}{n}(b-a)$,

par une fonction "simple". On a utilisé une fonction constante, ou polynôme de degré au plus 0, dans la Section 1.1 et une fonction affine, ou polynôme de degré au plus 1, dans la Section 1.2. Il est naturel de choisir un polynôme de degré au plus 2.

Étant donné un intervalle $[c,d]$ on considère le polynôme $P \in \mathbb{R}_2[X]$ qui interpole f aux points d'abscisse c, d et $m = \frac{c+d}{2}$, i.e. $P(X) = \alpha + \beta X + \gamma X^2$ est tel que $P(c) = f(c)$, $P(m) = f(m)$ et $P(d) = f(d)$.

Lemme 1.5. *Il existe un unique $P \in \mathbb{R}_2[X]$ qui interpole f aux points d'abscisse c, d et m. Il est donné par*

$$P(X) = f(c)\frac{(X-m)(X-d)}{(c-m)(c-d)} + f(m)\frac{(X-c)(X-d)}{(m-c)(m-d)} + f(d)\frac{(X-c)(X-m)}{(d-m)(d-c)}. \tag{1.5}$$

Cette formule est un cas particulier de polynôme d'interpolation de Lagrange. Nous y reviendrons dans la Section 3.1. Un calcul simple montre que si P est donné par (1.5) on a

$$\int_c^d P(x)\mathrm{d}x = \frac{d-c}{6}\left(f(c) + 4f(m) + f(d)\right). \tag{1.6}$$

Tout comme dans les sections précédentes, sur chaque intervalle $[a_k, a_{k+1}]$ on va approcher f par le polynôme d'interpolation correspondant et l'intégrale de f par celle du polynôme. La méthode de Simpson consiste donc à approcher $\int_a^b f(x)\mathrm{d}x$ par (pour chaque k on a $a_{k+1} - a_k = \frac{b-a}{n}$)

$$S_n := \frac{b-a}{6n}\sum_{k=0}^{n-1}\left(f(a_k) + 4f(m_k) + f(b_k)\right).$$

L'estimation de l'erreur est basée sur le lemme suivant

Lemme 1.6. *Soit $f : [c,d] \to \mathbb{R}$ une fonction de classe C^4 et $M_4 = \sup_{x\in[c,d]} |f^{(4)}(x)|$, alors*

$$\left|\int_c^d f(x)\mathrm{d}x - \frac{d-c}{6}\left(f(c) + 4f\left(\frac{c+d}{2}\right) + f(d)\right)\right| \leq \frac{M_4(d-c)^5}{2880}. \tag{1.7}$$

Remarque 1.6. *On peut en fait montrer qu'il existe $\xi \in]c,d[$ tel que*

$$\int_c^d f(x)\mathrm{d}x - \frac{d-c}{6}\left(f(c) + 4f\left(\frac{c+d}{2}\right) + f(d)\right) = \frac{(d-c)^5}{2880}f^{(4)}(\xi).$$

Le Lemme précédent en découle alors directement.

En appliquant le lemme sur chacun des intervalles $[a_k, a_{k+1}]$ on obtient immédiatement

Proposition 1.7. *Soit $f : [a,b] \to \mathbb{R}$ de classe C^4 et $M_4 = \sup\limits_{x\in[a,b]} |f^{(4)}(x)|$. On a, pour tout n,*

$$\left|\int_a^b f(x)\mathrm{d}x - S_n\right| \leq \frac{M_4(b-a)^5}{2880n^4}. \tag{1.8}$$

Démonstration du Lemme. On définit la fonction g sur l'intervalle $[0, \frac{d-c}{2}]$ par

$$g(t) = \int_{m-t}^{m+t} f(x)\mathrm{d}x - \frac{t}{3}\left(f(m-t) + 4f(m) + f(m+t)\right),$$

où $m = \dfrac{c+d}{2}$. La fonction g correspond au membre de gauche de (1.7) si on remplace c par $m-t$ et d par $m+t$. On veut donc montrer que

$$\left| g\left(\frac{d-c}{2}\right)\right| \leq \frac{M_4(d-c)^5}{2880}. \tag{1.9}$$

Comme f est C^4 la fonction g l'est aussi, et on calcule les dérivées successives de g :

$$\begin{aligned}
g'(t) &= \frac{1}{3}\left(2f(m+t) + 2f(m-t) - 4f(m) + tf'(m-t) - tf'(m+t)\right),\\
g''(t) &= \frac{1}{3}\left(f'(m+t) - f'(m-t) - tf''(m-t) - tf''(m+t)\right),\\
g^{(3)}(t) &= \frac{t}{3}\left(f^{(3)}(m-t) - f^{(3)}(m+t)\right).
\end{aligned}$$

On peut en particulier remarquer que $g(0) = g'(0) = g''(0) = 0$. Par ailleurs, le Théorème des accroissements finis appliqué à la fonction $f^{(3)}$ donne l'existence de $\xi \in]m-t, m+t[$ tel que

$$g^{(3)}(t) = \frac{t}{3}\left(f^{(3)}(m-t) - f^{(3)}(m+t)\right) = -\frac{2t^2}{3}f^{(4)}(\xi).$$

On en déduit que pour tout $t \in [0, \frac{d-c}{2}]$ on a

$$|g^{(3)}(t)| \leq \frac{2M_4}{3}t^2.$$

On effectue ensuite des intégrations successives. On obtient

$$|g''(t)| = |g''(t) - g''(0)| \leq \int_0^t |g^{(3)}(u)|\mathrm{d}u \leq \frac{2M_4}{3}\int_0^t u^2\mathrm{d}u = \frac{2M_4}{9}t^3.$$

Et de même

$$|g'(t)| \leq \frac{M_4}{18}t^4, \qquad |g(t)| \leq \frac{M_4}{90}t^5.$$

En prenant $t = \frac{d-c}{2}$ on obtient bien (1.9). □

Le Lemme ci-dessus montre que, de façon a priori inattendue, la formule d'approximation de Simpson est exacte non seulement pour les polynômes de degré au plus 2 mais en fait aussi pour les polynômes de degré 3. En effet, la dérivée quatrième d'un polynôme de degré 3 est nulle et la constante M_4 est donc nulle aussi. Ce fait est à rapprocher de ce qui se passe dans la méthode du point milieu qui est exacte pour des fonctions affines alors qu'on approche f par une fonction constante par morceaux. Si f est une fonction polynôme de degré (au plus) 3 et P le polynôme

d'interpolation de degré (au plus) 2 alors $f - P$ est un polynôme de degré au plus 3 qui s'annule en c, d et $m = \frac{c+d}{2}$. On en déduit qu'il existe $\alpha \in \mathbb{R}$ tel que $f(x) - P(x) = \alpha(x-c)(x-m)(x-d) = Q(x)$. On peut vérifier que $Q(c+d-x) = -Q(x)$ et donc, en effectuant le changement de variable $x = c + d - t$,

$$\int_c^d Q(t)\mathrm{d}t = -\int_c^d Q(c+d-t)\mathrm{d}t = -\int_c^d Q(x)\,\mathrm{d}x,$$

ce qui prouve que $\displaystyle\int_c^d Q(t)\,\mathrm{d}t = 0$ et donc $\displaystyle\int_c^d f(t)\mathrm{d}t = \int_c^d P(t)\mathrm{d}t$. (C'est en fait un cas particulier de la Proposition 5.6.)

Chapitre 2

Résolution approchée de $f(x) = 0$

2.1 Introduction

Dans de nombreux théorèmes d'analyse on trouve des résultats tels que "il existe un réel x tel que $f(x) = 0$" où f est une fonction donnée. On pourra penser au Théorème des Valeurs Intermédiaires (si f est continue), au Théorème de Rolle (ici ce sera plutôt "$f'(x) = 0$") ou encore au théorème du point fixe (on résout ici $g(x) = x$ ce qui équivaut à $f(x) = g(x) - x = 0$, voir TD). On dit que x est un zéro ou une racine de f et la question naturelle est alors "que vaut ce x ?". Mis à part quelques cas particuliers simples on ne pourra pas résoudre l'équation, autrement dit trouver explicitement la valeur de x. Même dans le cas où f est bijective et où on peut alors écrire $x = f^{-1}(0)$, il est souvent difficile voir impossible d'avoir une expression simple pour f^{-1} et dire que $x = f^{-1}(0)$ ne dit pas grand chose sur la valeur de x. On est donc amené à trouver une valeur *approchée* de cette dernière. Dans ce chapitre on se restreint au cas où f est une fonction d'une variable réelle.

Dans tout le chapitre on supposera que f est une fonction continue définie sur un intervalle I et à valeurs dans $\mathbb{R}$.

2.2 Méthode de la dichotomie

La méthode de la dichotomie consiste à déterminer une valeur approchée d'une solution de l'équation $f(x) = 0$ en "coupant en deux" l'intervalle dans lequel on cherche cette dernière. On commence par déterminer un intervalle $[a, b] \subset I$ sur lequel la fonction f change de signe, c'est-à-dire tel que $f(a)f(b) < 0$. Comme f est continue, le théorème des valeurs intermédiaires permet d'affirmer qu'il y a au moins une racine entre a et b. On coupe l'intervalle en deux et on calcule la valeur de f en $c = \dfrac{a+b}{2}$:

- Si $f(c) = 0$ c'est terminé on a trouvé une racine.
- Sinon, soit $f(c)$ est du même signe que $f(a)$ et on remplace l'intervalle $[a, b]$ par $[c, b]$ (on a alors $f(b)f(c) < 0$) soit il est du signe de $f(b)$ et on remplace cette fois $[a, b]$ par $[a, c]$ (on a maintenant $f(a)f(c) < 0$).

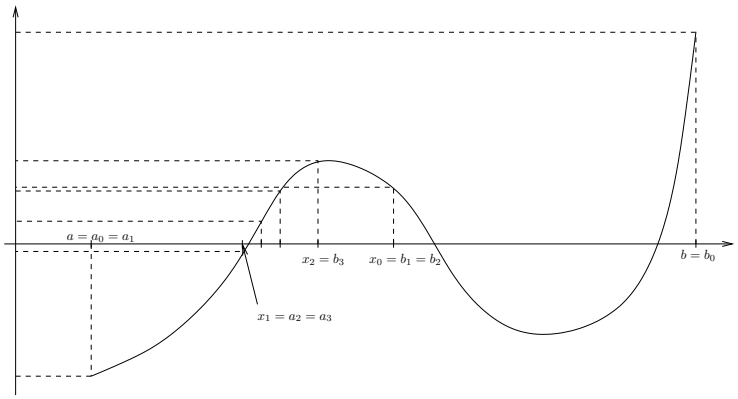

FIGURE 2.1 – La méthode de la dichotomie

Le théorème des valeurs intermédiaires affirme cette fois qu'il y a une solution dans le nouvel intervalle et on itère la procédure.

Remarque 2.1. *C'est cette méthode qui a été utilisée en L1 pour démontrer le théorème des valeurs intermédiaires.*

On définit ainsi par récurrence les suites $(a_n)_n$, $(b_n)_n$ et $(x_n)_n$:

- On prend $a_0 = a$, $b_0 = b$ et $x_0 = \frac{a+b}{2}$.
- Pour tout $n \in \mathbb{N}$, on pose $a_{n+1} = a_n$, $b_{n+1} = x_n$ si $f(a_n)f(x_n) \leq 0$ et $a_{n+1} = x_n$, $b_{n+1} = b_n$ sinon, puis $x_{n+1} = \frac{a_{n+1}+b_{n+1}}{2}$.

On peut montrer que les suites $(a_n)_n$ et $(b_n)_n$ sont adjacentes (voir cours de L1) donc convergent vers un certain ξ. Par ailleurs $a_n \leq x_n \leq b_n$ donc $(x_n)_n$ converge également vers ξ. Comme $f(a_n)f(b_n) \leq 0$ pour tout n on obtient, en passant à la limite (f est continue) que $f(\xi)^2 \leq 0$ et donc $f(\xi) = 0$.

On note également que pour tout n on a $b_n - a_n = \frac{b-a}{2^n}$ (vous pouvez le vérifier par récurrence, l'idée étant qu'on divise l'intervalle en 2 à chaque étape). Par ailleurs $x_n, \xi \in [a_n, b_n]$ pour tout n donc $|x_n - \xi| \leq b_n - a_n$.

Conclusion : pour tout $n \in \mathbb{N}$ on a $|x_n - \xi| \leq \dfrac{b-a}{2^n}$. Si ϵ est la précision désirée, on s'arrête dès qu'on a trouvé exactement la racine ou plus généralement dès que $\dfrac{b-a}{2^n} \leq \epsilon$, i.e.

$$n \geq \log_2\left(\frac{b-a}{\epsilon}\right) = \frac{\ln\left(\frac{b-a}{\epsilon}\right)}{\ln(2)}.$$

Le nombre d'étapes nécessaire pour atteindre une précision ϵ donnée est donc de l'ordre de $|\ln(\epsilon)|$. On parle ici de convergence linéaire, ou d'ordre 1 (voir la Définition 2.1).

Remarque 2.2. *Essentiellement le nombre d'étapes de calcul nécessaire pour avoir une précision ϵ est proportionnel à $-\ln(\epsilon)$. Si on souhaite une précision de k chiffres après la virgule on a alors $\epsilon = 10^{-k}$, et on constate que le nombre d'étapes de calcul est proportionnel au nombre de chiffres souhaité, d'où le terme linéaire.*

De façon plus précise, l'ordre de convergence d'une suite est défini de la façon suivante.

Définition 2.1. *Soit $(x_n)_n$ une suite de réels qui converge vers $\xi \in \mathbb{R}$, et $e_n := |x_n - \xi|$. On dit que la convergence est d'ordre $p \geq 1$ s'il existe des constantes $0 < c_- \leq c_+$ telles que à partir d'un certain rang*

$$c_- e_n^p \leq e_{n+1} \leq c_+ e_n^p.$$

On parle de convergence linéaire si $p = 1$, quadratique si $p = 2$.

Remarque 2.3. *Une condition suffisante pour que la convergence soit d'ordre p est que $\lim_{n\to\infty} \frac{e_{n+1}}{e_n^p} = c \neq 0$. L'idée étant que plus p est grand, mieux c'est. Prenons par exemple $c_+ = 1$. On montre alors facilement par récurrence que pour n assez grand on a alors, pour tout $m \in \mathbb{N}$,*

$$e_{n+m} \leq e_n^{p^m}.$$

Si x_n est une valeur approchée de ξ à 10^{-k}, i.e. $e_n \leq 10^{-k}$, alors x_{n+1} est une valeur approchée à 10^{-pk}. Autrement dit, à chaque étape le nombre de décimales exactes est multiplié par p.

Lorsque $p = 1$, le nombre de décimales exactes progresse en fait au plus arithmétiquement. En écrivant $c_\pm = 10^{-\gamma_\pm}$ le nombre de décimales gagnées à chaque étape est au moins γ_+ et au plus γ_-. On peut cependant avoir $\gamma_+ < 0$ comme dans la dichotomie : on ne gagne pas forcément une décimale à chaque étape (on ne fait "que" diviser la précision par 2).

2.3 Fonctions convexes

La méthode de dichotomie a l'avantage de ne pas nécessiter beaucoup d'hypothèses (la continuité de la fonction f suffit). Pour les autres méthodes que l'on va voir, on demandera un peu plus à la fonction f, essentiellement pour s'assurer que l'algorithme d'approximation converge, i.e. si $(x_n)_n$ est la suite des valeurs approchées obtenues on veut s'assurer que $x_n \to \xi$. Une des hypothèses concerne la notion de fonction convexe que vous avez vue en L2. On rappelle juste ici la définition ainsi que les quelques propriétés que l'on utilisera par la suite.

Définition 2.2. *Soit $f : I \to \mathbb{R}$. On dit que f est* convexe *(sur I) si*

$$\forall (x, y) \in I^2,\ \forall \lambda \in [0, 1], \quad f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y).$$

On dit que f est concave *si $-f$ est convexe.*

La fonction f est dite strictement convexe si

$$\forall x \neq y \in I,\ \forall \lambda \in]0, 1[, \quad f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda) f(y).$$

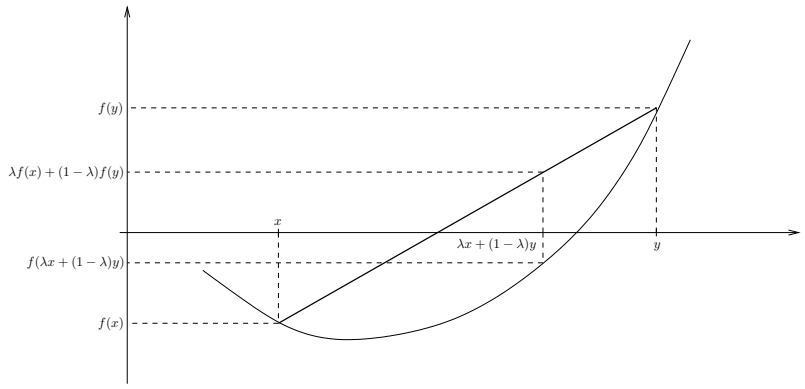

FIGURE 1.2 - Fonction convexe

Remarque 2.4. *a) Dans la définition ci-dessus il suffit de prendre $x < y$ (quitte à intervertir leur rôle) et $\lambda \in]0, 1[$.*

b) Géométriquement, f est convexe si le graphe de f est en dessous de toutes les cordes qui joignent deux points de ce graphe, et f est concave si le graphe de f est au dessus de toutes les cordes, voir la Figure 1.2.

Le théorème suivant permet de vérifier facilement si une fonction est convexe

Théorème 2.3. (Caractérisation des fonctions convexes dérivables)

1. *Si f est dérivable, alors f est convexe, resp. concave, si et seulement f' est croissante, resp. décroissante.*
2. *Si f est deux fois dérivable, alors f est convexe, resp. concave, si et seulement si $f'' \geq 0$, resp. $f'' \leq 0$.*

Lemme 2.4. (Inégalité des 3 pentes) *Soit $f : I \to \mathbb{R}$. La fonction f est convexe si et seulement si pour tous $x < y < z$ dans I on a*

$$\frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(x)}{z - x} \leq \frac{f(z) - f(y)}{z - y}. \tag{2.1}$$

Si on note A, B et C les points du graphe de f d'abscisse respective x, y et z, l'inégalité (2.1) dit que la pente de (AB) est inférieure à celle de (AC) qui est inférieure à celle de (BC).

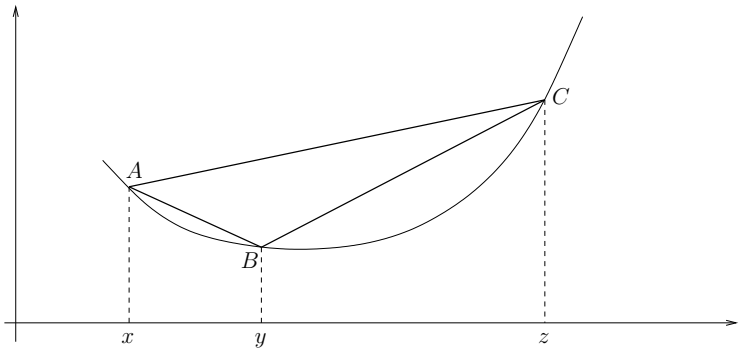

FIGURE 1.3 - Inégalité des trois pentes

Démonstration. Supposons que f est convexe. Soient $x < y < z$, on peut alors écrire $y = \lambda x + (1-\lambda)z$ avec $\lambda = \frac{z-y}{z-x} \in [0,1]$, et on a donc

$$f(y) = f(\lambda x + (1-\lambda)z) \leq \lambda f(x) + (1-\lambda)f(z) = \frac{z-y}{z-x}f(x) + \frac{y-x}{z-x}f(z),$$

d'où on déduit les deux inégalités de (2.1).

Réciproquement, supposons que (2.1) est vraie pour tous $x < y < z$ dans I. Soient $x < y$ et $\lambda \in]0,1[$. On a $x < \lambda x + (1-\lambda)y < y$, et donc

$$\begin{aligned}
& \frac{f(\lambda x + (1-\lambda)y) - f(x)}{\lambda x + (1-\lambda)y - x} \leq \frac{f(y) - f(x)}{y - x} \\
\iff \quad & \frac{f(\lambda x + (1-\lambda)y) - f(x)}{(1-\lambda)(y-x)} \leq \frac{f(y) - f(x)}{y - x} \\
\iff \quad & f(\lambda x + (1-\lambda)y) - f(x) \leq (1-\lambda)(f(y) - f(x)) \\
\iff \quad & f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y),
\end{aligned}$$

ce qui prouve que f est convexe. □

Démonstration du Théorème. On montre le 1., le 2. découle ensuite de l'équivalence entre f' croissante et $f'' \geq 0$.

Supposons que f est convexe. Soient $x < z$ dans I, alors pour tout $y \in]x,z[$ on a (2.1). En faisant tendre y vers x la première inégalité donne

$$f'(x) \leq \frac{f(z) - f(x)}{z - x},$$

et en faisant tendre y vers z la seconde inégalité donne

$$\frac{f(z) - f(x)}{z - x} \leq f'(z),$$

d'où on déduit que $f'(x) \leq f'(z)$, ce qui prouve bien que f' est croissante.

Réciproquement supposons que f' est croissante. Soient $x < y \in I$, $\lambda \in]0,1[$ et $z = \lambda x + (1-\lambda)y$. On a donc $x < z < y$. On applique ensuite le théorème des accroissements finis sur $[x,z]$ et sur $[z,y]$: il existe $x_1 \in]x,z[$ et $x_2 \in]z,y[$ tels que

$$f'(x_1) = \frac{f(z) - f(x)}{z - x} = \frac{f(z) - f(x)}{(1-\lambda)(y-x)} \quad \text{et} \quad f'(x_2) = \frac{f(y) - f(z)}{y - z} = \frac{f(y) - f(z)}{\lambda(y-x)}.$$

Comme f' est croissante on a $f'(x_1) \leq f'(x_2)$, d'où on déduit

$$\frac{f(z) - f(x)}{(1-\lambda)(y-x)} \leq \frac{f(y) - f(z)}{\lambda(y-x)} \quad \iff \quad f(z) \leq \lambda f(x) + (1-\lambda)f(y),$$

ce qui prouve que f est convexe. □

Proposition 2.5. *Soit $f : I \to \mathbb{R}$ une fonction convexe dérivable, alors pour tout $x_0 \in I$ le graphe de f est au dessus de la tangente au point d'abscisse x_0 : pour tout $x \in I$*

$$f(x) \geq f(x_0) + f'(x_0)(x - x_0).$$

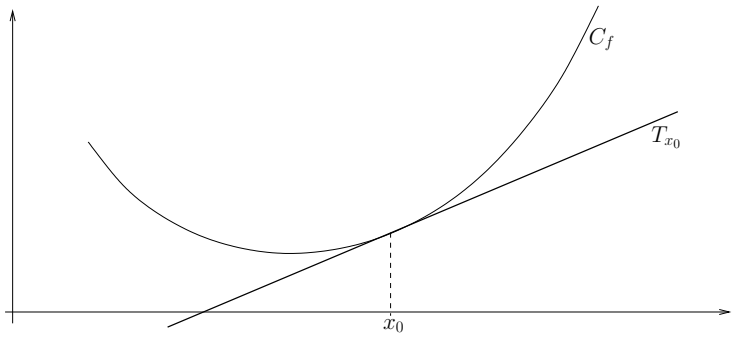

FIGURE 1.4 - Tangente au graphe d'une fonction convexe

Démonstration. Si $x < x_0$, d'après le Théorème des accroissement finis il existe $y \in]x, x_0[$ tel que

$$\frac{f(x_0) - f(x)}{x_0 - x} = f'(y).$$

Comme f est convexe, f' est croissante et donc

$$\frac{f(x_0) - f(x)}{x_0 - x} \leq f'(x_0) \quad \Longleftrightarrow \quad f(x) \geq f(x_0) + f'(x_0)(x - x_0).$$

De même, si $x > x_0$, il existe $y \in]x_0, x[$ tel que

$$\frac{f(x) - f(x_0)}{x - x_0} = f'(y).$$

La croissance de f' entraine que

$$f'(x_0) \leq \frac{f(x) - f(x_0)}{x - x_0} \quad \Longleftrightarrow \quad f(x) \geq f(x_0) + f'(x_0)(x - x_0).$$

□

2.4 Méthode de la sécante

2.4.1 Interpolation linéaire

À nouveau, on commence par se placer sur un intervalle $[a, b] \subset I$ sur lequel la fonction f change de signe, c'est-à-dire tel que $f(a)f(b) < 0$. L'idée est de remplacer la fonction f par une fonction

affine L et d'utiliser le zéro $\bar{\xi}$ de L comme valeur approchée de celui ξ de f. On interpole ici f en a et b, c'est-à-dire que L est la fonction affine telle que $L(a) = f(a)$ et $L(b) = f(b)$. Autrement dit on remplace le graphe de f par la corde passant par les points $A = (a, f(a))$ et $B = (b, f(b))$. Cette méthode est appelée *méthode de la fausse position* ou *regula falsi*.

On détermine facilement la fonction L :

$$L(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a),$$

ainsi que son unique zéro

$$\bar{\xi} = \frac{af(b) - bf(a)}{f(b) - f(a)}. \tag{2.2}$$

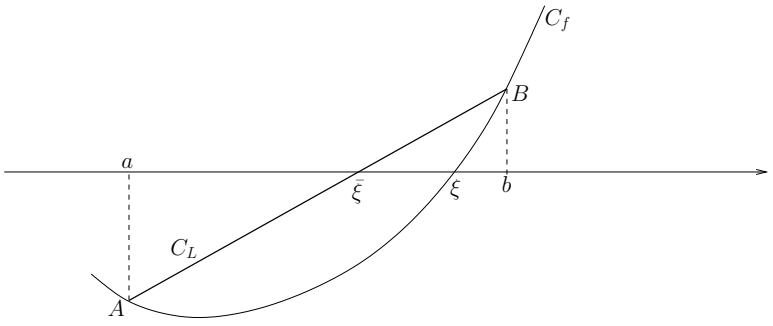

FIGURE 1.5 - Méthode de la fausse position

Si on suppose que de plus la fonction f est de classe C^2 on peut alors estimer l'erreur commise.

Lemme 2.6. *Pour tout $x \in]a, b[$ il existe $u \in]a, b[$ tel que*

$$f(x) - L(x) = \frac{1}{2}f''(u)(x - a)(x - b).$$

Démonstration. L'idée est similaire à la preuve de la formule de Taylor-Lagrange. Soit $x \in]a, b[$ fixé et $A \in \mathbb{R}$ tel que la fonction

$$g(y) = f(y) - L(y) - A(y - a)(y - b)$$

s'annule en x (un tel choix est possible puisque $(x-a)(x-b) \neq 0$). La fonction g s'annule en a, x et b donc d'après le Théorème de Rolle il existe $x_0 \in]a, x[$ et $x_1 \in]x, b[$ tels que $g'(x_0) = g'(x_1) = 0$, et en appliquant à nouveau le Théorème de Rolle il existe $u \in]x_0, x_1[\subset]a, b[$ tel que $g''(u) = 0$. Par ailleurs on calcule (L est affine donc $L'' = 0$)

$$g''(y) = f''(y) - 2A,$$

d'où $A = \frac{f''(u)}{2}$ et ainsi, puisque $g(x) = 0$,

$$0 = f(x) - L(x) - \frac{1}{2}f''(u)(x-a)(x-b).$$

□

Si par ailleurs f' ne s'annule pas (f est alors strictement monotone et a donc un unique zéro) on en déduit le résultat suivant

Proposition 2.7. *Soit ξ l'unique zéro de f et $\bar{\xi}$ celui de L. Il existe $u, v \in]a, b[$ tels que*

$$\bar{\xi} - \xi = \frac{1}{2}\frac{f''(u)}{f'(v)}(\bar{\xi} - a)(\bar{\xi} - b).$$

Si on note $m = \min_{x\in[a,b]} |f'(x)|$ et $M = \max_{x\in[a,b]} |f''(x)|$ alors

$$|\bar{\xi} - \xi| \leq \frac{M}{2m}|\bar{\xi} - a||\bar{\xi} - b|.$$

Démonstration. D'après le Lemme précédent appliqué en $\bar{\xi}$, il existe u tel que

$$f(\bar{\xi}) = \frac{1}{2}f''(u)(\bar{\xi} - a)(\bar{\xi} - b). \tag{2.3}$$

Par ailleurs, le Théorème des accroissements finis appliqué à f assure l'existence d'un $v \in]a, b[$ tel que

$$f(\bar{\xi}) = f(\bar{\xi}) - f(\xi) = f'(v)(\bar{\xi} - \xi).$$

En remplaçant cette identité dans (2.3) on obtient le résultat. □

Tout comme pour la méthode de la dichotomie, l'idée est ensuite d'itérer le processus (voir la section suivante).

Remarque 2.5. *De façon intuitive on peut voir cette approximation comme une amélioration de la dichotomie. Si $f(a)$ est plus proche de 0 que $f(b)$ on peut s'attendre à trouver le zéro de f plus proche de a que de b. C'est exactement ce que l'on fait en prenant pour approximation le zéro de L :*

$$\frac{\bar{\xi} - a}{\bar{\xi} - b} = \frac{f(a)}{f(b)}.$$

On verra que la méthode itérative correspondante (méthode de la sécante) n'est en fait pas bien meilleure que celle de la dichotomie.

2.4.2 Méthode de la sécante

On cherche maintenant à itérer la méthode de la fausse position. On va se placer dans le cas où la fonction f est strictement monotone sur $[a, b]$, ou plus précisément on suppose que f' ne s'annule pas qui est l'hypothèse utilisée dans la Proposition 2.7 pour avoir une estimation de l'erreur. On va de plus supposer que la fonction f'' ne s'annule pas non plus. Puisque f est C^2 cela signifie que $f'' > 0$ ou que $f'' < 0$ et donc que f est soit strictement convexe soit strictement concave sur $[a, b]$. Cette hypothèse permettra de justifier facilement que l'algorithme va converger.

Plus précisément, on supposera par la suite que $f' > 0$ (et donc que $f(a) < 0$ et $f(b) > 0$) et que $f'' > 0$. On peut toujours se ramener à ce cas quitte à remplacer $f(x)$ par $\pm f(\pm x)$:

- Si $f' > 0$ et $f'' > 0$ on garde f.
- Si $f' > 0$ et $f'' < 0$ alors la fonction $g(x) = -f(-x)$ vérifie $g' > 0$ et $g'' > 0$.
- Si $f' < 0$ et $f'' > 0$ alors la fonction $g(x) = f(-x)$ vérifie $g' > 0$ et $g'' > 0$.
- Si $f' < 0$ et $f'' < 0$ alors la fonction $g(x) = -f(x)$ vérifie $g' > 0$ et $g'' > 0$.

Le processus d'itération est alors décrit par, voir (2.2),

$$x_0 = a, \qquad x_{n+1} = \frac{x_n f(b) - b f(x_n)}{f(b) - f(x_n)}, \quad \forall n \in \mathbb{N}. \tag{2.4}$$

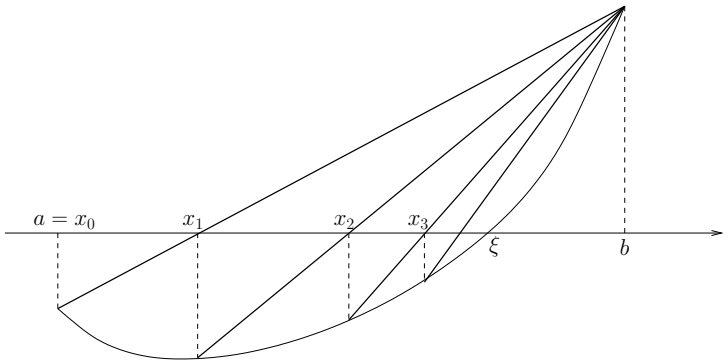

FIGURE 1.6 - Méthode de la sécante

Lemme 2.8. *Soit ξ l'unique zéro de f sur $[a, b]$. La suite $(x_n)_n$ définie par (2.4) est croissante et majorée par ξ.*

Démonstration. On montre par récurrence que pour tout n la propriété P_n : "$x_n < x_{n+1} < \xi$" est vraie.

- x_1 est l'unique zéro de la fonction affine L passant par $(a, f(a))$ et $(b, f(b))$. Comme $f(a) < 0$ et $f(b) > 0$ on a donc $a < x_1 < b$ et en particulier $x_0 < x_1$. Par ailleurs, f est strictement convexe donc le graphe de f se trouve en dessous de celui de L : $f(x_1) < L(x_1) = 0 = f(\xi)$. Et comme f est strictement croissante on en déduit que $x_1 < \xi$. Donc P_0 est vraie.

- Soit $n \in \mathbb{N}$, supposons que P_n est vraie. En particulier $f(x_{n+1}) < 0$ (on a $x_{n+1} < \xi$ et f est strictement croissante). On raisonne ensuite comme ci-dessus. x_{n+2} est l'unique zéro de la fonction affine L passant par $(x_{n+1}, f(x_{n+1}))$ et $(b, f(b))$. Comme $f(x_{n+1}) < 0$ et $f(b) > 0$ on a donc $x_{n+1} < x_{n+2} < b$. Par ailleurs, f est strictement convexe donc le graphe de f se trouve en dessous de celui de L : $f(x_{n+2}) < L(x_{n+2}) = 0$. Et comme f est strictement croissante on en déduit que $x_{n+2} < \xi$. Donc P_{n+1} est vraie.

□

Corollaire 2.9. *La suite $(x_n)_n$ converge vers ξ.*

Démonstration. La suite $(x_n)_n$ est croissante et majorée par ξ donc elle converge et sa limite ℓ vérifie $a < \ell \leq \xi < b$. Par ailleurs, en passant à la limite dans (2.4) on a

$$\ell = \frac{\ell f(b) - b f(\ell)}{f(b) - f(\ell)} \quad \Longleftrightarrow \quad (b - \ell) f(\ell) = 0.$$

Comme $\ell < b$ on a $f(\ell) = 0$ et donc $\ell = \xi$. □

On va maintenant chercher à estimer l'erreur entre x_n et ξ. La Proposition 2.7 donne

$$|x_{n+1} - \xi| \leq \frac{M}{2m} |x_{n+1} - x_n||x_{n+1} - b|,$$

ce qui permet de déterminer un test d'arrêt : si $\epsilon > 0$ est la précision souhaitée, on arrête dès que

$$\frac{M}{2m} |x_{n+1} - x_n||x_{n+1} - b| < \epsilon.$$

On peut en fait montrer que la convergence de cette méthode est linéaire (d'ordre 1, voir la Définition 2.1).

Proposition 2.10. *Si f est C^2 sur $[a, b]$ avec $f(a)f(b) < 0$, $f' > 0$ et $f'' > 0$, alors la convergence de $(x_n)_n$, définie en (2.4), vers l'unique zéro ξ de f dans $[a, b]$ est d'ordre 1.*

Démonstration. En écrivant $f(x_n) = f'(\xi)(x_n - \xi) + o(x_n - \xi)$, on a

$$x_{n+1} - \xi = \frac{(x_n - \xi) f(b) - (b - \xi) f(x_n)}{f(b) - f(x_n)} = \frac{(x_n - \xi)[f(b) - f'(\xi)(b - \xi)] + o(x_n - \xi)}{f(b) + o(1)}.$$

D'où

$$\frac{x_{n+1} - \xi}{x_n - \xi} \to \frac{f(b) - f'(\xi)(b - \xi)}{f(b)} =: c.$$

Comme f est strictement convexe, la Proposition 2.5 assure que $c > 0$ ce qui prouve que la convergence est bien d'ordre 1.

N.B. Comme $f' > 0$ on a $c < 1$, ce qui prouve à nouveau que $(x_n)_n$ tend vers ξ. On a en fait $|x_n - \xi| \simeq c^n$. □

2.5 Méthode de Newton

La méthode de Newton a une similitude avec la méthode de la sécante : à chaque itération on utilise une approximation affine. La différence est qu'on va utiliser la tangente à la courbe plutôt que la corde. Tout comme dans la section précédente on supposera que $f(a)f(b) < 0$, et que f' et f'' sont strictement positives. En particulier la fonction f possède une unique zéro $\xi \in [a, b]$.

Si $c \in [a, b]$, la tangente à la courbe C_f au point $(c, f(c))$ a pour équation

$$y = f'(c)(x - c) + f(c).$$

Elle coupe l'axe horizontal au point d'abscisse

$$x = c - \frac{f(c)}{f'(c)}. \tag{2.5}$$

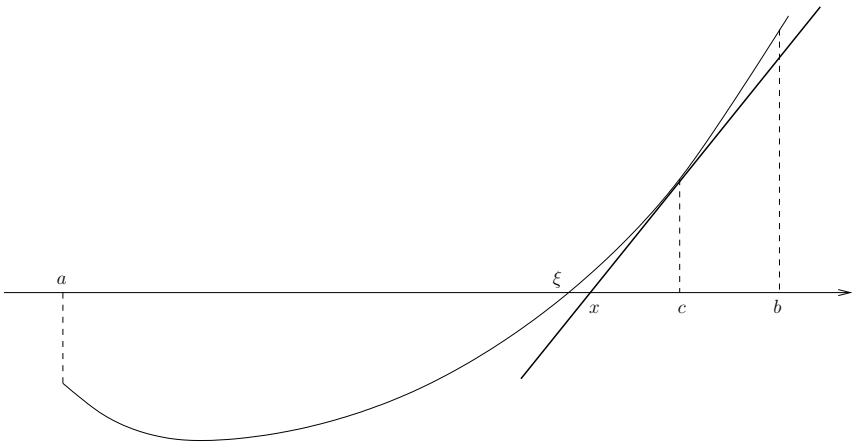

FIGURE 1.7 - Méthode de Newton

Comme pour la méthode de la sécante, l'idée est ensuite d'itérer la procédure en prenant comme nouvelle valeur de c l'abscisse x obtenue par (2.5). Il faut pour cela s'assurer que l'on reste bien dans $[a, b]$.

Lemme 2.11. *Soit f de classe C^2 sur $[a, b]$ telle que $f' > 0$, $f'' > 0$ et $f(a)f(b) < 0$. Soit $c \in [\xi, b]$ et $x := c - \frac{f(c)}{f'(c)}$. Alors $\xi \leq x \leq c \leq b$.*

Démonstration. Puisque $f' > 0$, f est strictement croissante et donc $f(c) \geq f(\xi) = 0$. On obtient donc $x \leq c$.

Par ailleurs, la fonction f est convexe ($f'' > 0$) donc, d'après la Proposition 2.5, son graphe est au-dessus de celui de sa tangente. Appliqué en ξ cela donne

$$f'(c)(\xi - c) + f(c) \leq f(\xi) = 0 \quad \Longleftrightarrow \quad \xi \leq c - \frac{f(c)}{f'(c)} = x.$$

□

On va maintenant itérer cette procédure en partant du point b. On considère donc la suite $(x_n)_n$ définie par

$$x_0 = b, \qquad x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad \forall n \in \mathbb{N}. \tag{2.6}$$

Le lemme ci-dessus montre que cette suite est bien définie, décroissante et minorée par l'unique zéro ξ de f dans $[a, b]$. La suite converge donc et sa limite ℓ vérifie

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

i.e. $f(\ell) = 0$. Autrement dit la suite $(x_n)_n$ converge vers ξ.

On va maintenant estimer l'erreur entre x_n et ξ. On montre le résultat suivant.

Proposition 2.12. *La convergence de $(x_n)_n$ vers ξ est d'ordre 2.*

Bien que la méthode semble à première vue similaire à celle de la sécante, on approche f par une fonction affine, on voit que la méthode de Newton converge en fait beaucoup plus vite.

Démonstration. On a

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

Comme $f(\xi) = 0$, la formule de Taylor-Lagrange assure l'existence de $\zeta_n \in [\xi, x_n]$ tel que

$$0 = f(\xi) = f(x_n) + f'(x_n)(\xi - x_n) + \frac{f''(\zeta_n)}{2}(\xi - x_n)^2,$$

d'où

$$x_{n+1} - \xi = \frac{f''(\zeta_n)}{2f'(x_n)}(x_n - \xi)^2.$$

Finalement, puisque f est C^2 et que $x_n \to \xi$ (et donc $\zeta_n \to \xi$) on a

$$\lim_{n\to+\infty} \frac{x_{n+1} - \xi}{(x_n - \xi)^2} = \frac{f''(\xi)}{2f'(\xi)} \neq 0,$$

ce qui prouve bien que la convergence est quadratique. □

Exemple 2.1. *Soit $a > 0$ et f définie par $f(x) = x^2 - a$. On peut vérifier que f' et f'' sont strictement positifs sur $\mathbb{R}_+^*$. Par ailleurs la fonction f a un seul zéro dans cet intervalle : $\sqrt{a}$. La méthode de Newton appliquée dans ce cas conduit à l'étude de la suite définie par récurrence*

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right).$$

Cette méthode de calcul approché d'une racine carrée est connue sous le nom d'algorithme de Babylone.

Remarque 2.6. *On donne ici un autre point de vue sur la méthode de Newton pour tenter d'éclairer la vitesse de convergence de celle-ci : quadratique au lieu de linéaire.*

Le théorème du point fixe affirme la chose suivante (voir TD et cours de Calcul Différentiel) : si I est fermé et $\varphi : I \to I$ est une fonction contractante, i.e. il existe $k \in [0,1[$ tel que pour tous $x, y \in I$ on a $|\varphi(y) - \varphi(x)| \leq k|y - x|$, alors φ admet un unique point fixe $\xi \in I$, i.e. tel que $\varphi(\xi) = \xi$. De plus la suite $(x_n)_n$ définie par récurrence par $x_0 \in I$ et $x_{n+1} = \varphi(x_n)$ converge vers ξ. Si φ est de classe C^1, la formule de Taylor montre alors que

$$|x_{n+1} - \xi| = |\varphi(x_n) - \varphi(\xi)| = |\varphi'(\xi)(x_n - \xi) + o(x_n - \xi)|.$$

Si $\varphi'(\xi) \neq 0$ la convergence est alors linéaire (on parle de point fixe attractif). Si par contre $\varphi'(\xi) = 0$ alors $\dfrac{x_{n+1} - \xi}{x_n - \xi} \to 0$ et on dit que le point fixe est super-attractif. Si par exemple φ est C^2, que $\varphi'(\xi) = 0$ mais $\varphi''(\xi) \neq 0$ alors la convergence est quadratique.

On s'intéresse ici aux racines d'une fonction f donnée. Pour se ramener au théorème du point fixe on peut bien sur prendre $\varphi(x) = x + f(x)$ mais on peut aussi prendre une fonction φ un peu plus générale, de la forme

$$\varphi(x) = x + f(x)g(x).$$

Si ξ est un zéro de f ce sera bien un point fixe de φ, et pourvu que la fonction g ne s'annule pas la réciproque est vraie, autrment dit $f(\xi) = 0$ ssi $\varphi(\xi) = \xi$. La suite $(x_n)_n$ définie par $x_{n+1} = \varphi(x_n)$ fournit donc une approximation de ξ, et cette approximation sera d'autant meilleure que $\varphi'(\xi)$ est petit. En particulier si $\varphi'(\xi) = 0$ (et que la fonction est C^2) la convergence sera au moins d'ordre 2. L'idée est donc de choisir la fonction g telle que $\varphi'(\xi)$ soit le plus petit possible en valeur absolue, et si possible soit nul. Or on a

$$\varphi'(\xi) = 1 + f'(\xi)g(\xi) + f(\xi)g'(\xi) = 1 + f'(\xi)g(\xi),$$

puisque $f(\xi) = 0$. Il suffit donc que g vérifie $g(\xi) = -\dfrac{1}{f'(\xi)}$. Comme on ne connait pas la valeur de ξ (on en cherche précisément une valeur approchée), et donc de $f'(\xi)$, la meilleure solution est de prendre $g = -\dfrac{1}{f'}$. La suite $(x_n)_n$ est alors définie par

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

ce qui est précisément la méthode de Newton.

Chapitre 3

Approximation polynomiale

Les méthodes de calcul approché d'intégrales vues au Chapitre 1 reposent sur le fait de "remplacer" la fonction f que l'on souhaite intégrer par un polynôme P et d'approcher l'intégrale de f par celle de P. Dans ce chapitre on étudie plus en détails la notion d'approximation polynomiale.

3.1 Interpolation de Lagrange

3.1.1 Existence et unicité du polynôme d'interpolation

Soient $f : [a, b] \to \mathbb{R}$ une fonction donnée et $a_0, a_1, \ldots, a_n$ dans $[a, b]$ deux à deux distincts. On notera $A = \{a_0, \ldots, a_n\}$ l'ensemble de ces points. On cherche un polynôme P tel que $P(a_i) = f(a_i)$ pour tout $a_i \in A$. Les a_i sont appelés les *noeuds d'interpolation*. Dans la mesure du possible on va chercher P de degré le plus bas possible. Si P est de degré k on a "à disposition" $k + 1$ paramètres (les coefficients de P). En d'autres termes on traduit simplement le fait que $\mathbb{R}_k[X]$ est de dimension $k + 1$. Comme ici on aura $n + 1$ contraintes, la valeura de P en chacun des noeuds d'interpolation a_i, il est donc raisonnable de chercher a priori P de degré (au plus) n.

Proposition 3.1. *Soient $a_0, a_1, \ldots, a_n$ dans $[a, b]$ deux à deux distincts et $y_0, \ldots, y_n$ des réels (pas nécessairement distincts). Il existe un unique polynôme $P_n \in \mathbb{R}_n[X]$ tel que*

$$P_n(a_i) = y_i, \qquad \forall i = 0, \ldots, n.$$

De plus, P_n est donné par

$$P_n = \sum_{i=0}^{n} y_i L_i, \qquad L_i(X) = \prod_{j \neq i} \frac{X - a_j}{a_i - a_j}. \tag{3.1}$$

Démonstration. Soit $\varphi : \mathbb{R}_n[X] \to \mathbb{R}^{n+1}$ définie par $\varphi(P) = (P(a_0), \ldots, P(a_n))$. On vérifie facilement que φ est linéaire et le résultat peut se reformuler de la façon suivante : φ est bijective.

Puisque $\mathbb{R}_n[X]$ et $\mathbb{R}^{n+1}$ sont tous les deux de dimension $n+1$, le théorème du rang montre qu'il suffit de vérifier que φ est injective ou surjective.

Soit donc $P \in \mathbb{R}_n[X]$ tel que $\varphi(P) = 0$. On a donc $P(a_i) = 0$ pour tout i. Les a_i étant tous distincts le polynôme P a donc au moins $n+1$ racines alors qu'il est de degré au plus n, c'est donc le polynôme nul et φ est injective.

On vérifie ensuite que pour tous i et j on a $L_i(x_j) = \delta_{ij}$ où $\delta_{ij} = 1$ si $i = j$ et vaut 0 sinon (symbole de Kronecker). D'où on vérifie qu'on a bien $P_n(a_j) = y_j$ pour tout j. □

Remarque 3.1. *Au lieu de montrer que φ est injective on aurait pu montrer qu'elle est surjective. Il suffit pour cela de vérifier que étant donné $y = (y_0, \dots, y_n) \in \mathbb{R}^{n+1}$ le polynôme P donné par 3.1 vérifie bien $\varphi(P) = y$, i.e. $P(a_i) = y_i$ pour tout i.*

Corollaire 3.2. *Soient $f : [a,b] \to \mathbb{R}$ une fonction donnée et $a_0, a_1, \dots, a_n$ dans $[a,b]$ deux à deux distincts. Il existe un unique $P_n \in \mathbb{R}_n[X]$ tel que $P_n(a_i) = f(a_i)$ pour tout $i \in \{0, \dots, n\}$.*

3.1.2 Erreur d'approximation

Si P_n est le polynôme d'interpolation de Lagrange d'une fonction f associé aux points $a_0, \dots, a_n$, il est naturel de savoir quelle est l'erreur commise en remplaçant f par P_n, l'idée étant qu'on utilisera ensuite P_n en prenant par exemple $\int_a^b P_n(x)\,\mathrm{d}x$ comme valeur approchée de $\int_a^b f(x)\,\mathrm{d}x$. Il faut bien sur préciser ici ce qu'on entend par erreur. On supposera toujours que la fonction f est au moins continue sur $[a,b]$. Sur l'espace des fonctions continues sur $[a,b]$ on considère (voir le cours de Calcul Différentiel) la norme infini $\|f\|_\infty = \sup_{x\in[a,b]} |f(x)|$, et on cherche donc à estimer l'erreur $\|f - P_n\|_\infty$.

Si la fonction f est juste continue on ne peut pas espérer avoir la moindre estimation. En effet, il n'y a alors aucune contrainte sur les variations de la fonction f entre deux noeuds d'interpolation consécutifs. Pour avoir une estimation de l'erreur il faut donc imposer plus de régularité à la fonction f.

Proposition 3.3. *Soit $f : [a,b] \to \mathbb{R}$ de classe C^{n+1} et $A = \{a_0, \dots, a_n\}$ un ensemble de points deux à deux distincts de $[a,b]$. Pour tout $x \in [a,b]$, il existe $\xi \in [a,b]$ tel que*

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

où $\omega_A(x) = \prod_{i=0}^{n}(x - a_i)$. En conséquence on a l'erreur d'approximation suivante

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

Lemme 3.4. *Si $g : I \to \mathbb{R}$ est de classe C^n et $a_0, \dots, a_n$ deux à deux distincts sont tels que $g(a_i) = 0$ pour tout i alors il existe $\xi \in I$ tel que $g^{(n)}(\xi) = 0$.*

Démonstration. On montre le résultat par récurrence sur n. Si $n = 1$ c'est le Théorème de Rolle. Si $n \geq 2$, on peut toujours supposer que les a_i sont rangés dans l'ordre croissant, et on applique le théorème de Rolle sur chacun des intervalles $[a_i, a_{i+1}]$. Il existe donc des $b_i \in]a_i, a_{i+1}[$ tels que $g'(b_i) = 0$. La fonction g' est de classe C^{n-1} et s'annule en chacun des points b_i, $i = 0, \ldots, n-1$ donc par hypothèse de récurrence il existe ξ tel que $g^{(n)}(\xi) = (g')^{(n-1)}(\xi) = 0$. □

Démonstration de la Proposition. On fixe $x \in [a, b]$. Si x est l'un des a_i le résultat est vrai pour tout ξ puisque dans ce cas $f(a_i) - P_n(a_i) = \omega_A(a_i) = 0$. On suppose donc que x est différent de tous les a_i. La preuve est inspirée de celle de la formule de Taylor-Lagrange. On considère la fonction

$$g(t) = f(t) - P_n(t) - \alpha\omega_A(t),$$

où α est choisi tel que $g(x) = 0$, i.e. $\alpha = \frac{f(x)-P_n(x)}{\omega_A(x)}$. La fonction g est de classe C^{n+1} et s'annule en $n+2$ points : x et les a_i. Ces points sont deux à deux distincts donc, d'après le lemme, il existe $\xi \in [a, b]$ tel que

$$g^{(n+1)}(\xi) = f^{(n+1)}(\xi) - P_n^{(n+1)}(\xi) - \alpha\omega_A^{(n+1)}(\xi) = 0.$$

Par ailleurs P_n est un polynôme de degré au plus n donc $P_n^{(n+1)}$ est nul, et ω_A est un polynôme unitaire de degré $n+1$ donc $\omega_A^{(n+1)}(\xi) = (n+1)!$. On en déduit que

$$\frac{f(x) - P_n(x)}{\omega_A(x)} = \alpha = \frac{f^{(n+1)}(\xi)}{(n+1)!} \quad \Longleftrightarrow \quad f(x) - P_n(x) = \frac{\omega_A(x)}{(n+1)!} f^{(n+1)}(\xi).$$

□

Remarque 3.2. *Le résultat ci-dessus montre que l'erreur d'approximation dépend des variations de la fonction f, via le terme $\|f^{(n+1)}\|_\infty$: plus la fonction oscille moins bonne est l'approximation, ce qui n'est pas surprenant. Elle dépend aussi du choix des noeuds d'interpolation, c'est le terme $\|\omega_A\|_\infty$. En particulier, il n'est pas toujours vrai que l'erreur d'approximation diminue lorsqu'on augmente le nombre de points d'interpolation : les dérivées successives de f peuvent être de plus en plus grandes et le terme $\|f^{(n+1)}\|_\infty$ ne pas être compensé par par le terme $\frac{\|\omega_A\|_\infty}{(n+1)!}$: c'est le phénomène dit de Runge (voir par exemple [1], Module IV.8. - Section 3).*

3.1.3 Stabilité du polynôme d'approximation

De façon générale, l'étude de la stabilité consiste à savoir comment varie le résultat lorsqu'on change (perturbe) les données initiales. On y reviendra dans le Chapitre 4.

Dans la pratique les valeurs de la fonction f que l'on interpole aux points a_i ne seront pas forcément connues avec exactitude. Elles auront par exemple été calculées avec un ordinateur qui n'en donnera ainsi qu'une valeur approchée ou bien simplement obtenues à partir d'instruments de mesures. La question est de savoir comment une erreur sur les $f(a_i)$ se répercute au niveau des polynômes d'interpolation. En d'autre termes quel est l'écart entre le polynôme d'interpolation P_n

associé à f et le polynôme d'interpolation Q_n calculé à partir de valeurs y_i, en fonction des écarts $f(a_i) - y_i$.

On a, avec les L_i définis par (3.1),

$$P_n(x) = \sum_{i=0}^{n} f(a_i) L_i(x) \quad \text{et} \quad Q_n(x) = \sum_{i=0}^{n} y_i L_i(x).$$

On en déduit la majoration suivante :

$$\begin{aligned} |P_n(x) - Q_n(x)| &= \left| \sum_{i=0}^{n} (f(a_i) - y_i) L_i(x) \right| \\ &\leq \sum_{i=0}^{n} |f(a_i) - y_i| \, |L_i(x)| \\ &\leq \max_i |f(a_i) - y_i| \times \sum_{i=0}^{n} |L_i(x)| \, . \end{aligned}$$

Si on note $\Lambda_n := \left\| \sum_{i=0}^{n} |L_i| \right\|_\infty$ on a alors

$$\|P_n - Q_n\|_\infty \leq \Lambda_n \max_i |f(a_i) - y_i|.$$

L'erreur commise sur les valeurs $f(a_i)$ est a priori amplifiée par la constante Λ_n, appelée *constante de Lebesgue*. Cette dernière s'exprime en terme des fonctions L_i, elle dépend donc uniquement du choix des points a_i. De façon générale on ne peut pas faire mieux que l'estimation ci-dessus. C'est précisément ce que dit la proposition suivante.

Proposition 3.5. *Soient $a_0, \ldots, a_n$ donnés et distincts, et soit $\varphi_n : C^0([a,b]) \to \mathbb{R}_n[X]$ l'application qui à f associe son polynôme d'interpolation aux points a_i, les espaces $C^0([a,b])$ et $\mathbb{R}_n[X]$ étant munis de la norme $\| \cdot \|_\infty$. Alors φ_n est une application linéaire continue et $|||\varphi_n||| = \Lambda_n$ où $\Lambda_n := \left\| \sum_{i=0}^{n} |L_i| \right\|_\infty$, i.e. Λ_n est la plus petite constante $C \geq 0$ telle que $\|\varphi_n(f)\|_\infty \leq C \|f\|_\infty$ pour tout $f \in C^0([a,b])$.*

Démonstration. La linéarité de φ_n est évident et laissée à titre d'exercie. La continuité se montre

ensuite de la même façon que ci-dessus. Soit $f \in C^0([a,b])$. On a pour tout x

$$\begin{aligned} |\varphi_n(f)(x)| &= \left|\sum_{i=0}^{n} f(a_i)L_i(x)\right| \\ &\leq \sum_{i=0}^{n} |f(a_i)|\,|L_i(x)| \\ &\leq \max_i |f(a_i)| \sum_{i=0}^{n} |L_i(x)| \\ &\leq \|f\|_\infty \sum_{i=0}^{n} |L_i(x)| \\ &\leq \Lambda_n \|f\|_\infty, \end{aligned}$$

et donc $\|\varphi_n(f)\|_\infty \leq \Lambda_n \|f\|_\infty$, ce qui prouve que φ_n est continue et que $|||\varphi_n||| \leq \Lambda_n$.

Pour montrer l'égalité on va chercher une fonction f telle que $\|\varphi_n(f)\|_\infty = \Lambda_n \|f\|_\infty$. Pour avoir égalité dans le calcul précédent on prend $x_0 \in [a,b]$ tel que $\sum_{i=0}^{n} |L_i(x_0)| = \Lambda_n$ (c'est toujours possible puisqu'on a une fonction continue sur $[a,b]$ qui est un segment). On choisit ensuite une fonction f telle que pour tout i on ait $|f(a_i)| = \|f\|_\infty$ (cela impose juste la valeur de f au a_i. Finalement on choisit le signe de $f(a_i)$ de façon à ce que $f(a_i)L_i(x_0)$ soit positif pour tout i. Au point x_0 toutes les inégalités ci-dessus sont des égalités, autrement dit $|\varphi_n(f)(x_0)| = \Lambda_n \|f\|_\infty$, et donc $\|\varphi_n(f)\|_\infty = \Lambda_n \|f\|_\infty$. Il reste à s'assurer que f est continue, il suffit pour cela de la prendre affine par morceaux, affine sur chacun des intervalles $[a_i, a_{i+1}]$ (on peut toujours supposer les a_i dans l'ordre croissant) et constante à gauche de a_0 et à droite de a_n.

Remarque 3.3. *La valeur de la constante Λ_n dépend uniquement du choix des noeuds d'interpolation. On peut montrer par exemple que si on choisit des points équidistants alors $\Lambda_n \sim \dfrac{2^{n+1}}{en\ln(n)}$: la constante Λ_n croit très rapidement avec le nombre de points, voir par exemple [1, 3].*

□

3.2 Approximation L^2 et polynômes orthogonaux

3.2.1 Généralités sur l'approximation polynomiale

On a vu dans la section précédente une façon d'approcher une fonction f donnée par un polynôme ainsi qu'une estimation de l'erreur commise. La première question qui se pose naturellement est "ce choix de polynôme est-il le meilleur ?" Il faut également préciser dans quel sens il serait le meilleur. La seconde question est "si j'augmente le nombre n de points l'estimation est-elle meilleure ? Tend-elle toujours vers 0 lorsque n augmente ?" Les réponses à ces deux questions

sont non et non ! Le choix n'est pas forcément le meilleur et l'erreur ne tend pas forcément vers 0 lorsque n augmente. Pire, cette dernière peut même augmenter (phénomène de Runge).

Le premier résultat est assez général et concerne l'existence d'une meilleure approximation.

Théorème 3.6. *Soit $\|\cdot\|$ une norme quelconque sur $C^0([a,b])$. Pour tout $f \in C^0([a,b])$ et pour tout $n \in \mathbb{N}$ il existe au moins un polynôme de meilleur approximation dans $\mathbb{R}_n[X]$, i.e. il existe $P_n \in \mathbb{R}_n[X]$ tel que*

$$\|f - P_n\| = \inf_{Q \in \mathbb{R}_n[X]} \|f - Q\|.$$

Démonstration. Tout repose sur la continuité de la norme et le fait que $\mathbb{R}_n[X]$ est de dimension finie. On considère la fonction $Q \mapsto \|f - Q\|$ de $\mathbb{R}_n[X]$ dans $\mathbb{R}$. Elle est continue (continuité de la norme, voir le cours de Calcul Différentiel). On remarque ensuite que

$$\inf_{Q \in \mathbb{R}_n[X]} \|f - Q\| \leq \|f - 0\| = \|f\|.$$

Si $\|Q\| > 2\|f\|$ on a $\|Q - f\| \geq \|Q\| - \|f\| > \|f\|$ d'où on déduit que

$$\inf_{Q \in \mathbb{R}_n[X]} \|f - Q\| = \inf_{\substack{Q \in \mathbb{R}_n[X] \\ B(0, 2\|f\|)}} \|f - Q\|.$$

où $B(0, 2\|f\|)$ est la boule fermée de $\mathbb{R}_n[X]$ de centre 0 et de rayon $2\|f\|$. Comme $\mathbb{R}_n[X]$ est de dimension finie, $B(0, 2\|f\|)$ est un compact et donc (fonction continue sur un compact) il existe $P \in B(0, 2\|f\|) \subset \mathbb{R}_n[X]$ tel que

$$\|f - P_n\| = \inf_{\substack{Q \in \mathbb{R}_n[X] \\ B(0, 2\|f\|)}} \|f - Q\| = \inf_{Q \in \mathbb{R}_n[X]} \|f - Q\|.$$

□

Remarque 3.4. *Le théorème assure que pour tout choix de norme il existe un polynôme de meilleure approximation, mais il n'y a pas toujours unicité. Cela dépend du choix de la norme. C'est vrai pour la norme $\|\cdot\|_\infty$, c'est également vrai pour la norme $\|\cdot\|_2$ (voir la Section 3.2.2).*

On termine cette section avec le Théorème de Weierstrass qui concerne l'approximation uniforme d'une fonction continue par des polynômes.

Théorème 3.7 (Weierstrass)**.** *Soit $f \in C^0([a,b])$. Il existe une suite de polynômes $(P_n)_n$ telle que* $\lim_{n \to \infty} \|f - P_n\|_\infty = 0$.

En d'autres termes, ce théorème dit que pour le choix de la norme $\|\cdot\|_\infty$ la suite des polynômes de meilleure approximation converge vers f : l'erreur tend vers 0.

Démonstration. On va donner une preuve constructive (la suite $(P_n)_n$ sera explicite) de ce théorème en utilisant les polynômes dits de Bernstein.

On commence par remarquer qu'il suffit de considérer le cas où $[a,b] = [0,1]$. En effet, si $f \in C^0([a,b])$ alors la fonction $g(t) = f(a + t(b-a))$ est continue sur $[0,1]$. Si $Q_n(t)$ est une suite de polynôme qui converge uniformément vers g sur $[0,1]$ alors la suite de polynômes $P_n(x) = Q_n\left(\frac{x-a}{b-a}\right)$ converge uniformément vers f sur $[a,b]$. En effet, pour tout $x \in [a,b]$

$$|P_n(x) - f(x)| = \left|Q_n\left(\frac{x-a}{b-a}\right) - g\left(\frac{x-a}{b-a}\right)\right| \leq \|Q_n - g\|_\infty.$$

Soit donc $f \in C^0([0,1])$. Pour tout $n \geq 1$ on définit le polynôme

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

On va montrer que $\|B_n - f\|_\infty \to 0$. Soit donc $\epsilon > 0$, on montre que pour n assez grand $\|B_n - f\|_\infty < \epsilon$, en utilisant une approche probabiliste. La fonction f est continue sur $[0,1]$ donc d'après le Théorème de Heine elle est uniformément continue. Il existe donc $\delta > 0$ tel que

$$|x-y| < \delta \implies |f(x) - f(y)| < \epsilon. \tag{3.2}$$

Soit maintenant $x \in [0,1]$. Pour tout n on considère une variable aléatoire S_n suivant la loi binomiale $Bin(n,x)$. Elle prend ses valeurs dans $\{0,\ldots,n\}$ et pour tout k on a $\mathbb{P}(S_n = k) = \binom{n}{k} x^k (1-x)^{n-k}$. On peut ainsi écrire que

$$B_n(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\mathbb{P}(S_n = k) = \mathbb{E}\left(f\left(\frac{S_n}{n}\right)\right).$$

On a alors, en utilisant que $\sum_{k=0}^{n} \mathbb{P}(S_n = k) = 1$,

$$\begin{aligned}
|B_n(x) - f(x)| &= \left|\sum_{k=0}^{n}\left(f\left(\frac{k}{n}\right) - f(x)\right)\mathbb{P}(S_n = k)\right| \\
&\leq \sum_{k=0}^{n}\left|f\left(\frac{k}{n}\right) - f(x)\right|\mathbb{P}(S_n = k) \\
&\leq \sum_{|\frac{k}{n}-x|<\delta}\left|f\left(\frac{k}{n}\right) - f(x)\right|\mathbb{P}(S_n = k) + \sum_{|\frac{k}{n}-x|\geq\delta}\left|f\left(\frac{k}{n}\right) - f(x)\right|\mathbb{P}(S_n = k) \\
&\leq \epsilon \sum_{|\frac{k}{n}-x|<\delta}\mathbb{P}(S_n = k) + 2\|f\|_\infty \sum_{|\frac{k}{n}-x|\geq\delta}\mathbb{P}(S_n = k) \\
&\leq \epsilon + 2\|f\|_\infty \mathbb{P}\left(\left|\frac{S_n}{n} - x\right| \geq \delta\right).
\end{aligned}$$

Comme $\mathbb{E}(S_n) = nx$ et $Var(S_n) = nx(1-x)$, d'après l'inégalité de Bienaymé-Tchebychev on a

$$\mathbb{P}\left(\left|\frac{S_n}{n} - x\right| \geq \delta\right) = \mathbb{P}(|S_n - nx| \geq n\delta) \leq \frac{Var(S_n)}{n^2\delta^2} = \frac{x(1-x)}{n\delta^2}.$$

On montre facilement que pour tout $x \in [0,1]$ on a $x(1-x) \leq \frac{1}{4}$, et ainsi pour tout $x \in [0,1]$ et tout n on a

$$|B_n(x) - f(x)| \leq \epsilon + \frac{\|f\|_\infty}{2n\delta^2},$$

et donc

$$\|B_n - f\|_\infty \leq \epsilon + \frac{\|f\|_\infty}{2n\delta^2}. \tag{3.3}$$

Pour n assez grand on a donc bien $\|B_n - f\|_\infty < 2\epsilon$. □

Remarque 3.5. *Si f est L-Lipschitzienne, i.e. $|f(x) - f(y)| \leq L|x-y|$ pour tous $x, y \in [a,b]$, (c'est le cas par exemple si f est C^1 avec $L = \|f'\|_\infty$ d'après l'inégalité des accroissements finis) on peut prendre $\delta = \frac{\epsilon}{L}$ et (3.3) devient*

$$\|B_n - f\|_\infty \leq \epsilon + \frac{L^2\|f\|_\infty}{2n\epsilon^2}, \quad \forall \epsilon > 0,\ \forall n \in \mathbb{N}^*.$$

Pour un n donné, on peut chercher pour quelle valeur de $\epsilon > 0$ le membre de droite est minimum, ce qui donnera la meilleure estimation possible de l'erreur entre f et B_n que l'on puisse obtenir par le calcul précédent. Une étude rapide de la fonction $g(\epsilon) = \epsilon + \frac{L^2\|f\|_\infty}{2n\epsilon^2}$ sur $]0, +\infty[$ montre que le minimum est atteint pour $\epsilon = \left(\frac{L^2\|f\|_\infty}{n}\right)^{1/3}$ et vaut alors $\frac{3}{2}\left(\frac{L^2\|f\|_\infty}{n}\right)^{1/3}$. Autrement dit, on a l'erreur d'approximation suivante

$$\|B_n - f\|_\infty \leq \frac{3L^{2/3}\|f\|_\infty^{1/3}}{2n^{1/3}}, \quad \forall n \in \mathbb{N}^*.$$

3.2.2 Polynôme de meilleure approximation L^2

On va s'intéresser ici au polynôme de meilleure approximation pour la norme L^2 :

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

Cette norme a le gros avantage de provenir d'un produit scalaire.

Proposition 3.8. *La forme bilinéaire $\langle f, g\rangle := \int_a^b f(x)g(x)\,\mathrm{d}x$ définit un produit scalaire sur $C^0([a,b])$ et $\|f\|_2$ est la norme associée.*

Exercice 3.1. Démontrer la propostion ci-dessus.

On ne le précisera pas dans la suite mais le produit scalaire dépend de l'intervalle $[a, b]$.

Théorème 3.9. *Pour tout $f \in C^0([a, b])$ et tout $n \in \mathbb{N}$ il existe un unique $P_n \in \mathbb{R}_n[X]$ tel que*

$$\|f - P_n\|_2 = \inf_{Q \in \mathbb{R}_n[X]} \|f - Q\|_2. \tag{3.4}$$

On a la caractérisation suivante : P_n est l'unique polynôme $P \in \mathbb{R}_n[X]$ vérifiant

$$\langle f - P, Q\rangle = 0, \qquad \forall Q \in \mathbb{R}_n[X], \tag{3.5}$$

i.e. c'est l'unique élément de $\mathbb{R}_n[X]$ tel que $f - P \in (\mathbb{R}_n[X])^\perp$.

Remarque 3.6. *Ce résultat est un cas particulier du résultat plus général suivant : si E est un espace vectoriel (ici $C^0([a, b])$) muni d'un produit scalaire et F (ici $\mathbb{R}_n[X]$) un sous-espace vectoriel fermé de E alors pour tout $u \in E$ il existe un unique $v \in F$ tel que $\|u - v\| = \inf_{w\in F} \|u - w\|$, et c'est l'unique élément de F tel que $\langle u - v, w\rangle = 0$ pour tout $w \in F$.*

Démonstration. L'existence de P_n est un cas particulier du Théorème 3.6. Il reste à montrer l'unicité et la caractérisation (3.5).

On montre d'abord la caractérisation de P_n, c'est-à-dire que P_n vérifie (3.4) si et seulement si P_n vérifie (3.5). Supposons d'abord que P_n vérifie (3.5). Soit $Q \in \mathbb{R}_n[X]$, on écrit

$$\|f - Q\|_2^2 = \|f - P_n + P_n - Q\|_2^2 = \|f - P_n\|_2^2 + \|P_n - Q\|_2^2 + 2\langle f - P_n, P_n - Q\rangle,$$

où on a utilisé (bilinéarité du produit scalaire)

$$\|a + b\|_2^2 = \langle a + b, a + b\rangle = \|a\|_2^2 + 2\langle a, b\rangle + \|b\|_2^2.$$

On a $P_n - Q \in \mathbb{R}_n[X]$ donc par hypothèse $\langle f - P_n, P_n - Q\rangle = 0$. On en déduit que pour tout $Q \in \mathbb{R}_n[X]$ on a

$$\|f - Q\|_2^2 = \|f - P_n\|_2^2 + \|P_n - Q\|_2^2 \geq \|f - P_n\|_2^2,$$

et donc P_n vérifie (3.4).

Réciproquement, supposons que P_n vérifie (3.4). Pour tout $Q \in \mathbb{R}_n[X]$ on a

$$\begin{aligned} & \|f - P_n\|_2^2 \leq \|f - Q\|_2^2 = \|f - P_n\|_2^2 + \|P_n - Q\|_2^2 + 2\langle f - P_n, P_n - Q\rangle \\ \iff\ & \|P_n - Q\|_2^2 + 2\langle f - P_n, P_n - Q\rangle \geq 0. \end{aligned}$$

On applique cette inégalité au polynôme $Q_t = P_n - tQ$ où $t \in \mathbb{R}$. On a bien $Q_t \in \mathbb{R}_n[X]$ et $P_n - Q_t = tQ$ donc

$$t^2\|Q\|_2^2 + 2t\langle f - P_n, Q\rangle \geq 0.$$

Cette inégalité est vraie pour tout $t \in \mathbb{R}$ donc le discriminant Δ du polynôme du segond degré (en t) $t^2\|Q\|_2^2 + 2t\langle f - P_n, Q\rangle$ vérifie $\Delta \leq 0$. Or $\Delta = 4\langle f - P_n, Q\rangle^2$, on a donc $\langle f - P_n, Q\rangle = 0$ ce qui prouve que P_n vérifie (3.5).

On montre finalement l'unicité. D'après ce qui précède il suffit pour cela de vérifier qu'il y a un unique polynôme P vérifiant (3.5). Soient P et $\tilde{P}$ deux éléments de $\mathbb{R}_n[X]$ vérifiant (3.5). Pour tout $Q \in \mathbb{R}_n[X]$ on a

$$\langle f - P, Q\rangle = 0 \quad \text{et} \quad \langle f - \tilde{P}, Q\rangle = 0.$$

En faisant la différence de ces deux équations on obtient

$$\langle \tilde{P} - P, Q\rangle = 0, \qquad \forall Q \in \mathbb{R}_n[X].$$

On applique cette identité au polynôme $Q = \tilde{P} - P$ ce qui donne $\|\tilde{P} - P\|_2^2 = 0$ et donc $\tilde{P} = P$.

□

Tout comme pour la norme uniforme on peut alors montrer que lorsque n augmente la suite des polynômes de meilleure approximation L^2 tend vers f.

Théorème 3.10. *Pour tout $n \in \mathbb{N}$ on note P_n l'unique élément de $\mathbb{R}_n[X]$ vérifiant*

$$\|f - P_n\|_2 = \inf_{Q\in\mathbb{R}_n[X]} \|f - Q\|_2.$$

Alors $\lim_{n\to\infty} \|f - P_n\|_2 = 0$.

Démonstration. Soit $Q_n \in \mathbb{R}_n[X]$ telle que $\|f - Q_n\|_\infty \to 0$. Une telle suite existe d'après le Théorème de Weierstrass, on peut par exemple prendre les polynômes de Bernstein. Par définition de P_n on a

$$\|f - P_n\|_2 \leq \|f - Q_n\|_2.$$

Par ailleurs,

$$\|f - Q_n\|_2^2 \;=\; \int_a^b (f(x) - Q_n(x))^2 \,\mathrm{d}x \;\leq\; \|f - Q_n\|_\infty^2 \int_a^b \mathrm{d}x \;=\; (b-a)\|f - Q_n\|_\infty^2.$$

On obtient donc

$$\|f - P_n\|_2 \leq \sqrt{b-a}\|f - Q_n\|_\infty,$$

et le résultat découle du Théorème des gendarmes. □

3.2.3 Polynômes orthogonaux

On voudrait maintenant pouvoir calculer le polynôme P_n de meilleure approximation L^2. Là encore le fait que la norme provienne d'un produit scalaire facilite les choses. On sait que $P_n \in \mathbb{R}_n[X]$. Pour connaitre P_n il suffit d'avoir ses coordonnées dans une base donnée. La première base qui vienne à l'esprit est bien sûr la base $(1, X, X^2, \ldots, X^n)$. Ce n'est cependant pas la base la plus adaptée ici. Quand on a un produit scalaire le mieux est d'utiliser une base orthonormée pour ce produit scalaire. Pour tout n l'espace vectoriel $\mathbb{R}_n[X]$ est de dimension finie donc il possède des bases orthonormées (voir le cours d'Algèbre Bilinéaire de L2). De plus on peut en construire explicitement à l'aide du procédé d'orthogonalisation de Gram-Schmidt.

Proposition 3.11. *Il existe une unique suite $(p_n)_n$ de polynômes telle que*

1. *pour tout $n \in \mathbb{N}$, p_n est de degré n et son coefficient dominant est 1.*
2. *pour tous $n \neq m \in \mathbb{N}$ on a $\langle p_n, p_m \rangle = 0$ (famille orthogonale).*

Remarque 3.7. *La famille $(p_n)_n$ n'est pas orthonormée puisqu'on n'a pas nécessairement $\|p_n\|_2 = 1$. C'est un choix de convention dans la définition des p_n. Pour obtenir une famille orthonormée il suffit de considérer $\tilde{p}_n = \dfrac{p_n}{\|p_n\|_2}$.*

Remarque 3.8. *Le produit scalaire dépend de l'intervalle $[a, b]$, les polynômes p_n en dépendent donc également.*

Démonstration. L'existence provient de l'application du procédé de Gram-Schmidt à partir de la famille $(X^n)_{n\in\mathbb{N}}$. On rappelle que par construction on aura, pour tout $n \in \mathbb{N}$,

$$\mathrm{Vect}(p_0, \ldots, p_n) = \mathrm{Vect}(1, X, \ldots, X^n) = \mathbb{R}_n[X],$$

d'où on déduit que p_n est de degré exactement n et, quitte à le multiplier par une constante, on peut donc toujours se ramener au cas où son coefficient dominant est 1.

On montre l'unicité. Soient $(p_n)_n$ et $(q_n)_n$ deux telles suites. Etant donné $n \in \mathbb{N}$ on a $\langle p_n, p_k \rangle$ pour tout $k < n$ donc $p_n \in (\mathrm{Vect}(p_0, \ldots, p_{n-1}))^\perp = \mathbb{R}_{n-1}[X]^\perp$. De même $q_n \in \mathbb{R}_{n-1}[X]^\perp$ d'où $p_n - q_n \in \mathbb{R}_{n-1}[X]^\perp$. Or p_n et q_n sont de degré n et ont coefficient dominant 1. Donc $p_n - q_n \in \mathbb{R}_{n-1}[X]$. Ainsi $p_n - q_n \in \mathbb{R}_{n-1}[X] \cap \mathbb{R}_{n-1}[X]^\perp$ et donc $\|p_n - q_n\|_2^2 = \langle p_n - q_n, p_n - q_n \rangle = 0$ et donc $p_n - q_n = 0$. □

Exemple 3.1. *On prend $[a, b] = [-1, 1]$. On montre comment construire les premiers polynômes orthogonaux. On a clairement $p_0 = 1$. Le polynôme p_1 est de la forme $p_1(x) = x + \alpha$ où $\alpha \in \mathbb{R}$ est tel que*

$$0 = \langle p_1, p_0 \rangle = \int_{-1}^{1} (x + \alpha)\,\mathrm{d}x = 2\alpha,$$

d'où $p_1(x) = x$.

Le polynôme p_2 est de la forme $p_2(x) = x^2 + \alpha x + \beta$ où $\alpha, \beta \in \mathbb{R}$ sont tels que

$$0 = \langle p_2, p_0 \rangle = \int_{-1}^{1} x^2 + \alpha x + \beta\,\mathrm{d}x = \frac{2}{3} + 2\beta \quad \textit{et} \quad 0 = \langle p_2, p_1 \rangle = \int_{-1}^{1} x^3 + \alpha x^2 + \beta x\,\mathrm{d}x = \frac{2\alpha}{3}.$$

On en déduit que $p_2(x) = x^2 - \frac{1}{3}$.

On peut montrer que pour tout $n \in \mathbb{N}$ on a $p_n(x) = \dfrac{n!}{(2n)!}\dfrac{\mathrm{d}^n}{\mathrm{d}x^n}\left((x^2-1)^n\right)$. Ces polynômes sont appelés polynômes de Legendre.

Exercice 3.2. On note p_n les polynômes de Legendre de l'exemple précédent. Montrer que la famille $(q_n)_n$ définie par

$$q_n(x) = \left(\frac{b-a}{2}\right)^n p_n\left(\frac{2x-a-b}{b-a}\right)$$

est la famille des polynômes orthogonaux relatifs à l'intervalle $[a, b]$.

Remarque 3.9. *La notion de polynômes orthogonaux est plus générale que celle vue ici, voir par exemple le Module VI.1 de [1]. Etant donnés un intervalle $]a,b[$ et une fonction $w :]a,b[\to \mathbb{R}$ continue, positive, non identiquement nulle et telle que pour tout $n \in \mathbb{N}$ l'intégrale $\int_a^b x^n w(x)\,\mathrm{d}x$ converge, une telle fonction est appelé un poids, l'application $\langle P,Q\rangle_w := \int_a^b P(x)Q(x)w(x)\,\mathrm{d}x$ définit alors un produit scalaire sur $\mathbb{R}[X]$. On peut ainsi définir la famille de polynômes orthogonaux relative à ce produit scalaire. Les polynômes de Legendre correspondent au cas $w \equiv 1$ sur l'intervalle $]-1,1[$. Par exemple, la fonction $w(x) = \mathrm{e}^{-x^2}$ définit un poids sur $\mathbb{R}$ et les polynômes orthogonaux correspondant sont appelés polynômes de Hermite.*

Théorème 3.12. *Soit $f \in C^0([a,b])$. Alors pour tout $n \in \mathbb{N}$ le polynôme de meilleure approximation L^2 de f dans $\mathbb{R}_n[X]$ est donné par*

$$P_n = \sum_{k=0}^{n} \frac{\langle f, p_k\rangle}{\|p_k\|_2^2} p_k, \tag{3.6}$$

où $(p_n)_n$ est la suite de polynômes orthogonaux donnée par la Proposition 3.11.

Démonstration. D'après le Théorème 3.9, P_n est caractérisé par le fait que $\langle f, Q\rangle = \langle P_n, Q\rangle$ pour tout $Q \in \mathbb{R}_n[X]$. Comme $\mathbb{R}_n[X] = \mathrm{Vect}(p_0,\dots,p_n)$ il suffit donc de montrer que pour tout $j = 0,\dots,n$ on a

$$\langle f, p_j\rangle = \left\langle \sum_{k=0}^{n} \frac{\langle f, p_k\rangle}{\|p_k\|_2^2} p_k, p_j \right\rangle.$$

En utilisant le fait que les p_k sont orthogonaux on a, pour tout $j = 0,\dots,n$,

$$\left\langle \sum_{k=0}^{n} \frac{\langle f, p_k\rangle}{\|p_k\|_2^2} p_k, p_j \right\rangle = \sum_{k=0}^{n} \frac{\langle f, p_k\rangle}{\|p_k\|_2^2} \langle p_k, p_j\rangle = \frac{\langle f, p_j\rangle}{\|p_j\|_2^2} \langle p_j, p_j\rangle = \langle f, p_j\rangle.$$

□

Remarque 3.10. *En combinant le Théorème ci-dessus avec le Théorème 3.10 on pourra écrire*

$$f = \sum_{k=0}^{\infty} \frac{\langle f, p_k\rangle}{\|p_k\|_2^2} p_k.$$

Cette égalité est à comprendre dans le sens $\lim_{n\to\infty} \left\| f - \sum_{k=0}^{n} \frac{\langle f, p_k\rangle}{\|p_k\|_2^2} p_k \right\|_2 = 0$. Elle est à rapprocher de ce qu'on fait en séries de Fourier quand on écrit

$$f(x) = \sum_{k\in\mathbb{N}} a_k(f)\cos(kx) + b_k(f)\sin(kx).$$

Attention, cela ne veut pas dire que les ploynômes orthogonaux forment une base au sens algébrique du terme. Il faudrait pour cela pouvoir écrire f comme une somme finie de p_k ce qui n'est possible que si la fonction f elle-même est un polynôme.

Chapitre 4

Résolution numérique des équations différentielles

4.1 Quelques aspects théoriques

Avant de s'intéresser à la résolution approchée d'équations différentielles on va commencer par présenter quelques résultats généraux, en particulier concernant l'existence et l'unicité de solutions : avant d'essayer d'approcher la solution il est important de savoir si on "approche quelque chose" et si oui "ce qu'on approche".

Dans tout le chapitre I est un intervalle ouvert de $\mathbb{R}$, Ω un ouvert de $\mathbb{R}^d$ et $f : I \times \Omega \to \mathbb{R}^d$ une fonction. On s'intéresse aux équations de la forme

$$y'(t) = f(t, y(t)), \tag{4.1}$$

où y est une fonction dérivable de t à valeurs dans Ω. Une telle équation est dite du premier ordre et mise sous forme résolue (il n'y a pas de coefficient devant le y').

Définition 4.1. *On appelle solution de l'équation différentielle (4.1) tout couple (y, J) où $J \subset I$ est un intervalle d'intérieur non-vide et $y : J \to \Omega$ est une fonction dérivable vérifiant (4.1) pour tout $t \in \overset{\circ}{J}$.*

Remarque 4.1. *L'intervalle J peut a priori être plus petit que l'intervalle I sur lequel est définie la fonction f (voir l'exemple ci-dessous).*

Exemple 4.1. *Soit f la fonction définie sur $I \times \Omega = \mathbb{R} \times \mathbb{R}$ par $f(t, y) = y^2$. On peut montrer que les solutions non nulles sont les fonctions de la forme $y(t) = \frac{1}{a-t}$, où $a \in \mathbb{R}$, définies (au maximum) soit sur l'intervalle $]-\infty, a[$ soit sur $]a, +\infty[$. Mise à part la fonction nulle, en aucun cas elles ne peuvent être définies sur $I = \mathbb{R}$.*

Remarque 4.2. *Une équation différentielle de la forme $y^{(n)} = f(t, y, y', \ldots, y^{(n-1)})$ est dite (résolue) d'ordre n. Elle peut toujours se ramener à une équation du premier ordre. En effet, en posant $Y(t) = \big(y(t), y'(t), \ldots, y^{(n-1)}(t)\big)$ alors Y est solution de l'équation $Y'(t) = F(t, Y(t))$ où $F(t, y_0, \ldots, y_{n-1}) = (y_1, \ldots, y_{n-1}, f(t, y_0, \ldots, y_{n-1}))$.*

Pour une équation différentielle du premier ordre on précise souvent en plus une *condition initiale* du type $y(t_0) = y_0$ où $(t_0, y_0) \in I \times \Omega$. Trouver une solution (y, J) de (4.1) vérifiant une condition du type $y(t_0) = y_0$ avec J voisinage de t_0 est appelé *problème de Cauchy*. Un des principaux résultats sur les équations différentielles est le théorème dit de Cauchy-Lipschitz concernant l'existence et l'unicité des solutions du problème de Cauchy. On rappelle d'abord la notion de fonction (localement) Lipschitzienne.

Définition 4.2. *1. Une fonction $f : I \times \Omega \to \mathbb{R}^d$ est dite (globalement) Lipshitzienne par rapport à la seconde variable s'il existe $k \geq 0$ tel que pour tout $t \in I$ et tous $y_1, y_2 \in \Omega$ on ait*

$$\|f(t, y_1) - f(t, y_2)\| \leq k\|y_1 - y_2\|.$$

La constante k est appelée constante de Lipschitz, on dit aussi que f est k-Lipschitzienne.

2. Une fonction $f : I \times \Omega \to \mathbb{R}^d$ est dite localement Lipshitzienne par rapport à la seconde variable si pour tout $(t_0, y_0) \in I \times \Omega$ il existe un voisinage V de (t_0, y_0) et $k \geq 0$ tel que pour tous (t, y_1) et (t, y_2) dans V on ait

$$\|f(t, y_1) - f(t, y_2)\| \leq k\|y_1 - y_2\|.$$

Proposition 4.3. *Si f est de classe C^1 alors elle est localement Lipshitzienne par rapport à la seconde variable.*

Démonstration. C'est une conséquence de l'inégalité des accroissements finis (voir le cours de Calcul Différentiel). Il suffit en fait que f possède une différentielle partielle par rapport à la seconde variable et que cette dernière soit continue. □

Théorème 4.4 (Cauchy-Lipschitz)**.** *Soit $f : I \to \Omega$ continue et localement Lipshitzienne par rapport à la seconde variable. Pour tout $(t_0, y_0) \in I \times \Omega$ il existe une solution (y, J), avec $t_0 \in \overset{\circ}{J}$, au problème de Cauchy*

$$\begin{cases} y'(t) = f(t, y(t)), \\ y(t_0) = y_0. \end{cases}$$

Cette solution est localement unique *dans le sens suivant : si (y_1, J_1) et (y_2, J_2) sont deux telles solutions alors $y_1(t) = y_2(t)$ pour tout $t \in J_1 \cap J_2$ (qui est un intervalle non-vide contenant t_0).*

Remarque 4.3. *Le théorème reste vrai si on remplace $\mathbb{R}^d$ par E espace de Banach.*

Remarque 4.4. *Le théorème ci-dessus affirme l'existence d'une solution définie sur un voisinage de t_0. Par ailleurs, si on a deux solutions elles ne peuvent différer que par l'intervalle sur lequel elles sont définies. Si on a deux solutions (y_1, J_1) et (y_2, J_2) on peut facilement voir que si $J = J_1 \cup J_2$ et y est définie par $y(t) = y_i(t)$ si $t \in J_i$ (il n'y a pas d'ambiguité puisque $y_1 = y_2$ sur $J_1 \cap J_2$) alors (y, J) est aussi une solution.*

Étant donnée la remarque précédente, il est naturel de chercher une solution dont l'intervalle J de définition soit le plus grand possible, c'est la notion de solution maximale.

Définition 4.5. *Une solution (y, J) de l'équation différentielle $y'(t) = f(t, y(t))$ est dite maximale si toute solution $(\tilde{y}, \tilde{J})$ telle que $J \subset \tilde{J}$ et $y(t) = \tilde{y}(t)$ pour tout $t \in J$ (on dit que $(\tilde{y}, \tilde{J})$ est un prolongement de (y, J)) vérifie $\tilde{J} = J$. Autrement dit, il n'existe pas de solution qui coincide avec y sur J et définie sur un intervalle plus grand que J. Ou encore, le seul prolongement de la solution (y, J) qui soit aussi solution est elle-même.*

On peut alors reformuler le Théorème de Cauchy-Lipschitz à l'aide de la notion de solution maximale.

Théorème 4.6 (Cauchy-Lipschitz)**.** *Soit $f : I \to \Omega$ continue et localement Lipshitzienne par rapport à la seconde variable. Pour tout $(t_0, y_0) \in I \times \Omega$ il existe une unique solution maximale (y, J) au problème de Cauchy.*

Démonstration. On montre simplement comment passer de l'existence et unicité locale à celle de solution maximale. On note $S := \{(y_i, J_i)$ solution du problème de Cauchy$\}$. On prend $J = \cup J_i$ et on définit, pour $t \in J$, $y(t) = y_i(t)$ si $t \in J_i$. A nouveau il n'y a pas d'ambiguité dans la définition de y puisque $y_i(t) = y_j(t)$ si $t \in J_i \cap J_j$. On vérifie alors facilement que (y, J) est une solution et qu'elle est maximale par définition de J. □

Remarque 4.5. *On peut montrer que l'intervalle de défintion J d'une solution maximale est toujours ouvert, i.e. $J =]T_-, T_+[$ avec éventuellement $T_- = -\infty$ et/ou $T_+ = +\infty$. Par ailleurs, l'unicité permet d'affirmer que deux solutions maximales distinctes de l'équation différentielle ne peuvent jamais prendre la même valeur, i.e.*

$$\exists t \in I,\ y_1(t) \neq y_2(t) \quad \Rightarrow \quad \forall t \in I,\ y_1(t) \neq y_2(t).$$

En effet, s'il existe $t_0 \in I$ tel que $y_1(t_0) = y_2(t_0) =: y_0$, alors y_1 et y_2 sont toutes les deux solutions du problème de Cauchy avec condition initiale $y(t_0) = y_0$ donc par unicité elles coincident. Pour résoudre l'équation $y' = y^2$ on peut par exemple affirmer que si une solution y s'annule alors c'est la fonction nulle. Pour chercher les solutions non-nulles on peut alors légitimement écrire $\frac{y'}{y^2} = 1$ puis intégrer des deux côtés.

La démonstration du Théorème de Cauchy-Lipschitz n'est pas évidente, en particulier à cause du problème de l'intervalle de définition de la solution qui peut être plus petit que l'intervalle de définition de la fonction f, voir l'Exemple 4.1. On va démontrer le cas particulier ci-dessous. L'esprit de la démonstration est le même que dans le cas général mais évite le problème de l'intervalle de définition.

Théorème 4.7 (Cauchy linéaire)**.** *Soient $A : I \to M_d(\mathbb{R})$ et $B : I \to \mathbb{R}^d$ des fonctions continues. Pour tout $(t_0, y_0) \in I \times \mathbb{R}^d$ le problème de Cauchy* $\begin{cases} y'(t) = A(t)y(t) + B(t), \\ y(t_0) = y_0, \end{cases}$ *admet une unique solution globale, i.e. définie sur I tout entier.*

Démonstration. On commence par remarquer qu'une fonction dérivable y est solution du problème de Cauchy si et seulement si elle vérifie, pour tout $t \in I$,

$$y(t) = y_0 + \int_{t_0}^{t} (A(s)y(s) + B(s))\,\mathrm{d}s. \tag{4.2}$$

Existence : On définit par récurrence la suite de fonctions $(y_n)_n$ par

$$y_0(t) = y_0, \qquad y_{n+1}(t) = y_0 + \int_{t_0}^{t} (A(s)y_n(s) + B(s))\,\mathrm{d}s. \tag{4.3}$$

On va montrer que la suite $(y_n)_n$ converge vers une fonction y solution de (4.2).

On commence par montrer la convergence uniforme sur tout segment $J \subset I$ contenant t_0. Soit J un tel segment. La fonction A étant continue, elle est bornée sur J. On note $M_A := \max_{t\in J} \|A(t)\|$ (ici $M_d(\mathbb{R})$ est muni de la norme matricielle associée à la norme choisie sur $\mathbb{R}^d$) et $M_B := \max_{t\in J} \|B(t)\|$. On a pour tout $n \geq 1$ et tout $t \in J$, $t \geq t_0$,

$$\begin{aligned} \|y_{n+1}(t) - y_n(t)\| &= \left\| \int_{t_0}^{t} A(s)(y_n(s) - y_{n-1}(s))\,\mathrm{d}s \right\| \\ &\leq M_A \int_{t_0}^{t} \|y_n(s) - y_{n-1}(s)\|\,\mathrm{d}s, \end{aligned} \tag{4.4}$$

tandis que

$$\|y_1(t) - y_0(t)\| = \left\| \int_{t_0}^{t} (A(s)y_0 + B(s))\,\mathrm{d}s \right\| \leq (M_A\|y_0\| + M_B)(t - t_0). \tag{4.5}$$

En raisonnant par récurrence on montre alors que, pour tout $n \in \mathbb{N}$,

$$\|y_{n+1}(t) - y_n(t)\| \leq (M_A\|y_0\| + M_B)\frac{M_A^n (t - t_0)^{n+1}}{(n+1)!} \tag{4.6}$$

En effet, (4.5) montre que le résultat est vrai pour $n = 0$. Supposons le résultat vrai au rang n. D'après (4.4) on a alors

$$\begin{aligned} \|y_{n+2}(t) - y_{n+1}(t)\| &\leq M_A \int_{t_0}^{t} (M_A\|y_0\| + M_B)\frac{M_A^n (s - t_0)^{n+1}}{(n+1)!}\,\mathrm{d}s \\ &\leq (M_A\|y_0\| + M_B)\frac{M_A^{n+1} (t - t_0)^{n+2}}{(n+2)!}. \end{aligned}$$

On montre de la même manière que si $t \in J$, $t < t_0$, alors

$$\|y_{n+1}(t) - y_n(t)\| \leq (M_A\|y_0\| + M_B)\frac{M_A^n |t - t_0|^{n+1}}{(n+1)!}.$$

Finalement, pour tout $t \in J$ et tout entier n on a

$$\|y_{n+1}(t) - y_n(t)\| \leq (M_A\|y_0\| + M_B)\frac{M_A^n|J|^{n+1}}{(n+1)!},$$

où $|J|$ est la longueur de J et où on a utilisé $|t - t_0| \leq |J|$. On en déduit que la série de fonction $\sum(y_{n+1} - y_n)$ converge normalement sur J et donc uniformément. Comme $y_n = y_0 + \sum_{k=0}^{n-1}(y_{k+1} - y_k)$, cela prouve la convergence uniforme de la suite $(y_n)_n$ sur J. On note y sa limite, et en passant à la limite dans (4.3) (on peut puisque la convergence est uniforme), la fonction y vérifie bien (4.2) pour tout $t \in J$. Le résultat étant vrai pour segment $J \subset I$ il est aussi vrai pour tout $t \in I$.

Unicité : Soient y_1 et y_2 deux solutions. On note $z = y_1 - y_2$. La fonction z vérifie alors $z(t_0) = 0$ et pour tout $t \in I$

$$z(t) = \int_{t_0}^{t} A(s)z(s)\,\mathrm{d}s. \tag{4.7}$$

On montre que z est nulle sur tout segment $J \subset I$. Soit J un tel segment. On note $M_z := \max_{t\in J}\|z(t)\|$ et $M_A := \max_{t\in J}\|A(t)\|$. On montre alors par récurrence en utilisant (4.7) que pour tout $n \in \mathbb{N}$ et tout $t \in J$ on a

$$\|z(t)\| \leq M_z M_A^n \frac{|t - t_0|^n}{n!},$$

et donc $M_z \leq M_z M_A^n \frac{|J|^n}{n!}$. En faisant tendre n vers l'infini on en déduit que $M_z = 0$ et donc z est nulle sur J. Ceci étant vrai pour tout segment $J \subset I$ on obtient que z est nulle sur I et donc $y_1 = y_2$, ce qui prouve l'unicité. □

Du point de vue de la recherche de solutions approchées, le Théorème de Cauchy-Lipschitz nous permet de savoir ce qu'on approche : pourvu que la fonction f soit continue et localement Lipschitzienne par rapport à la seconde variable (par exemple si f est C^1) il y a une unique solution. On est par contre confronté à la difficulté de connaitre l'intervalle J sur lequel elle est définie. Il serait donc utile d'avoir des résultats permettant d'affirmer que celui-ci est l'intervalle I de départ.

Théorème 4.8. *Soit $f : I \times \Omega \to \mathbb{R}^d$ continue, localement Lipschitzienne par rapport à la seconde variable, $(t_0, y_0) \in I \times \Omega$ et y l'unique solution maximale du problème de Cauchy. On note $]T_-, T_+[$ l'intervalle sur lequel elle est définie. Si $T_+ < \sup I$, resp. $T_- > \inf I$, alors pour tout compact $K \subset \Omega$ il existe $\delta > 0$ tel que pour tout $t \in]T_+ - \delta, T_+[$, resp. $t \in]T_-, T_- + \delta[$, on a $y(t) \notin K$. En particulier si $\Omega = \mathbb{R}^d$ alors $\lim_{t\to T_+}\|y(t)\| = +\infty$, resp. $\lim_{t\to T_-}\|y(t)\| = +\infty$.*

Le théorème ci-dessus affirme que si l'intervalle de définition d'une solution maximale n'est pas I tout entier la solution y doit approcher le bord de l'ensemble de définition Ω de la fonction f (sortie de tout compact). Dans le cas où $\Omega = \mathbb{R}^d$ on parle aussi d'explosion en temps fini.

Exemple 4.2. *On reprend la fonction $f(y) = y^2$ de l'Exemple 4.1. On considère le problème de Cauchy associé à la condition initiale $y(0) = y_0 > 0$. La solution est alors $y(t) = \frac{1}{y_0^{-1} - t}$ définie sur $J =]-\infty, y_0^{-1}[$. On a $\sup J = y_0^{-1} < \sup I = +\infty$ et $\Omega = \mathbb{R}$. On peut vérifier que $\lim_{t \to y_0^{-1}} y(t) = +\infty$.*

Démonstration. On traite le cas $T_+ < \sup I$ et on montre la version plus faible suivante : pour tout compact $K \subset \Omega$ il existe $t_K \in [t_0, T_+[$ tel que $y(t_K) \notin K$, autrement dit y sort de tout compact K au moins une fois (le théorème affirme qu'à un moment donné elle n'y revient plus). On raisonne par l'absurde. Supposons qu'il existe K tel que pour tout $t \in [t_0, T_+[$ on ait $y(t) \in K$. La fonction f est continue sur le compact $[t_0, T_+] \times K$ donc y est bornée. La fonction $y'(t) = f(t, y(t))$ est donc bornée sur $[t_0, T_+[$. D'après le théorème des accroissements finis, pour tous $s, t \in [t_0, T_+[$ on a

$$\|y(t) - y(s)\| \leq \sup_{[t_0, T_+[} \|y'(u)\| \times |t - s|,$$

et le critère de Cauchy permet d'affirmer que la fonction y admet une limite y_+ lorsque $t \to T_+$.

On considère maintenant le problème de Cauchy $\begin{cases} y'(t) = f(t, y(t)), \\ y(T_+) = y_+. \end{cases}$ Il admet une unique solution maximale $(\tilde{y}, \tilde{J})$. On note alors $J_{\text{ext}} =]T_-, T_+[\cup \tilde{J}$ et on définit sur J_{ext} la fonction

$$y_{\text{ext}}(t) = \begin{cases} y(t) & \text{si } t < T_+ \\ \tilde{y}(t) & \text{si } t \geq T_+ \end{cases}.$$

On vérifie alors que y_{ext} est une fonction dérivable sur J_{ext} qui est un prolongement strict de y dans le sens où $J \not\subset J_{\text{ext}}$. Cela contredit la maximalité de y.

Si $\Omega = \mathbb{R}^d$ il suffit de prendre $K = B_0(R)$ la boule de centre 0 et de rayon R. On a alors : pour tout R il existe $\delta > 0$ tel que $\|y(t)\| > R$ pour tout $t > T_+ - \delta$. C'est précisément la définition de $\lim_{t \to T_+} \|y(t)\| = +\infty$.

Remarque : la version plus faible démontrée ci-dessus s'écrirait $\limsup_{t \to T_+} \|y(t)\| = +\infty$. □

On va utiliser le Théorème 4.8 pour donner deux conditions suffisantes pour que la solution maximale soit bien définie sur I tout entier.

Proposition 4.9. *Soit $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ continue, localement Lipschitzienne par rapport à la seconde variable. Si f est bornée alors pour tout $(t_0, y_0) \in I \times \mathbb{R}^d$ le problème de Cauchy $\begin{cases} y'(t) = f(t, y(t)), \\ y(t_0) = y_0, \end{cases}$ admet une unique solution globale, i.e. définie sur I tout entier.*

Démonstration. Si f est bornée alors pour tout $t \in J$ (intervalle de définition de la solution maximale) on a

$$\|y(t)\| = \left\| y_0 + \int_{t_0}^{t} f(s, y(s)) \, \mathrm{d}s \right\| \leq \|y_0\| + \|f\|_\infty |t - t_0|.$$

Cela prouve que si $\sup J < \sup I$ alors y est bornée au voisinage de $\sup J$, contredisant le Théorème 4.8. □

Proposition 4.10. *Soit $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ continue, globalement Lipschitzienne par rapport à la seconde variable dans le sens où il existe $k : I \to \mathbb{R}_+$ continue telle que pour tout $t \in I$ et $y_1, y_2 \in \mathbb{R}^d$ on ait*

$$\|f(t, y_1) - f(t, y_2)\| \leq k(t)\|y_1 - y_2\|,$$

(f est globalement Lipschitzienne dans le sens où la constante k peut dépendre de t mais pas de y). Alors pour tout $(t_0, y_0) \in I \times \mathbb{R}^d$ le problème de Cauchy $\begin{cases} y'(t) = f(t, y(t)), \\ y(t_0) = y_0, \end{cases}$ *admet une unique solution globale*

La preuve utilise le résultat suivant qui est très utile en soi et que l'on rencontre souvent.

Lemme 4.11 (Gronwall)**.** *Soient $\varphi, \psi : [a, b[\to \mathbb{R}$ et $\alpha \in \mathbb{R}_+$. On suppose que ψ est positive et pour tout $t \geq a$*

$$0 \leq \varphi(t) \leq \alpha + \int_a^t \psi(s)\varphi(s)\,\mathrm{d}s.$$

Alors pour tout $t \in [a, b[$ on a $\varphi(t) \leq \alpha \exp\left(\int_a^t \psi(s)\,\mathrm{d}s\right)$.

Démonstration. On note $f(t) = \alpha + \int_a^t \psi(s)\varphi(s)\,\mathrm{d}s$. On a $f'(t) = \psi(t)\varphi(t)$ et puisque ψ est positive on obtient, pour tout t,

$$f'(t) \leq \psi(t)f(t).$$

On définit alors $g(t) = f(t)\exp\left(-\int_a^t \psi(s)\,\mathrm{d}s\right)$. On a

$$g'(t) = (f'(t) - \psi(t)f(t))\exp\left(-\int_a^t \psi(s)\,\mathrm{d}s\right) \leq 0.$$

La fonction g est donc décroissante. Comme $g(a) = f(a) = \alpha$ on en déduit que

$$g(t) = f(t)\exp\left(-\int_a^t \psi(s)\,\mathrm{d}s\right) \leq \alpha \quad \Longleftrightarrow \quad f(t) \leq \alpha \exp\left(\int_a^t \psi(s)\,\mathrm{d}s\right).$$

Puisque $\varphi(t) \leq f(t)$ cela prouve le résultat. □

Remarque 4.6. *Un cas particulier important est lorsque ψ est constante égale à $\beta > 0$. L'hypothèse s'écrit alors $\varphi(t) \leq \alpha + \beta \int_a^t \varphi(s)\,\mathrm{d}s$ et la conclusion $\varphi(t) \leq \alpha \mathrm{e}^{\beta(t-a)}$.*

Démonstration de la Proposition. Soit (y, J) la solution maximale du problème de Cauchy. On a pour tout t

$$y(t) = y(t_0) + \int_{t_0}^t f(s, y(s))\,\mathrm{d}s,$$

et donc pour $t > t_0$

$$\begin{aligned}\|y(t)\| &\leq \|y_0\| + \int_{t_0}^{t} \|f(s,y(s)) - f(s,0) + f(s,0)\| \, \mathrm{d}s \\ &\leq \|y_0\| + \int_{t_0}^{t} \|f(s,0)\| \, \mathrm{d}s + \int_{t_0}^{t} \|f(s,y(s)) - f(s,0)\| \, \mathrm{d}s \\ &\leq \|y_0\| + \int_{t_0}^{t} \|f(s,0)\| \, \mathrm{d}s + \int_{t_0}^{t} k(s)\|y(s)\| \, \mathrm{d}s.\end{aligned}$$

Supposons que $T_+ := \sup J < \sup I$. Pour tout $t \in [t_0, T_+[$ on a alors

$$\|y(t)\| \leq \underbrace{\|y_0\| + \int_{t_0}^{T_+} \|f(s,0)\| \, \mathrm{d}s}_{=\alpha} + \int_{t_0}^{t} k(s)\|y(s)\| \, \mathrm{d}s,$$

et donc d'après le Lemme de Gronwall on a

$$\|y(t)\| \leq \alpha \exp\left(\int_{t_0}^{t} k(s) \, \mathrm{d}s\right).$$

Cela montre que y reste bornée lorsque $t \to T_+$ en contradiction avec le Théorème 4.8. □

4.2 Le schéma d'Euler

Dans la suite du chapitre on va s'intéresser à la résolution approchée du problème de Cauchy

$$\begin{cases} y'(t) = f(t, y(t)), \\ y(t_0) = y_0. \end{cases} \tag{4.8}$$

On cherchera à résoudre l'équation sur un intervalle du type $[t_0, t_0 + T]$. C'est ici que les résultats d'existence globale de solution sont utiles, ils permettent de savoir a priori sur quel intervalle on pourra chercher une solution approchée. Dans toute la suite on supposera que la fonction f est définie et continue sur $[t_0, t_0 + T] \times \mathbb{R}^d$ et est globalement lipschitzienne en y de constante de Lipschitz L (la Proposition 4.10 assure ainsi l'existence globale de la solution).

On se donne une suite de points $t_0 < t_1 < \cdots < t_N = T$ et on va essayer d'obtenir des valeurs approchées de la solution en chacun de ces points, i.e. des valeurs y_n proches des $y(t_n)$. Pour simplifier on ne se placera que dans le cas d'une répartition uniforme des t_n : $t_n = t_0 + nh$ où $h := \dfrac{T}{N}$ est appelé le *pas* (on a alors $t_{n+1} - t_n = h$ pour tout n).

La méthode d'Euler consiste à approcher la dérivée de y au point t_n par le taux d'accroissement $\dfrac{y(t_{n+1}) - y(t_n)}{t_{n+1} - t_n}$. L'équation différentielle est alors remplacée, pour tout $n = 0, \ldots, N-1$, par

$$\frac{y(t_{n+1}) - y(t_n)}{t_{n+1} - t_n} \simeq f(t_n, y(t_n)) \qquad \Longleftrightarrow \qquad y(t_{n+1}) \simeq y(t_n) + hf(t, y(t_n)).$$

La méthode d'Euler consiste donc à définir par récurrence

$$y_{n+1} = y_n + hf(t_n, y_n), \tag{4.9}$$

y_0 étant donné (typiquement la donnée initiale du problème de Cauchy).

Il y a une deuxième façon d'appréhender le schéma d'Euler. Elle consiste à d'abord intégrer l'équation différentielle en écrivant

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(s, y(s))\,\mathrm{d}s,$$

puis à remplacer l'intégrale par une formule d'intégration numérique. Si on choisit la méthode des rectangles à gauche, on approche $\displaystyle\int_{t_n}^{t_{n+1}} f(s, y(s))\,\mathrm{d}s$ par $(t_{n+1} - t_n)f(t_n, y(t_n))$ et on retrouve

$$y(t_{n+1}) \simeq y(t_n) + hf(t, y(t_n)).$$

Cette méthode est dite à *un pas*, le calcul de y_{n+1} ne fait intervenir que la valeur de y_n et pas les valeurs antérieures, et explicite (une fois connu y_n on calcule directement y_{n+1}). Si on utilisait la méthode des rectangles à droite dans l'approximation de $\displaystyle\int_{t_n}^{t_{n+1}} f(s, y(s))\,\mathrm{d}s$ on obtiendrait le schéma

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1}).$$

C'est la méthode d'Euler dite implicite. Pour obtenir y_{n+1} à partir de y_n il faut pouvoir "inverser" la fonction $y \mapsto y - hf(t_{n+1}, y)$. On peut montrer que c'est le cas dès que $hL < 1$.

On voudrait maintenant estimer l'erreur commise par la méthode d'Euler, c'est-à-dire majorer (en norme) la quantité

$$e_n = y(t_n) - y_n,$$

appelée *erreur de discrétisation*, et en particulier voir si cette erreur tend vers 0 lorsqu'on diminue la taille du pas (convergence du schéma). On va supposer que le point de départ y_0 est bien la donnée de Cauchy et donc que $e_0 = 0$. On notera également

$$\epsilon_n := y(t_{n+1}) - y(t_n) - hf(t_n, y(t_n)). \tag{4.10}$$

Cette erreur est appelée *erreur de consistance*. C'est l'erreur commise au point t_{n+1} si la valeur approchée au point t_n était exacte, i.e. si $y_n = y(t_n)$.

Remarque 4.7. *La définition de l'erreur de consistance n'est pas la même dans tous les ouvrages. On peut également rencontrer* $\tilde{\epsilon}_n = \dfrac{y(t_{n+1}) - y(t_n)}{h} - f(t_n, y_n)$, *i.e.* $\tilde{\epsilon}_n = \dfrac{\epsilon_n}{h}$.

On écrit alors

$$\begin{aligned} e_{n+1} &= y(t_{n+1}) - y_{n+1} \\ &= [\epsilon_n + y(t_n) + hf(t_n, y(t_n))] - [y_n + hf(t_n, y_n)] \\ &= \epsilon_n + e_n + h\left[f(t_n, y(t_n)) - f(t_n, y_n)\right]. \end{aligned}$$

Comme f est L-Lipschitzienne par rapport à y, on en déduit que pour tout $n = 0, \dots, N-1$

$$\|e_{n+1}\| \leq (1 + hL)\|e_n\| + \|\epsilon_n\|. \tag{4.11}$$

On va ensuite utiliser une version discrète du Lemme de Gronwall (Lemme 4.11).

Lemme 4.12 (Lemme de Gronwall discret)**.** *Soient $(a_n)_n$ et $(b_n)_n$ des nombres positifs et $C \geq 0$ tels que pour tout n*

$$a_{n+1} \leq (1 + C)a_n + b_n.$$

Alors on a, pour tout n,

$$a_n \leq (1 + C)^n a_0 + \sum_{j=0}^{n-1} b_j (1 + C)^{n-1-j}.$$

Démonstration. On montre le résultat par récurrence sur n. Pour $n = 0$ le résultat est évident.

Soit $n \in \mathbb{N}$, supposons le résultat vrai au rang n. On a alors

$$\begin{aligned} a_{n+1} &\leq (1 + C)\left((1 + C)^n a_0 + \sum_{j=0}^{n-1} b_j (1 + C)n - 1 - j\right) + b_{n+1} \\ &\leq (1 + C)^{n+1} a_0 + \sum_{j=0}^{n-1} b_j (1 + C)^{n-j} + b_{n+1} \\ &= (1 + C)^{n+1} a_0 + \sum_{j=0}^{n} b_j (1 + C)^{n-j}. \end{aligned}$$

□

Remarque 4.8. *On rencontre souvent ce Lemme avec comme conclusion*

$$a_n \leq \mathrm{e}^{nC} a_0 + \sum_{j=0}^{n-1} b_j \mathrm{e}^{C(n-1-j)}.$$

Ca découle directement de la version ci-dessus en utilisant l'inégalité $1 + C \leq \mathrm{e}^C$.

En utilisant (4.11) et le lemme, on en déduit que pour tout $n = 0, \dots, N-1$ on a (on rappelle que $e_0 = 0$)

$$\|e_n\| \leq \sum_{j=0}^{n-1} (1 + hL)^{n-1-j} \|\epsilon_j\|. \tag{4.12}$$

Pour voir si $\|e_n\| \to 0$ il faut pouvoir estimer les erreurs de consistance $\|\epsilon_j\|$. Afin de simplifier les calculs on suppose que f est de classe C^1. Toute solution y de l'équation différentielle est alors de classe C^2. La formule de Taylor avec reste intégral à l'ordre 2 donne

$$y(t_{n+1}) = y(t_n) + hy'(t_n) + \int_{t_n}^{t_{n+1}} (t_{n+1} - s)y''(s)\,\mathrm{d}s,$$

et comme y est solution de l'équation différentielle on a $y'(t_n) = f(t_n, y(t_n))$ et donc

$$\epsilon_n = y(t_{n+1}) - y(t_n) - hf(t_n, y(t_n)) = \int_{t_n}^{t_{n+1}} (t_{n+1} - s)y''(s)\,\mathrm{d}s.$$

On a ainsi, pour tout $n \in \mathbb{N}$,

$$\|\epsilon_n\| \leq \sup_{t\in[t_0,t_0+T]} \|y''(t)\| \times \int_{t_n}^{t_{n+1}} (t_{n+1} - s)\,\mathrm{d}s = \frac{h^2}{2} \sup_{t\in[t_0,t_0+T]} \|y''(t)\|. \tag{4.13}$$

et donc

$$\|e_n\| \leq \frac{h^2}{2} \sup_{t\in[t_0,t_0+T]} \|y''(t)\| \times \sum_{j=0}^{n-1} (1+hL)^{n-1-j} = \frac{h^2}{2} \sup_{t\in[t_0,t_0+T]} \|y''(t)\| \frac{(1+hL)^n - 1}{hL}.$$

Pour tout n on a $nh \leq Nh = T$ et par ailleurs $1 + hL \leq \mathrm{e}^{hL}$, d'où

$$(1+hL)^n \leq \mathrm{e}^{nhL} \leq \mathrm{e}^{TL}.$$

Ainsi, pour tout n on a

$$\|e_n\| \leq h\frac{\mathrm{e}^{TL} - 1}{2L} \sup_{t\in[t_0,t_0+T]} \|y''(t)\|, \tag{4.14}$$

et donc $\sup_{n=0,\ldots,N-1} \|e_n\| \to 0$ lorsque $h \to 0$. On dit que la méthode est convergente. On a même une estimation de l'erreur en $O(h)$. On verra par la suite que l'on ne peut pas faire mieux, on dit que la méthode est d'ordre 1 (voir la Définition 4.20).

Si on souhaite avoir une fonction qui soit solution approchée définie sur $[t_0, t_0 + T]$ on peut alors définir la fonction y_h affine par morceaux telle que $y_h(t_n) = y_n$ pour tout $n = 0, \ldots, N$. On montre alors que $(y_h)_h$ converge uniformément sur $[t_0, t_0 + T]$ vers la solution de (4.8), c'est-à-dire

$$\lim_{h\to 0} \|y - y_h\|_\infty = 0.$$

4.3 Méthodes à un pas

4.3.1 Définition et exemples

On reste, pour simplifier, dans le cas d'un pas $h = \frac{T}{N}$ constant. Les méthodes à un pas [a] sont des schémas numériques de la forme

$$y_{n+1} = y_n + h\Phi(t_n, y_n, h), \tag{4.15}$$

où $\Phi : [t_0, t_0 + T] \times \mathbb{R} \times [0, h_{\max}]$ est une fonction continue. La méthode d'Euler en est un cas particulier en prenant $\Phi(t, y, h) = f(t, y)$. En voici deux autres exemples :

a. La terminologie vient du fait que y_{n+1} se calcule uniquement à partir de y_n et pas des valeurs antérieures, i.e. des y_k pour $k < n$.

1. *Méthodes de Taylor.* Dans la méthode d'Euler on a approché la dérivée $y'(t_n)$ par le taux d'accroissement $\frac{1}{h}(y(t_{n+1}) - y(t_n))$, ce qui revient à négliger le reste dans le développement de Taylor à l'ordre 1 :

$$y(t_{n+1}) = y(t_n) + hy'(t_n) + o(h) \simeq y(t_n) + hy'(t_n) = y(t_n) + hf(t_n, y(t_n)).$$

On peut choisir d'effectuer un développement à un ordre plus élevé. Par exemple à l'ordre 2 on écrit

$$y(t_{n+1}) = y(t_n) + hy'(t_n) + \frac{h^2}{2}y''(t_n) + o(h^2).$$

En dérivant l'équation différentielle on a

$$y''(t) = \frac{\partial f}{\partial t}(t, y(t)) + D_y f(t, y(t))(y'(t)) = \frac{\partial f}{\partial t}(t, y(t)) + D_y f(t, y(t))(f(t, y(t))),$$

où $D_y f$ note la différentielle partielle de f par rapport à $y \in \mathbb{R}^d$. Si $d = 1$ l'équation ci-dessus s'écrit juste

$$y''(t) = \frac{\partial f}{\partial t}(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t)) \times f(t, y(t)).$$

En négligeant le $o(h^2)$ dans le développement de Taylor on obtient alors, pour $d = 1$, le schéma

$$y_{n+1} = y_n + hf(t_n, y_n) + \frac{h^2}{2}\left(\frac{\partial f}{\partial t}(t_n, y_n) + \frac{\partial f}{\partial y}(t_n, y_n) \times f(t_n, y_n)\right).$$

C'est bien un schéma de la forme (4.15) avec

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

2. *Méthode du point milieu.* Au lieu de la méthode des rectangles à gauche on utilise celle du point milieu pour approcher $\displaystyle\int_{t_n}^{t_{n+1}} f(s, y(s))\,\mathrm{d}s$:

$$\int_{t_n}^{t_{n+1}} f(s, y(s))\,\mathrm{d}s \simeq hf\left(t_n + \frac{h}{2}, y\left(t_n + \frac{h}{2}\right)\right).$$

Cependant on ne connait pas $y\left(t_n + \frac{h}{2}\right)$, on utilise pour cela la méthode d'Euler. Au final le schéma s'écrit

$$y_{n+1} = y_n + hf\left(t_n + \frac{h}{2}, y_n + \frac{h}{2}f(t_n, y_n)\right).$$

La fonction Φ est ici $\Phi(t, y, h) = f\left(t + \frac{h}{2}, y + \frac{h}{2}f(t, y)\right)$.

Dans toute la suite de la section on considère donc un schéma numérique de la forme (4.15) où Φ est une fonction donnée. On considèrera également le cas où la condition initiale du schéma n'est pas nécessairement la donnée initiale $y(t_0)$ du problème de Cauchy. On la notera $y_{0,h}$ pour indiquer qu'elle peut dépendre de la taille du pas (celle-ci peut être obtenue via une mesure, un autre calcul approché, etc).

4.3.2 Consistance, stabilité et convergence d'un schéma

La première notion est celle de schéma consistant. Elle relie l'équation différentielle au schéma numérique considéré : ce dernier approche-t-il correctement l'équation lorsque la taille du pas tend vers 0 ? On verra dans la Section 4.3.3 un critère simple pour vérifier si un schéma est consistant ou non.

Définition 4.13. *On appelle erreurs de consistance d'un schéma, associées à une solution y de l'équation différentielle, les quantités*

$$\epsilon_n := y(t_{n+1}) - y(t_n) - h\Phi(t_n, y(t_n), h).$$

C'est la même quantité que dans (4.10), on y a juste remplacé f par Φ : ϵ_n est l'erreur commise au point t_{n+1} si la valeur approchée au point t_n est exacte.

Remarque 4.9. *Selon les ouvrages l'erreur de consistance est parfois définie par*

$$\tilde{\epsilon}_n = \frac{y(t_{n+1}) - y(t_n)}{h} - \Phi(t_n, y(t_n), h), \text{ i.e. } \epsilon_n = h\tilde{\epsilon}_n.$$

Définition 4.14. *On dit que le schéma est consistant pour l'équation différentielle si pour* toute *solution y de celle-ci on a*

$$\lim_{h\to 0} \sum_{n=0}^{N-1} \|\epsilon_n\| = 0, \qquad h = \frac{T}{N}.$$

Remarque 4.10. *Lorsque l'erreur de consistance est définie comme dans la Remarque 4.9 la notion de schéma consistant est alors remplacée par : pour toute solution y de l'équation différentielle*

$$\lim_{h\to 0} \sup_{n=0,\dots,N-1} \|\tilde{\epsilon}_n\| = 0, \qquad h = \frac{T}{N}.$$

Il est facile de voir que cette notion est a priori plus forte celle de la Définition 4.14. Cela découle de la majoration

$$\sum_{n=0}^{N-1} \|\epsilon_n\| = \frac{1}{h} \sum_{n=0}^{N-1} \|\tilde{\epsilon}_n\| \leq hN \sup_{n=0,\dots,N-1} \|\tilde{\epsilon}_n\| = T \sup_{n=0,\dots,N-1} \|\tilde{\epsilon}_n\|.$$

On peut en fait montrer que ces deux définitions sont équivalentes.

La deuxième notion est intrinsèque au schéma, c'est celle de stabilité. Au cours du calcul les erreurs vont s'accumuler (erreur d'approximation, d'arrondis, etc.) et la notion de stabilité permet de contrôler ce cumul d'erreurs en fonctions des erreurs individuelles.

Définition 4.15. *Un schéma est dit stable s'il existe une constante S, indépendante de h et appelée constante de stabilité, telle que pour tout h et toutes suites $(y_n)_n$, $(\tilde{y}_n)_n$ et $(\delta_n)_n$ telles que*

$$\begin{aligned} y_{n+1} &= y_n + h\Phi(t_n, y_n, h), \qquad \forall n = 0, \dots, N-1, \\ \tilde{y}_{n+1} &= \tilde{y}_n + h\Phi(t_n, \tilde{y}_n, h) + \delta_n, \quad \forall n = 0, \dots, N-1, \end{aligned}$$

on a

$$\sup_{n=0,\dots,N-1} \|y_n - \tilde{y}_n\| \leq S \left(\|y_0 - \tilde{y}_0\| + \sum_{n=0}^{N-1} \|\delta_n\| \right).$$

On définit enfin la notion de schéma convergent.

Définition 4.16. *Un schéma est dit convergent si pour tout y_0 et toute suite $(y_{0,h})_h$ telle que $y_{0,h} \to y_0$ lorsque $h \to 0$ on a*

$$\lim_{h\to 0} \sup_{n=0,\dots,N-1} \|y(t_n) - y_n\| = 0,$$

où y est la solution du problème de Cauchy avec condition initiale y_0 et $(y_n)_n$ la suite de valeurs approchées obtenues par le schéma à partir de $y_{0,h}$.

On a alors le résultat suivant

Théorème 4.17. *Si un schéma est consistant et stable alors il est convergent.*

Démonstration. Par définition de l'erreur de consistance, la suite $(y(t_n))_n$ vérifie

$$y(t_{n+1}) = y(t_n) + h\Phi(t_n, y(t_n), h) + \epsilon_n.$$

Comme le schéma est stable on a

$$\sup_{n=0,\dots,N-1} \|y(t_n) - y_n\| \leq S \left(\|y_{0,h} - y_0\| + \sum_{n=0}^{N-1} \|\epsilon_n\| \right). \tag{4.16}$$

Par hypothèse le premier terme, $\|y_{0,h} - y_0\|$, tend vers 0 et comme le schéma est consistant le second également. □

Exemple 4.3. *Le schéma d'Euler est consistant et stable. En effet, si $y_{0,h} \neq y_0$ l'équation (4.12) devient*

$$\|y(t_n) - y_n\| \leq (1 + hL)^n \|y_0 - y_{0,h}\| + \sum_{j=0}^{n-1} (1 + hL)^{n-1-j} \|\epsilon_j\|.$$

Pour tout $n \leq N$ et tout $j = 0, \dots, n-1$ on a $nh \leq T$ et $(n-1-j)h \leq T$ ce qui prouve que le schéma est stable avec $S = \mathrm{e}^{TL}$.

Par ailleurs, l'équation (4.13) entraine

$$\sum_{n=0}^{N-1} \|\epsilon_n\| \leq N \times \frac{h^2}{2} \sup_{t\in[t_0,t_0+T]} \|y''(t)\| = \frac{hT}{2} \sup_{t\in[t_0,t_0+T]} \|y''(t)\|,$$

ce qui prouve que la méthode est consistante.

4.3.3 Critères de consistance et stabilité

Afin de montrer la convergence d'un schéma on utilise en général, voir tout le temps, les notions de consistance et stabilité. Il est donc utile d'avoir des critères simples pour montrer qu'un schéma est consistant et qu'il est stable.

Théorème 4.18. *Un schéma à un pas est consistant si et seulement si*

$$\Phi(t, y, 0) = f(t, y), \qquad \forall t \in [0, T],\ \forall y \in \mathbb{R}^d.$$

Remarque 4.11. *L'idée derrière ce théorème est la suivante. Le schéma peut se réécrire*

$$\frac{y_{n+1} - y_n}{h} = \Phi(t_n, y_n, h).$$

Lorsque h tend vers 0 le membre de gauche se rapproche d'une dérivée y' et comme Φ est continue celui de droite se rapproche de $\Phi(t, y, 0)$. Pour "coller" à l'équation différentielle il faut donc que cette dernière quantité soit précisément $f(t, y)$.

Démonstration. On fait la preuve dans le cas $d = 1$.

Soit y une solution de l'équation différentielle. D'après le théorème des accroissements finis appliqué sur l'intervalle $[t_n, t_{n+1}]$ (c'est ici qu'on utilise $d = 1$), il existe $c_n \in]t_n, t_{n+1}[$ tel que

$$\begin{aligned}
\epsilon_n &= hy'(c_n) - h\Phi(t_n, y(t_n), h) \\
&= hf(c_n, y(c_n)) - h\Phi(t_n, y(t_n), h) \\
&= \underbrace{h\left(f(c_n, y(c_n)) - \Phi(c_n, y(c_n), 0)\right)}_{=:a_n} + \underbrace{h\left(\Phi(c_n, y(c_n), 0) - \Phi(t_n, y(t_n), h)\right)}_{=:b_n}.
\end{aligned}$$

Toute solution y de l'équation différentielle est continue sur $[0, T]$ donc y est uniformément continue et bornée : il existe M tel que $|y(t)| \leq M$. De plus la fonction Φ est continue sur $[0, T] \times \bar{B}(0, M) \times [0, h_{\max}]$ qui est compact donc elle y est uniformément continue. Soit $\varepsilon > 0$. L'uniforme continuité de Φ entraine que

$$\exists \delta > 0,\ (|t - s| < \delta,\ |y - z| < \delta,\ 0 \leq h \leq \delta) \implies |\Phi(t, y, 0) - \Phi(s, z, h)| < \varepsilon.$$

Celle de y entraine que

$$\exists \zeta > 0,\ |t - s| < \zeta \implies |y(s) - y(t)| < \delta.$$

Puisque $|t_n - c_n| < h$, on rappelle que $c_n \in]t_n, t_{n+1}[$, en prenant $\eta = \min(\delta, \zeta)$ on obtient

$$0 \leq h \leq \eta \implies |\Phi(c_n, y(c_n), 0) - \Phi(t_n, y(t_n), h)| < \varepsilon, \quad \forall n = 0, \ldots, N - 1.$$

En sommant sur n on a ainsi montré

$$\forall \varepsilon > 0,\ \exists \eta > 0,\ 0 \leq h \leq \eta \implies \sum_{n=0}^{N-1} |b_n| < Nh\varepsilon = T\varepsilon.$$

Autrement dit $\lim_{h\to 0}\sum_{n=0}^{N-1}|b_n| = 0$. On en déduit que le schéma est consistant si et seulement si, pour toute solution y,

$$\lim_{h\to 0}\sum_{n=0}^{N-1}|a_n| = 0, \quad h = \frac{T}{N}.$$

Or

$$\sum_{n=0}^{N-1}|a_n| = h\sum_{n=0}^{N-1}\left|f(c_n, y(c_n)) - \Phi(c_n, y(c_n), 0)\right|.$$

Comme $c_n \in]t_n, t_{n+1}[$ pour tout n on reconnait une somme de Riemann et donc

$$\lim_{h\to 0}\sum_{n=0}^{N-1}|a_n| = \int_{t_0}^{t_0+T}|f(s, y(s)) - \Phi(s, y(s), 0)|\ \mathrm{d}s.$$

Par continuité de y, f et Φ, le schéma est consistant si et seulement si pour tout solution de l'équation différentielle on a

$$f(s, y(s)) - \Phi(s, y(s), 0) = 0, \quad \forall s \in [t_0, t_0 + T].$$

Si $f(t, y) = \Phi(t, y, 0)$ pour tout $(t, y) \in [0, T]\times\mathbb{R}$ le schéma est donc bien consistant. Réciproquement, si le schéma est consistant, étant donné $(t, y) \in [0, T] \times \mathbb{R}$ le théorème de Cauchy-Lipschitz assure l'existence d'une solution telle que $y(t) = y$ et donc en particulier $f(t, y) = \Phi(t, y, 0)$. □

Concernant la stabilité on a la condition suffisante suivante.

Théorème 4.19. *Si Φ est globalement lipschitzienne en y sur $[t_0, t_0 + T] \times \mathbb{R}^d \times [0, h_{\max}]$, i.e. il existe $\Lambda \geq 0$ tel que*

$$\forall t \in [t_0, t_0 + T],\ \forall y_1, y_2 \in \mathbb{R}^d,\ \forall h \in [0, h_{\max}],\ \|\Phi(t, y_1, h) - \Phi(t, y_2, h)\| \leq \Lambda\|y_1 - y_2\|,$$

alors le schéma est stable avec constante de stabilité $S = \mathrm{e}^{\Lambda T}$.

Démonstration. Soient $(y_n)_n$, $(\tilde{y}_n)_n$ et $(\delta_n)_n$ telles que pour tout n

$$y_{n+1} = y_n + h\Phi(t_n, y_n, h) \quad \text{et} \quad \tilde{y}_{n+1} = \tilde{y}_n + h\Phi(t_n, \tilde{y}_n, h) + \delta_n.$$

Pour tout n on a donc

$$\begin{aligned}\|y_{n+1} - \tilde{y}_{n+1}\| &\leq \|y_n - \tilde{y}_n\| + h\|\Phi(t_n, y_n, h) - \Phi(t_n, \tilde{y}_n, h)\| + \|\delta_n\| \\ &\leq (1 + h\Lambda)\|y_n - \tilde{y}_n\| + \|\delta_n\|.\end{aligned}$$

Le lemme de Gronwall discret, Lemme 4.12, permet alors d'affirmer que pour tout n on a

$$\|y_n - \tilde{y}_n\| \leq (1 + h\Lambda)^n\|y_0 - \tilde{y}_0\| + \sum_{j=0}^{n-1}\|\delta_j\|(1 + h\Lambda)^{n-1-j}.$$

Pour tout $n = 0, \dots, N-1$ et tout $j = 0, \dots, n-1$ on a

$$nh \leq Nh = T \quad \text{et} \quad (n-1-j)h \leq Nh = T,$$

et donc, en utilisant à nouveau $1 + h\Lambda \leq \mathrm{e}^{h\Lambda}$,

$$\|y_n - \tilde{y}_n\| \leq \mathrm{e}^{\Lambda T} \left(\|y_0 - \tilde{y}_0\| + \sum_{j=0}^{N-1} \|\delta_j\| \right),$$

ce qui prouve que le schéma est stable avec constante de stabilité $S = \mathrm{e}^{\Lambda T}$. □

4.3.4 Ordre d'un schéma

Une fois assurée la convergence d'un schéma il est naturel de s'intéresser à la "vitesse de convergence".

Définition 4.20. *On dit qu'un schéma est convergent à l'ordre (au moins) $p \geq 1$ si pour tout y_0 et toute suite $(y_{0,h})_h$ on a*

$$\|y_{0,h} - y_0\| \leq Ch^p \implies \sup_{n=0,\dots,N-1} \|y(t_n) - y_n\| \leq C'h^p,$$

où y est la solution du problème de Cauchy avec condition initiale y_0 et $(y_n)_n$ la suite de valeurs approchées obtenues par le schéma à partir de $y_{0,h}$.

Remarque 4.12. *L'équation (4.14) montre que, au moins dans le cas $y_{0,h} \equiv y_0$, la méthode d'Euler est d'ordre au moins 1.*

Définition 4.21. *On dit qu'un schéma est consistant à l'ordre (au moins) $p \geq 1$ si pour toute solution y de l'équation différentielle il existe $C > 0$ telle que pour tout $h \in [0, h_{\max}]$ l'erreur de consistance vérifie*

$$\|\epsilon_n\| \leq Ch^{p+1}, \quad \forall n = 0, \dots, N-1.$$

Remarque 4.13. *Attention ici à la puissance $p+1$. On peut comprendre cela de deux façons. D'un côté ce qui joue un rôle dans la notion de consistance est la somme $\sum_{n=0}^{N-1} \|\epsilon_n\|$ des erreurs de consistance, voir la Définition 4.14. Si le schéma est consistant à l'ordre p on a alors*

$$\sum_{n=0}^{N-1} \|\epsilon_n\| \leq NCh^{p+1} = CTh^p.$$

C'est parfois cette condition qui est prise comme définition de consistance à l'ordre p, cf [2].

D'autre part, la quantité

$$\frac{\epsilon_n}{h} = \frac{y(t_{n+1}) - y(t_n)}{h} - \Phi(t_n, y(t_n), h)$$

représente l'approximation de l'équation différentielle et on a alors

$$\left\| \frac{y(t_{n+1}) - y(t_n)}{h} - \Phi(t_n, y(t_n), h) \right\| \leq Ch^p.$$

On montre alors facilement

Théorème 4.22. *Si un schéma est consistant à l'ordre p et stable alors il est convergent à l'ordre p.*

Démonstration. C'est une conséquence directe de (4.16) et de la définition de consistance à l'ordre p. □

Tout comme pour la consistance, il existe une caractérisation de la consistance à l'ordre p.

Définition 4.23. *Soit $f : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ de classe C^p. On définit pour $k = 0, \ldots, p$ les fonctions $f^{[k]} : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ par récurrence par*

$$f^{[0]}(t,y) = f(t,y), \quad f^{[k+1]}(t,y) = \frac{\partial f^{[k]}}{\partial t}(t,y) + \sum_{j=1}^{d} \frac{\partial f^{[k]}}{\partial y_j}(t,y) \times f_j(t,y), \tag{4.17}$$

où $f = (f_1, \ldots, f_d)$.

Remarque 4.14. *Si $d = 1$ la définition (4.17) s'écrit simplement*

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

L'origine de cette définition est la suivante : si f est de classe C^p alors toute solution $y(t)$ de l'équation différentielle $y' = f(t,y)$ est de classe C^{p+1} et on a, pour tout $k = 0, \ldots, p$,

$$y^{(k+1)}(t) = f^{[k]}(t, y(t)). \tag{4.18}$$

Théorème 4.24. *Si f est de classe C^p et pour tout $k \leq p$ les dérivées partielles $\frac{\partial^k \Phi}{\partial h^k}$ existent et sont continues, la méthode est consistante d'ordre au moins p si et seulement si pour tout $(t,y) \in [0,T] \times \mathbb{R}^d$ on a*

$$\frac{\partial^k \Phi}{\partial h^k}(t,y,0) = \frac{1}{k+1} f^{[k]}(t,y), \qquad \forall 0 \leq k \leq p-1. \tag{4.19}$$

Dans ce cas on a

$$\|\epsilon_n\| \leq h^{p+1} \left[\frac{1}{(p+1)!} \|f^{[p]}\|_\infty + \frac{1}{p!} \left\| \frac{\partial^p \Phi}{\partial h^p} \right\|_\infty \right].$$

Démonstration. On montre uniquement qu'on a une condition suffisante. On pourra consulter par exemple [2, 3] pour la condition nécessaire.

D'après la formule de Taylor avec reste intégral on a

$$\begin{aligned} y(t_{n+1}) - y(t_n) &= \sum_{k=1}^{p} \frac{y^{(k)}(t_n)}{k!} h^k + \int_{t_n}^{t_{n+1}} \frac{(t_{n+1}-s)^p}{p!} y^{(p+1)}(s)\,\mathrm{d}s \\ &= \sum_{k=1}^{p} \frac{y^{(k)}(t_n)}{k!} h^k + \int_{0}^{h} \frac{(h-s)^p}{p!} y^{(p+1)}(t_n+s)\,\mathrm{d}s, \end{aligned}$$

$$\Phi(t_n, y(t_n), h) = \sum_{k=0}^{p-1} \frac{\partial^k \Phi}{\partial h^k}(t_n, y(t_n), 0) \frac{h^k}{k!} + \int_0^h \frac{(h-s)^{p-1}}{(p-1)!} \frac{\partial^p \Phi}{\partial h^p}(t_n, y(t_n), s)\,\mathrm{d}s.$$

En utilisant (4.18) on en déduit que

$$\begin{aligned} \epsilon_n &= \sum_{k=0}^{p-1} \left[\frac{f^{[k]}(t_n, y(t_n))}{(k+1)!} - \frac{1}{k!} \frac{\partial^k \Phi}{\partial h^k}(t_n, y(t_n), 0) \right] h^{k+1} \\ &\quad + \int_0^h \left[\frac{(h-s)^p}{p!} f^{[p]}(t_n+s, y(t_n+s)) - h \frac{(h-s)^{p-1}}{(p-1)!} \frac{\partial^p \Phi}{\partial h^p}(t_n, y(t_n), s) \right] \mathrm{d}s. \end{aligned}$$

Si (4.19) est vérifiée on a alors

$$\begin{aligned} \|\epsilon_n\| &= \left\| \int_0^h \left[\frac{(h-s)^p}{p!} f^{[p]}(t_n+s, y(t_n+s)) - h \frac{(h-s)^{p-1}}{(p-1)!} \frac{\partial^p \Phi}{\partial h^p}(t_n, y(t_n), s) \right] \mathrm{d}s \right\| \\ &\leq \frac{1}{p!} \|f^{[p]}\|_\infty \int_0^h (h-s)^p\,\mathrm{d}s + \frac{h}{(p-1)!} \left\| \frac{\partial^p \Phi}{\partial h^p} \right\|_\infty \int_0^h (h-s)^{p-1}\,\mathrm{d}s \\ &\leq h^{p+1} \left[\frac{1}{(p+1)!} \|f^{[p]}\|_\infty + \frac{1}{p!} \left\| \frac{\partial^p \Phi}{\partial h^p} \right\|_\infty \right]. \end{aligned}$$

□

Exemple 4.4. *Dans la méthode d'Euler on a* $\Phi(t, y, h) = f(t, y)$. *On en déduit que*

$$\frac{\partial \Phi}{\partial h}(t, y, 0) = 0 \neq f^{[1]}(t, y),$$

ce qui prouve que cette méthode n'est pas d'ordre 2.

Exemple 4.5. *Dans la méthode de Taylor à l'ordre* 2 *on a, voir la Section 4.3.1,*

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

On en déduit facilement que si f *est de classe* C^2 *cette méthode est consistante d'ordre* 2.

Exemple 4.6. *Dans la méthode du point milieu, voir la Section 4.3.1, on a*

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

On vérifie que $\Phi(t,y,0) = f(t,y)$ *et que*

$$\frac{\partial\Phi}{\partial h}(t,y,h) = \frac{1}{2}\frac{\partial f}{\partial t}\left(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)\right) + \frac{1}{2}f(t,y)\times\frac{\partial f}{\partial y}\left(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)\right),$$

et donc $\dfrac{\partial\Phi}{\partial h}(t,y,0) = \dfrac{1}{2}f^{[1]}(t,y)$. *Si* f *est de classe* C^2 *cette méthode est d'ordre* 2.

4.4 Méthodes de Runge-Kutta

4.4.1 Présentation des méthodes

Pour simplifier on se placera dans le cas $d = 1$. Les méthodes de Runge-Kutta, dans la suite nommées simplement RK, reposent sur la version intégrale de l'équation différentielle :

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(s,y(s))\,\mathrm{d}s.$$

On utilise ensuite des méthodes dites de quadratures, dont les méthodes vues au Chapitre 1 sont des cas particuliers (voir le Chapitre 5 pour plus de détails à ce sujet), pour approcher l'intégrale $\displaystyle\int_{t_n}^{t_{n+1}} f(s,y(s))\,\mathrm{d}s$. On se donne $c_1,\dots,c_q$ et on introduit les points intermédiaires

$$t_{n,i} = t_n + c_i h.$$

On utilise alors une formule de quadrature du type

$$\int_{t_n}^{t_{n+1}} f(s,y(s))\,\mathrm{d}s \simeq h\sum_{i=1}^{q} b_i f(t_{n,i}, y(t_{n,i})), \tag{4.20}$$

où les b_i sont des réels donnés.

Remarque 4.15. *L'idée est d'utiliser la "même" formule de quadrature pour tout* h *et sur chacun des intervalles* $[t_n, t_{n+1}]$. *On peut en effet écrire, pour tout* n,

$$\int_{t_n}^{t_{n+1}} f(s,y(s))\,\mathrm{d}s = h\int_0^1 f(t_n+hs, y(t_n+hs))\,\mathrm{d}s,$$

et on utilise ensuite la formule de quadrature suivante sur $[0,1]$ *:*

$$\int_0^1 g(s)\,\mathrm{d}s \simeq \sum_{i=1}^{q} b_i g(c_i). \tag{4.21}$$

En prenant $g(s) = f(t_n+hs, y(t_n+hs))$ *on retrouve bien (4.20).*

Remarque 4.16. *En général les c_i sont pris dans l'ordre croissant mais pas nécessairement strictement, i.e. $c_1 \leq c_2 \leq \cdots \leq c_q$.*

Cela conduit à un schéma de la forme

$$y_{n+1} = y_n + h \sum_{i=1}^{q} b_i k_{n,i},$$

où $k_{n,i}$ est une approximation de $f(t_{n,i}, y(t_{n,i})) = y'(t_{n,i})$, i.e. la pente de la solution en $t_{n,i}$. Il faut cependant encore calculer des approximations $y_{n,i}$ de $y(t_{n,i})$ afin d'obtenir les $k_{n,i}$. Cela se fait par une nouvelle méthode de quadrature, on écrit

$$y(t_{n,i}) = y(t_n) + \int_{t_n}^{t_{n,i}} f(s, y(s))\,\mathrm{d}s \simeq y(t_n) + h \sum_{j=1}^{q} a_{i,j} y(t_{n,j}).$$

Remarque 4.17. *Comme ci-dessus l'idée est d'utiliser la "même" formule de quadrature pour tous h et n. On peut en effet écrire*

$$\int_{t_n}^{t_{n,i}} f(s, y(s))\,\mathrm{d}s = h \int_0^{c_i} f(t_n + hs, y(t_n + hs))\,\mathrm{d}s,$$

et on utilise ensuite la formule de quadrature suivante sur $[0, c_i]$:

$$\int_0^{c_i} g(s)\,\mathrm{d}s \simeq \sum_{j=1}^{q} a_{i,j} g(c_j). \tag{4.22}$$

On obtient ainsi le schéma suivant, dit de Runge-Kutta,

$$\begin{cases} t_{n,i} &= t_n + c_i h, & \forall i = 1, \ldots, q, \\ y_{n,i} &= y_n + h \displaystyle\sum_{j=1}^{q} a_{i,j} f(t_{n,j}, y_{n,j}), & \forall i = 1, \ldots, q, \\ y_{n+1} &= y_n + h \displaystyle\sum_{i=1}^{q} b_i f(t_{n,i}, y_{n,i}). \end{cases}$$

Un schéma de Runge-Kutta est donc spécifié par la donnée des $a_{i,j}$, des b_j et des c_i que l'on résume souvent par le tableau suivant, appelé tableau de Butcher,

$$\begin{array}{c|ccc} c_1 & a_{1,1} & \cdots & a_{1,q} \\ \vdots & \vdots & & \vdots \\ c_q & a_{q,1} & \cdots & a_{q,q} \\ \hline & b_1 & \cdots & b_q \end{array}$$

On parle de méthode RK_q pour préciser le nombre de points intermédiaires. Afin de calculer les $y_{n,i}$ on se retrouve ainsi avec un système non-linéaire, très difficile à résoudre, dans lequel le calcul de chacun des points intermédiaires dépend de tous les autres points intermédiaires. Dans la pratique, afin de rendre les calculs réalisables, on impose souvent la condition supplémentaire suivante : le calcul d'un point intermédiaire ne dépend que des points précédents. Cela se traduit simplement par la condition

$$a_{i,j} = 0, \qquad \forall j \geq i,$$

autrement dit la matrice $A = (a_{i,j})_{i,j}$ est triangulaire inférieure stricte. On a en particulier $y_{n,1} = y_n$. On dit parfois alors que la méthode de Runge-Kutta est explicite. Tout schéma de type RK_q est en fait une méthode à un pas : on fait intervenir des points intermédiaires mais le calcul de y_{n+1} ne dépend au final, de façon plus ou moins compliquée, que de y_n et pas des valeurs antérieures. Plus précisément, la fonction $\Phi(t, y, h)$ d'un schéma RK explicite peut s'écrire de la façon suivante,

$$\begin{cases} \Phi_1(t,y,h) = 0, \quad P_1(t,y,h) = f(t + c_1 h, y), & \\ \Phi_i(t,y,h) = \displaystyle\sum_{j=1}^{i-1} a_{i,j} P_j(t,y,h), & i = 2, \dots, q, \\ P_i(t,y,h) = f(t + c_i h, y + h\Phi_i(t,y,h)), & i = 2, \dots, q, \\ \Phi(t,y,h) = \displaystyle\sum_{i=1}^{q} b_i P_i(t,y,h). & \end{cases} \tag{4.23}$$

En ce qui concerne la stabilité des méthodes RK explicites on peut montrer le résultat suivant, voir [3].

Proposition 4.25. *Si f est globalement L-lipschitzienne en y, les méthodes RK explicites sont stables avec constante de stabilité $S = \mathrm{e}^{\Lambda T}$ où $\Lambda = L \displaystyle\sum_{j=1}^{q} \sum_{i=0}^{j-1} |b_j| (\alpha L h_{\max})^i$ et $\alpha = \max_i \sum_j |a_{i,j}|$.*

Le choix des paramètres se fait alors de façon à ce que la méthode soit consistante, et d'ordre le plus élevé possible.

Proposition 4.26. *Une méthode RK est consistante si et seulement si $\displaystyle\sum_{i=1}^{q} b_i = 1$.*

Démonstration. D'après (4.23) on a

$$\Phi(t, y, 0) = \sum_{i=1}^{q} b_i P_i(t, y, 0) = f(t, y) \sum_{i=1}^{q} b_i,$$

et le résultat découle du Théorème 4.18. □

En utilisant le Théorème 4.24 on peut a priori donner des conditions sur les différents paramètres pour qu'un schéma RK soit d'ordre p. Naturellement on ne pourra a priori obtenir une méthode

d'ordre élevé que si le nombre q de points intermédiaires l'est aussi. Afin de limiter les calculs numériques, pour obtenir un ordre souhaité on choisira généralement une méthode avec le moins de points intermédiaires possibles. On donne ici juste la condition pour qu'une méthode soit d'ordre (au moins) 2.

Proposition 4.27. *Si f est de classe C^2 le schéma RK est d'ordre au moins 2 si et seulement si*

$$\sum_{i=1}^{q} b_i = 1, \quad \sum_{i=1}^{q} b_i c_i = \frac{1}{2}, \quad \sum_{i,j=1}^{q} b_i a_{i,j} = \frac{1}{2}.$$

Démonstration. La première condition est celle pour avoir une méthode d'ordre au moins 1, i.e. $\Phi(t,y,0) = f(t,y)$. On calcule ensuite, à partir de (4.23),

$$\begin{aligned}
\frac{\partial \Phi}{\partial h}(t,y,0) &= \sum_{i=1}^{q} b_i \frac{\partial P_i}{\partial h}(t,y,0), \\
\frac{\partial P_i}{\partial h}(t,y,0) &= c_i \frac{\partial f}{\partial t}(t,y) + \Phi_i(t,y,0)\frac{\partial f}{\partial y}(t,y), \\
\Phi_i(t,y,0) &= \sum_{j=1}^{i-1} a_{i,j} f(t,y).
\end{aligned}$$

On en déduit que

$$\frac{\partial \Phi}{\partial h}(t,y,0) = \left(\sum_{i=1}^{q} b_i c_i\right) \frac{\partial f}{\partial t}(t,y) + \left(\sum_{i,j=1}^{q} b_i a_{i,j}\right) \frac{\partial f}{\partial y}(t,y) f(t,y).$$

D'après le Théorème 4.24, la méthode est d'ordre 2 si pour toute fonction f

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

d'où le résultat. □

Remarque 4.18. *On se place également souvent dans le cas où les méthodes de quadratures sont exactes pour les fonctions constantes, i.e. on a égalité dans (4.21) et (4.22). Cela impose les conditions suivantes sur les paramètres :*

$$\sum_{i=1}^{q} b_i = 1, \qquad \sum_{j=1}^{q} a_{i,j} = c_i, \ \forall i = 1, \dots, q.$$

La seconde condition pour $i = 1$ entraine (pour une méthode RK explicite) que $c_1 = 0$. La première condition correspond en fait à celle de la consistance de la méthode. Si la seconde est vérifiée on peut alors voir que les deux conditions supplémentaires pour que la méthode soit d'ordre 2 reviennent au même : $\sum_{i=1}^{q} b_i c_i = \frac{1}{2}$.

Remarque 4.19. *Pour $q = 1$ on a, voir (4.23),*

$$\Phi(t, y, h) = b_1 f(t + c_1 h, y).$$

Si on veut que la méthode soit consistante et que la formule de quadrature soit exacte pour les fonctions constantes on obtient $\Phi(t, y, h) = f(t, y)$. C'est la méthode d'Euler explicite.

Dans les sections suivantes on étudie deux cas particuliers : les méthodes dites RK2 et RK4 qui correspondent à $q = 2$ et $q = 4$ respectivement.

4.4.2 La méthode RK2

On se place ici dans le cas $q = 2$, explicite, et où les méthodes de quadratures sont exactes pour les fonctions constantes. On cherche à quelle condition cette méthode est d'ordre 2. On a donc a priori 3 paramètres : c_2, b_1, b_2. En effet, $c_1 = 0$ et $a_{2,1} = c_2$ d'après la Remarque 4.18 et $a_{1,1} = a_{1,2} = a_{2,2} = 0$ pour que le schéma soit explicite. La méthode est d'ordre 1 si et seulement si $b_1 + b_2 = 1$, et elle est d'ordre 2 si de plus $b_1 c_1 + b_2 c_2 = b_2 c_2 = \frac{1}{2}$. En prennt $b_2 = \alpha$ comme paramètre, on trouve ainsi une famille de solutions

$$b_1 = 1 - \alpha,\ b_2 = \alpha,\ c_2 = \frac{1}{2\alpha}.$$

Les méthodes de quadratures sous-jacentes sont

$$\int_0^1 g(s)\,\mathrm{d}s \simeq b_1 g(c_1) + b_2 g(c_2) = (1-\alpha) g(0) + \alpha g\left(\frac{1}{2\alpha}\right),$$

pour le calcul de y_{n+1} et

$$\int_0^{1/2\alpha} g(s)\,\mathrm{d}s \simeq \frac{1}{2\alpha} g(0),$$

pour celui du point $y_{n,2}$ (on rappelle que $y_{n,1} = y_n$ dans les méthodes RK explicites). On y reconnait une méthode des rectangles à gauche, autrement dit on calcule $y_{n,2}$ à partir de y_n à l'aide de la méthode d'Euler.

La fonction Φ du schéma est alors

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

On peut vérifier en utilisant le Théorème 4.24 que la méthode n'est pas d'ordre 3.

- Pour $\alpha = 1$ on reconnait la méthode du point milieu et la première méthode de quadrature est bien entendu celle du point milieu.
- Pour $\alpha = \frac{1}{2}$ on a

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

  Elle est appelée méthode de Heun. Elle repose sur la méthode des trapèzes : $\displaystyle\int_0^1 g(s)\,\mathrm{d}s \simeq \frac{g(0) + g(1)}{2}$.

4.4.3 La méthode RK4

On ne va donner ici qu'un exemple de méthode RK4 qui est celle qui est la plus utilisée dans la pratique. Le tableau de Butcher correspondant est

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

La méthode s'écrit aussi

$$\Phi(t,y,h) = \frac{1}{6}(P_1 + 2P_2 + 2P_3 + P_4),$$

où

$$\left\{ \begin{array}{lcl} P_1(t,y,h) & = & f(t,y), \\ P_2(t,y,h) & = & f\left(t+\frac{h}{2}, y+\frac{h}{2}P_1(t,y,h)\right), \\ P_3(t,y,h) & = & f\left(t+\frac{h}{2}, y+\frac{h}{2}P_2(t,y,h)\right), \\ P_4(t,y,h) & = & f\left(t+h, y+hP_3(t,y,h)\right). \end{array} \right.$$

On vérifie facilement que les conditions de la Proposition 4.27 sont satisfaites et donc que la méthode est d'ordre au moins 2. On peut en fait montrer à l'aide du Théorème 4.24 qu'elle est d'ordre 4.

Cette méthode repose sur les formules de quadratures suivantes :

- pour le calcul de y_{n+1}, (4.21) s'écrit

$$\int_0^1 g(s)\,\mathrm{d}s \simeq \frac{1}{6}\left(g(c_1) + 2g(c_2) + 2g(c_3) + g(c_4)\right) = \frac{1}{6}\left(g(0) + 4g\left(\frac{1}{2}\right) + g(1)\right).$$

  On reconnait la méthode de Simpson.

- pour le calcul des points intermédiaires, (4.22) devient

$$\begin{array}{rcl} \displaystyle\int_0^{c_2} g(s)\,\mathrm{d}s & \simeq & c_2 g(0), \\ \displaystyle\int_0^{c_3} g(s)\,\mathrm{d}s & \simeq & c_3 g\left(c_2\right) = c_3 g(c_3), \\ \displaystyle\int_0^{c_4} g(s)\,\mathrm{d}s & \simeq & c_4 g\left(c_3\right) = c_4 g\left(\frac{c_4}{2}\right). \end{array}$$

  Ce sont des méthodes des rectangles à gauche, à droite et au point milieu.

Chapitre 5

Compléments sur le calcul approché d'intégrales : méthodes de quadrature

Le principe des méthodes de quadrature est d'approcher une intégrale $\displaystyle\int_a^b f(x)\,\mathrm{d}x$ par une formule du type $\displaystyle\sum_{j=0}^{N}\lambda_j f(x_j)$, appelée formule de quadrature simple, où les x_j sont des points de $[a,b]$ et les λ_j des réels, appelés poids. On parle alors de formule de *quadrature simple*. On utilisera souvent des formules dites de *quadrature composée*. On commence par subdiviser $[a,b]$ en des segments $[a_k,a_{k+1}]$, avec $a=a_0<a_1<\cdots<a_n=b$, et on utilise une formule de quadrature simple sur chacun des $[a_k,a_{k+1}]$ puis la relation de Chasles. C'est une généralisation de ce qui a été fait au Chapitre 1. En effet, si on prend par exemple la méthode de Simpson, on est dans le cas d'une méthode de quadrature composée avec $a_k=a+\dfrac{k}{n}(b-a)$ et sur chaque intervalle $[a_k,a_{k+1}]$ on utilise la formule de quadrature simple

$$\int_{a_k}^{a_{k+1}} f(x)\,\mathrm{d}x \simeq \frac{b-a}{6n}f(a_k)+\frac{4(b-a)}{6n}f\left(\frac{a_k+a_{k+1}}{2}\right)+\frac{b-a}{6n}f(a_{k+1}).$$

5.1 Méthodes de Newton-Cotes

5.1.1 Formules de quadrature simple

On commence par s'intéresser au cas des formules de quadratures simples. On se donne donc un intervalle $[a,b]$ et on souhaite une formule permettant d'approcher $\displaystyle\int_a^b f(x)\,\mathrm{d}x$.

Définition 5.1. *On appelle formule de quadrature simple, ou de Newton-Cotes, à $N+1$ points toute formule du type*

$$I_N(f)=\sum_{j=0}^{N}\lambda_j f(x_j)$$

où les x_j sont des points de $[a,b]$ et les λ_j des réels ne dépendant pas de f.

L'erreur d'intégration de f est alors $E_N(f) = \int_a^b f(x)\,\mathrm{d}x - I_N(f)$.

Attention !! Dans la définition de l'erreur celle-ci a un signe.

Remarque 5.1. *On peut remarquer que les formules de quadrature sont linéaires :* $I_N(f+g) = I_N(f) + I_N(g)$. *C'est donc également le cas pour l'erreur d'intégration :* $E_N(f+g) = E_N(f) + E_N(g)$.

Définition 5.2. *On dit qu'une formule de quadrature est d'ordre supérieur ou égal à K si elle est exacte pour toute fonction polynomiale de degré au plus K, i.e. si pour toute fonction* $f(x) = \sum_{i=0}^{K} \alpha_i x^i$ *on a* $E_N(f) = 0$. *On dit qu'elle est d'ordre exactement K si elle est d'ordre supérieur ou égal à K et s'il existe au moins une fonction polynomiale g de degré $K+1$ telle que* $E_N(g) \neq 0$.

Définition 5.3. *Une formule de quadrature I_N est dite de type interpolation si pour toute fonction f on a*

$$I_N(f) = \int_a^b P_N(f)(x)\,\mathrm{d}x,$$

où $P_N(f)$ est le polynôme d'interpolation de Lagrange de f associé aux $N+1$ points x_j de la formule de quadrature.

Remarque 5.2. *Si I_N est de type interpolation alors $E_N(f) = 0$ pour toute fonction polynomiale de degré au plus N. On peut également remarquer que dans ce cas $\sum_{j=0}^{N} \lambda_j = b - a$. Il suffit pour cela de considérer la fonction constante égale à* 1.

Proposition 5.4. *I_N est de type interpolation si et seulement si pour tout $i = 0, \ldots, N$ on a* $\lambda_i = \int_a^b L_i(x)\,\mathrm{d}x$ *où* $L_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}$.

Démonstration. Si I_N est de type interpolation on a pour tout i

$$I_N(L_i) = \int_a^b P_n(L_i)(x)\,\mathrm{d}x. \tag{5.1}$$

Or $L_i \in \mathbb{R}_N[X]$ donc $P_N(L_i) = L_i$, et pour tout j on a $L_i(x_j) = \delta_{ij}$ d'où

$$I_N(L_i) = \sum_{j=0}^{n} \lambda_j L_i(x_j) = \lambda_i.$$

L'égalité (5.1) devient donc $\lambda_i = \int_a^b L_i(x)\,\mathrm{d}x$.

Réciproquement, si pour tout i on a $\lambda_i = \int_a^b L_i(x)\,\mathrm{d}x$, alors quelque soit f on a

$$I_N(f) = \sum_{j=0}^{N} f(x_j) \int_a^b L_j(x)\,\mathrm{d}x = \int_a^b \left(\sum_{j=0}^{N} f(x_j) L_j(x) \right) \mathrm{d}x = \int_a^b P_N(f)(x)\,\mathrm{d}x,$$

où on a utilisé, voir la Section 3.1, que $P_N(f) = \sum_{i=0}^{N} f(x_i) L_i$ pour toute fonction f. I_N est donc bien de type interpolation. □

Exercice 5.1. On considère la formule d'interpolation à 3 points avec $x_0 = a$, $x_1 = \dfrac{a+b}{2}$ et $x_2 = b$. Montrer que la formule de quadrature est

$$I_2(f) = \frac{b-a}{6} f(a) + \frac{4(b-a)}{6} f\left(\frac{a+b}{2}\right) + \frac{b-a}{6} f(b).$$

Que reconnaissez-vous ?

Même question avec les formules d'interpolation à 1 point avec $x_0 = a$, resp. $x_0 = b$ puis $x_0 = \dfrac{a+b}{2}$, et avec la formule d'interpolation à 2 points avec $x_0 = a$ et $x_1 = b$.

Proposition 5.5. *Une formule de quadrature à $N+1$ points est d'ordre supérieur ou égal à N si et seulement si elle est de type interpolation.*

Démonstration. Si la formule est de type interpolation alors elle est bien d'ordre au moins N puisque si f est un polynôme de degré au plus N on a $P_N(f) = f$.

Réciproquement, supposons que I_N est d'ordre au moins N. Puisque L_i est de degré N pour tout i on a

$$I_N(L_i) = \int_a^b L_i(x)\,\mathrm{d}x \quad \Longleftrightarrow \quad \sum_{j=0}^{N} \lambda_j L_i(x_j) = \int_a^b L_i(x)\,\mathrm{d}x \quad \Longleftrightarrow \quad \lambda_i = \int_a^b L_i(x)\,\mathrm{d}x.$$

D'après la proposition précédente I_N est donc de type interpolation. □

Remarque 5.3. *Si on fixe les points x_j dans la formule de quadrature il reste $N+1$ paramètres libres : les λ_j. L'espace des polynômes de degré au plus K étant de dimension $K+1$ il est donc raisonnable de pouvoir espérer choisir les λ_j de façon à ce que la formule soit d'ordre au moins N. C'est précisément ce que dit la proposition ci-dessus. Il y a alors une unique façon de choisir les λ_j. Le fait d'avoir ou non une formule d'ordre strictement plus grand que N ne pourra alors dépendre que du choix des points d'interpolation x_j (voir la Section 5.2).*

Un cas particulier important est celui où on choisit les x_j uniformément répartis dans $[a,b]$, i.e. pour tout $j=0,\dots,N$ on prend $x_j = a + \dfrac{j}{N}(b-a)$.

Proposition 5.6. *On considère la formule d'interpolation à $N+1$ points avec les x_j uniformément répartis dans $[a,b]$. Alors*

1. *Les poids vérifient la symétrie $\lambda_{N-j} = \lambda_j$ pour tout j.*
2. *Si N est pair la méthode est d'odre supérieur ou égal à $N+1$.*

Remarque 5.4. *La deuxième propriété montre que le phénomène observé pour la méthode de Simpson (voir Section 1.3), à savoir que la méthode est exacte non seulement pour les polynômes de degré au plus 2 mais aussi pour ceux de degré au plus 3, n'est pas un phénomène isolé. Il se retrouve pour toutes les formules de Newton-Cotes avec un nombre impair de points équirépartis.*

Remarque 5.5. *On peut en fait montrer que si les points sont uniformément répartis la méthode est d'ordre exactement N pour N impair et d'ordre exactement $N+1$ pour N pair.*

Démonstration. 1. On remarque que le polynôme $\tilde{L}_j(x) := L_j(a+b-x)$ est de degré N et vérifie

$$\begin{aligned}\tilde{L}_j(x_i) &= L_j(a+b-x_i) = L_j\left(a+b-a-\frac{i}{N}(b-a)\right) = L_j\left(a+\frac{N-i}{N}(b-a)\right)\\ &= L_j(x_{N-i}) = \delta_{N-i,j} = \delta_{i,N-j}.\end{aligned}$$

Par unicité des polynômes d'interpolation on en déduit que $\tilde{L}_j = L_{N-j}$. On a donc

$$\lambda_{N-j} = \int_a^b L_{N-j}(x)\,\mathrm{d}x = \int_a^b L_j(a+b-x)\,\mathrm{d}x = \int_a^b L_j(t)\,\mathrm{d}t = \lambda_j,$$

où on a utilisé le changement de variable $t = a+b-x$ à la troisième égalité.

2. On sait déjà que la formule est d'ordre supérieur ou égal à N. Par linéarité de l'erreur il suffit de montrer que cette dernière est nulle pour un polynôme de degré $N+1$. Prenons $P(X) = \prod_{j=0}^{N}(X-x_j)$. On a évidemment $P(x_j)=0$ pour tout j donc $E(P) = \int_a^b P(x)\,\mathrm{d}x$. En utilisant $x_{N-j} = a+b-x_j$ pour tout j et la parité de N on a

$$P(a+b-X) = \prod_{j=0}^{N}(a+b-X-x_j) = \prod_{j=0}^{N}(x_{N-j}-X) = \prod_{k=0}^{N}(x_k-X) = (-1)^{N+1}P(X) = -P(X).$$

On en déduit que

$$E(P) = \int_a^b P(x)\,\mathrm{d}x = \int_a^b P(a+b-t)\,\mathrm{d}t = -\int_a^b P(t)\,\mathrm{d}t = -E(P),$$

et donc $E(P)=0$.

N.B. : Cet argument est le même que celui utilisé à la fin de la Section 1.3 pour expliquer pourquoi la méthode de Simpson est exacte pour les polynômes de degré au plus 3. □

Étant donnée une formule de quadrature, la question suivante est celle d'estimer l'erreur commise.

Proposition 5.7. *Si I_N est de type interpolation, on a alors pour toute fonction f de classe C^{N+1}*

$$|E_N(f)| \leq \frac{b-a}{(N+1)!}\|\omega\|_\infty \|f^{(N+1)}\|_\infty,$$

où $\omega(x) = \prod_{j=0}^{N}(x-x_j)$ avec les x_j les points d'interpolation et $\|g\|_\infty = \sup_{x\in[a,b]} |g(x)|$.

Démonstration. Le résultat découle directement de la Proposition 3.3 et de l'inégalité

$$E_N(f) = \left|\int_a^b f(x)\,\mathrm{d}x - \int_a^b P_N(f)(x)\,\mathrm{d}x\right| \leq \int_a^b |f(x)-P_N(f)(x)|\,\mathrm{d}x \leq (b-a)\|f-P_N(f)\|_\infty.$$

□

Remarque 5.6. *Pour tout $x \in [a,b]$ on a $|x-x_j| \leq b-a$ quelque soit le choix des points d'interpolation. On en déduit qu'on a toujours $\|\omega\|_\infty \leq (b-a)^{N+1}$ et donc pour tout formule de type interpolation à $N+1$ points on a*

$$|E_N(f)| \leq \frac{(b-a)^{N+2}}{(N+1)!}\|f^{(N+1)}\|_\infty. \tag{5.2}$$

Remarque 5.7. *On peut ici faire le même constat que dans la Remarque 3.2, il n'est pas garantit que l'erreur tende vers 0 lorsque N augmente. C'est pour cela qu'on utilisera souvent des méthodes de quadrature composées.*

On peut en fait améliorer en général un peu le résultat ci-dessus, voir la Remarque 5.9 à la fin de la Section 5.1.2. On introduit pour cela la notion de noyau de Peano.

Définition 5.8. *Etant donnés une méthode de quadrature à $N+1$ points et $m \in \mathbb{N}$, le noyau de Peano d'ordre m associé à la méthode de quadrature est la fonction $K_m : [a,b] \to \mathbb{R}$ définie par*

$$K_m(t) = E_N(f_m),$$

où $f_m(x) = (x-t)_+^m$, i.e. $f_m(x) = (x-t)^m$ si $x-t \geq 0$ et $f_m(x) = 0$ sinon.

On a alors l'estimation suivante de l'erreur.

Théorème 5.9. *Si la méthode de quadrature est d'ordre m alors pour toute fonction f de classe C^{m+1} on a*

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

En particulier,

$$|E(f)| \leq \frac{\|f^{(m+1)}\|_\infty}{m!}\int_a^b |K_m(t)|\,\mathrm{d}t. \tag{5.3}$$

Démonstration. Soit f de classe C^{m+1}. La formule de Taylor avec reste intégral s'écrit

$$\begin{aligned} f(x) &= \sum_{k=0}^{m} \frac{f^{(k)}(a)}{k!}(x-a)^k + \int_a^x \frac{(x-t)^m}{m!} f^{(m+1)}(t)\,\mathrm{d}t \\ &= \underbrace{\sum_{k=0}^{m} \frac{f^{(k)}(a)}{k!}(x-a)^k}_{=P(x)} + \underbrace{\int_a^b \frac{(x-t)_+^m}{m!} f^{(m+1)}(t)\,\mathrm{d}t}_{=R(x)}. \end{aligned}$$

On peut alors écrire

$$E(f) = E(P+R) = E(P) + E(R) = E(R),$$

où on a utilisé la linéarité de l'erreur (voir Remarque 5.1), le fait que P est de degré m et que la méthode est d'ordre m (on a donc $E(P) = 0$). On a alors

$$\begin{aligned} E(R) &= \int_a^b R(x)\,\mathrm{d}x - \sum_{j=0}^{m} \lambda_j R(x_j) \\ &= \int_a^b \left(\int_a^b \frac{(x-t)_+^m}{m!} f^{(m+1)}(t)\,\mathrm{d}t \right) \mathrm{d}x - \sum_{j=0}^{N} \lambda_j \int_a^b \frac{(x_j-t)_+^m}{m!} f^{(m+1)}(t)\,\mathrm{d}t \\ &= \int_a^b \left(\int_a^b (x-t)_+^m\,\mathrm{d}x \right) \frac{f^{(m+1)}(t)}{m!}\,\mathrm{d}t - \int_a^b \left(\sum_{j=0}^{N} \lambda_j (x_j-t)_+^m \right) \frac{f^{(m+1)}(t)}{m!}\,\mathrm{d}t \\ &= \int_a^b \left(\int_a^b (x-t)_+^m\,\mathrm{d}x - \sum_{j=0}^{N} \lambda_j (x_j-t)_+^m \right) \frac{f^{(m+1)}(t)}{m!}\,\mathrm{d}t \\ &= \int_a^b K_m(t) \frac{f^{(m+1)}(t)}{m!}\,\mathrm{d}t. \end{aligned}$$

□

Exercice 5.2. Montrer que le noyau de Peano K_3 pour la méthode de Simpson ($N = 2$ et $x_0 = a$, $x_1 = \frac{a+b}{2}$, $x_2 = b$) est donné par

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

En remarquant que K_3 est toujours négatif calculer $\displaystyle\int_a^b |K_3(t)|\,\mathrm{d}t$ puis retrouver la majoration de l'erreur d'intégration (1.7).

5.1.2 Formules de quadrature composée

Comme on l'a remarqué dans la section précédente, le fait d'augmenter le nombre de points ne garantit pas une diminution de l'erreur. Cela amène aux formules de quadrature composées. On

commence par diviser l'intervalle $[a, b]$ en n sous-intervalles $[a_k, a_{k+1}]$ avec $a = a_0 < a_1 < \cdots < a_n = b$ et on utilise ensuite une formule de quadrature simple sur chacun des sous-intervalles. C'est exactement ce qu'on fait dans les méthodes des rectangles, resp. des trapèzes ou de Simpson : on découpe $[a, b]$ et sur chaque intervalle on approche f par une fonction constante, resp. affine ou polynômiale de degré au plus 2.

Définition 5.10. *On appelle formule de quadrature composée toute formule du type*

$$I(f) = \sum_{k=0}^{n-1} \sum_{j=0}^{N_k} \lambda_{k,j} f(x_{k,j}),$$

où pour tout k les $\lambda_{k,j}$ sont les poids et les $x_{k,j}$ les points de la formule de quadrature simple sur l'intervalle $[a_k, a_{k+1}]$, avec $a = a_0 < \cdots < a_n = b$.

Remarque 5.8. *Pour k fixé, $\displaystyle\sum_{j=0}^{N_k} \lambda_{k,j} f(x_{k,j})$ est une formule de quadrature simple à $N_k + 1$ points sur l'intervalle $[a_k, a_{k+1}]$.*

Dans la pratique on considère souvent le cas où les a_k sont équidistants, i.e. $a_k = a + \dfrac{k}{n}(b-a)$, et où sur chaque intervalle $[a_k, a_{k+1}]$ on utilise une même formule de quadrature (c'est exactement ce qu'on a fait au Chapitre 1), en particulier $N_k \equiv N$ (le nombre de points est le même pour chaque intervalle). Précisons ce qu'on entend par une même formule de quadrature puisque les $[a_k, a_{k+1}]$ sont distincts :

1. Dans le cas où les a_k sont équidistants, les intervalles sont juste les translatés de l'intervalle $[a_0, a_1]$: $[a_k, a_{k+1}] = \left[a_0 + \dfrac{k}{n}(b-a), a_1 + \dfrac{k}{n}(b-a)\right]$, et on procède de même sur les $x_{k,j}$ tout en gardant des poids identiques : pour tous j et k on a $x_{k,j} = x_{0,j} + \dfrac{k}{n}(b-a)$ et $\lambda_{k,j} = \lambda_{0,j}$ (on écrira simplement λ_j).
2. Si les a_k ne sont pas nécessairement équidistants, la fonction affine

$$\varphi_k(x) := \frac{a_{k+1} - a_k}{a_1 - a_0}(x - a_0) + a_k$$

envoie l'intervalle $[a_0, a_1]$ sur $[a_k, a_{k+1}]$ (dans le cas équidistants on retrouve la translation précédente $\varphi_k(x) = x - a_0 + a_k = x + \frac{k}{n}(b-a)$). On prend alors, pour tous k et j, $x_{k,j} = \varphi_k(x_{0,j})$ et $\lambda_{k,j} = \dfrac{a_{k+1} - a_k}{a_1 - a_0}\lambda_{0,j}$.

Une remarque importante est que si la formule est de type interpolation sur $[a_0, a_1]$ elle l'est aussi sur chacun des $[a_k, a_{k+1}]$ (c'est l'origine du facteur $\dfrac{a_{k+1} - a_k}{a_1 - a_0}$ dans la définition des $\lambda_{k,j}$: dans une formule de type interpolation la somme des poids est égale à la longueur de l'intervalle). En

effet, si f est une fonction polynomiale de degré au plus N sur $[a_k, a_{k+1}]$ alors $f_k = f \circ \varphi_k$ est une fonction polynomiale de degré au plus N sur $[a_0, a_1]$. On a donc

$$\begin{aligned} 0 = E(f_k) &= \int_{a_0}^{a_1} f_k(x)\,\mathrm{d}x - \sum_{j=0}^{N} \lambda_{0,j} f_k(x_{0,j}) \\ &= \int_{a_0}^{a_1} f \circ \varphi_k(x)\,\mathrm{d}x - \sum_{j=0}^{N} \lambda_{0,j} f\left(\varphi_k(x_{0,j})\right) \\ &= \frac{a_1 - a_0}{a_{k+1} - a_k} \int_{a_k}^{a_{k+1}} f(t)\,\mathrm{d}t - \sum_{j=0}^{N} \frac{a_1 - a_0}{a_{k+1} - a_k} \lambda_{k,j} f(x_{k,j}) \\ &= \frac{a_1 - a_0}{a_{k+1} - a_k} E(f), \end{aligned}$$

où on a utilisé à la troisième ligne le changement de variable $t = \varphi_k(x)$ et le fait que $\varphi_k'(x) \equiv \frac{a_{k+1}-a_k}{a_1-a_0}$. La formule est donc d'ordre supérieur ou égal à N, elle est donc de type interpolation.

L'estimation de l'erreur suivante est une conséquence directe de la formule de Chasles et de l'estimation (5.2).

Proposition 5.11. *Si les a_k sont équidistants et sur chaque intervalle $[a_k, a_{k+1}]$ la formule de quadrature est identique (dans le sens ci-dessus) et de type interpolation alors pour tout f de classe C^{N+1} l'erreur d'intégration*

$$E(f) = \int_a^b f(x)\,\mathrm{d}x - \sum_{k=0}^{n-1} \sum_{j=0}^{N} \lambda_j f(x_{k,j})$$

vérifie

$$|E(f)| \leq \frac{(b-a)^{N+2}}{(N+1)! n^{N+1}} \|f^{(N+1)}\|_\infty.$$

On remarque immédiatement que l'erreur tend vers 0 lorsque n, le nombre d'intervalles dans la subdivision, tend vers l'infini (et pas le nombre de points dans les méthodes de quadratures simples).

Démonstration. En utilisant la relation de Chasles on écrit $E(f) = \sum_{k=0}^{n-1} E_k(f)$ où

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

L'estimation (5.2) donne alors

$$|E_k(f)| \leq \frac{1}{(N+1)!} \left(\frac{b-a}{n}\right)^{N+2} \sup_{x \in [a_k, a_{k+1}]} |f^{(N+1)}(x)| \leq \frac{(b-a)^{N+2}}{(N+1)! n^{N+2}} \|f^{(N+1)}\|_\infty.$$

Le résultat découle alors de l'inégalité triangulaire. □

Remarque 5.9. *Si on prend une méthode de quadrature dans laquelle les poids sont tous positifs (c'est le cas par exemple de la méthode de Simpson), en majorant (très grossièrement) le noyau de Peano, on peut montrer que dans (5.3) on a*

$$\int_a^b |K_m(t)|\,\mathrm{d}t \leq C(b-a)^{m+2},$$

où m est l'ordre de la méthode. Ainsi, si on utilise l'estimation (5.3) au lieu de (5.2) dans chacun des intervalles $[a_k, a_{k+1}]$ on peut voir que l'erreur est alors en $O\left(n^{-m-1}\right)$ où m est l'ordre de la méthode. Ce dernier étant toujours supérieur ou égal à $N+1$ cela donne une meilleure estimation que celle de la Proposition 5.11. C'est en ce sens que l'estimation (5.3) est meilleure que (5.2).

5.2 Méthode de Gauss

Comme mentionné dans la Remarque 5.9, on peut montrer que l'erreur dans une méthode de quadrature composée, dans laquelle la méthode de quadrature simple utilisée dans chaque intervalle est d'ordre m et pourvu que les poids soient positifs, est en $O(n^{-m-1})$ pour toute fonction f de classe C^{m+1} et où n est le nombre d'intervalles dans la subdivision de $[a,b]$. En particulier, plus l'ordre de la méthode de quadrature simple est élevé, i.e. plus m est grand, plus la méthode de quadrature composée convergera rapidement lorsque n tendra vers l'infini.

Une fois les $N+1$ points de la méthode de quadrature choisis, on a vu que le mieux était de prendre une formule de type interpolation (condition nécessaire pour que l'ordre de la méthode soit au moins N). Le choix le plus simple des points d'interpolation est de les choisir uniformément répartis dans l'intervalle et on a vu qu'alors la méthode est d'ordre exactement N si N est impair et exactement $N+1$ si N est pair. Il est naturel de se demander si un autre choix de points d'interpolation ne serait pas meilleur, et dans ce cas quel ordre on peut espérer.

Si la méthode est d'ordre m elle doit être exact pour tout polynôme de degré au plus m, en particulier pour tous les x^k avec $k \leq m$, ce qui s'écrit

$$\int_a^b x^k\,\mathrm{d}x = \sum_{j=0}^{N} \lambda_j x_j^k \quad \Longleftrightarrow \quad \frac{b^{k+1}-a^{k+1}}{k+1} = \sum_{j=0}^{N} \lambda_j x_j^k.$$

On a alors un système (non-linéaire à cause des x_j^k) de $m+1$ équations ($k=0,\ldots,m$) à $2(N+1)$ inconnues : les λ_j et les x_j. On peut a priori espérer une solution tant que $m+1 \leq 2(N+1)$, i.e. $m \leq 2N+1$, et donc l'ordre maximum raisonnable est $2N+1$. Il faudrait également s'assurer que les x_j soient bien 2 à 2 distincts. Le Théorème 5.13 ci-dessous dit que c'est précisément le cas et donne les points x_j correspondants (les λ_j sont alors forcément donnés par la Proposition 5.4). Il repose sur les polynômes orthogonaux considérés au Chapitre 3. Pour tout entier n, p_n désigne l'unique polynôme unitaire de degré n telle que la famille $(p_m)_m$ soit orthogonale pour le produit scalaire $\int_a^b f(x)g(x)\,\mathrm{d}x$ (voir la Proposition 3.11).

On aura besoin du résultat suivant concernant les racines de ces polynômes orthogonaux.

Proposition 5.12. *Pour tout intervalle $[a,b]$ et tout $n \in \mathbb{N}$ le polynôme p_n admet n racines réelles distincts qui sont toutes dans l'intervalle $[a,b]$.*

Démonstration. On commence par montrer que p_n a exactement n racines réelles dans $[a,b]$ (comptées éventuellement avec multiplicité). On note $x_1, \dots, x_k$ les racines de p_n dans l'intervalle $[a,b]$ et on note $p(x) = (x-x_1)\cdots(x-x_k)$ (si p_n n'a aucune racine dans $[a,b]$ alors $p(x) = 1$). On peut ainsi écrire $p_n = p \times q$ avec q n'ayant aucune racine dans $[a,b]$. En particulier q est de signe constant sur $[a,b]$.

On a alors, puisque q est de signe constant sur $[a,b]$,

$$\int_a^b p(x)p_n(x)\,\mathrm{d}x = \int_a^b p(x)^2 q(x)\,\mathrm{d}x \neq 0$$

Comme p_n est orthogonal à $\mathbb{R}_{n-1}[X]$ cela prouve que p est de degré au moins n et donc que $p = p_n$: toutes les racines de p_n sont donc réelles et dans $[a,b]$.

Il reste à montrer qu'elles sont toutes simples. Supposons que p_n ait au moins une racine double x_0. On peut donc écrire $p_n(x) = (x-x_0)^2 q(x)$ avec q de degré $n-2$. En particulier p_n et q sont orthogonaux, c'est-à-dire

$$\int_a^b p_n(x)q(x)\,\mathrm{d}x = \int_a^b (x-x_0)^2 q(x)^2\,\mathrm{d}x = 0,$$

ce qui est impossible. □

Théorème 5.13. *Soit $[a,b]$ fixé. Pour tout $N \in \mathbb{N}$ il existe un unique choix de points $x_0, \dots, x_N$ et de poids $\lambda_0, \dots, \lambda_N$ tels que la méthode de quadrature simple associée soit d'ordre supérieur ou égal à $2N+1$. Elle est alors d'ordre exactement $2N+1$ et les points x_j sont les racines de p_{N+1} (appelés points de Gauss-Legendre) et les λ_j sont alors donnés par la Proposition 5.4.*

Démonstration. On raisonne par analayse/syntèse.

Supposons que les x_j et λ_j soient tels que la méthode est d'ordre supérieur ou égal à $2N+1$. On note $w(x) = \prod_{j=0}^{N}(x-x_j)$. Il est de degré $N+1$ donc pour tout $P \in \mathbb{R}_N[X]$ on a $wP \in \mathbb{R}_{2N+1}[X]$. Puisque la méthode est d'ordre supérieur ou égal à $2N+1$ on a alors

$$\int_a^b w(x)P(x)\,\mathrm{d}x = \sum_{j=0}^{N} \lambda_j \underbrace{w(x_j)}_{=0} P(x_j) = 0.$$

Ainsi w est un polynôme unitaire, de degré $N+1$ et orthogonal à $\mathbb{R}_N[X]$. C'est donc nécessairement p_{N+1} et les x_j sont les racines de ce dernier. Puisque $2N+1 \geq N$, la méthode est de type interpolation et cela fixe donc aussi les λ_j.

On prend donc comme points x_j les racines de p_{N+1} et comme poids λ_j tels que la formule soit de type interpolation. D'après la Proposition 5.12 les x_j sont bien deux à deux distincts et dans $[a,b]$. On montre que la méthode est alors d'ordre exactement $2N+1$. Il faut pour cela montrer que pour tout $P \in \mathbb{R}_{2N+1}[X]$ on a

$$\int_a^b P(x)\,\mathrm{d}x = \sum_{j=0}^{N} \lambda_j P(x_j),$$

et qu'il y a au moins un polynôme de degré $2N+2$ pour lequel l'identité ci-dessus est fausse.

On effectue la division euclidienne de P par p_{N+1} : $P = Qp_{N+1} + R$ avec $\deg(R) < \deg(p_{N+1})$, i.e. $R \in \mathbb{R}_N[X]$. On remarque par ailleurs que puisque $P \in \mathbb{R}_{2N+1}[X]$ on a aussi $Q \in \mathbb{R}_N[X]$. On a donc

$$\int_a^b P(x)\,\mathrm{d}x = \int_a^b Q(x)p_{N+1}(x)\,\mathrm{d}x + \int_a^b R(x)\,\mathrm{d}x.$$

Le premier terme du membre de droite est nul par définition de p_{N+1} et puisque $Q \in \mathbb{R}_N[X]$. Par ailleurs R est de degré au plus N et la méthode est d'ordre au moins N (elle est de type interpolation). On obtient ainsi

$$\int_a^b P(x)\,\mathrm{d}x = \int_a^b R(x)\,\mathrm{d}x = \sum_{j=0}^{N} \lambda_j R(x_j) = \sum_{j=0}^{N} \lambda_j P(x_j),$$

où on a utilisé $p_{N+1}(x_j) = 0$ pour tout j à la dernière égalité. La méthode est donc bien d'ordre supérieur ou égal à $2N+1$.

Finalement p_{N+1}^2 est de degré $2N+2$, $\displaystyle\sum_{j=0}^{N} \lambda_j p_{N+1}^2(x_j) = 0$ tandis que $\displaystyle\int_a^b p_{N+1}^2(x)\,\mathrm{d}x \neq 0$, ce qui prouve que la méthode est d'ordre exactement $2N+1$. □

Remarque 5.10. *On peut montrer que dans la méthode de Gauss les poids sont positifs. En effet, la méthode est d'ordre $2N+1$ et le polynôme L_j^2 est de degré $2N$ donc*

$$0 \leq \int_a^b L_j^2(x)\,\mathrm{d}x = \sum_{k=0}^{N} \lambda_k L_k^2(x_j) = \lambda_j.$$

Si on utilise une méthode de quadrature composée dans laquelle la méthode de quadrature simple est une méthode de Gauss, la Remarque 5.9 montre que l'erreur est en $O(n^{-(2N+2)})$.

On termine ce chapitre avec le résultat suivant. On a vu qu'une méthode de quadrature simple ne convergeait pas forcément en augmentant le nombre de points (de la même façon que la suite des polynômes d'interpolation de Lagrange d'une fonction f ne converge pas nécessairement vers la fonction f lorsqu'on augmente le nombre de points). Si on choisit la méthode de Gauss on peut cependant montrer, voir [1] - Module VI.1,

Théorème 5.14. *On note $E_N(f)$ l'erreur d'intégration obtenue par la méthode de Gauss à $N+1$ points. Pour toute fonction $f \in C^0([a,b])$ on a $\lim_{N\to\infty} E_N(f) = 0$.*

Bibliographie

[1] Buff X., Garnier J., Halberstadt E., Moulin F., Ramis M., Sauloy J. : *Mathématiques Tout-en-un pour la Licence 2.* Dunod, Paris, 2007.

[2] Crouzeix M., Mignot A.L. : *Analyse Numérique des équations différentielles. Collection Mathématiques Appliquées pour la Maîtrise.* Masson, Paris, 1984.

[3] Demailly J.-P. : *Analyse Numérique et équations différentielles.* Collection Grenoble Sciences, Presse Universitaire de Grenoble, 1991.