Chapitre 1

Résolution des équations non linéaires

1.1 Rappels d'analyse

Théorème 1.1 (de la valeur intermédiaire) *Soit $f : [a,b] \subset \mathbb{R} \to \mathbb{R}$ une fonction continue (i.e. $f \in C^0([a,b])$). Alors, f atteint sa borne inférieure $\min_{x\in[a,b]} f(x)$, sa borne supérieure $\max_{x\in[a,b]} f(x)$ et toute valeur intermédiaire entre les deux bornes.*

Autrement dit : $f([a,b]) = [\min_{x\in[a,b]} f(x), \max_{x\in[a,b]} f(x)]$.

Théorème 1.2 (de Bolzano) *Soit $f : [a,b] \subset \mathbb{R} \to \mathbb{R}$ une fonction continue (i.e. $f \in C^0([a,b])$) telle que $f(a) \cdot f(b) \leq 0$. Alors, il existe (au moins) un nombre réel $c \in [a,b]$ tel que $f(c) = 0$.*

Définition 1.1 *Une suite réelle de Cauchy est, par définition, une suite réelle (x_n) pour laquelle à tout nombre réel ε on peut associer un nombre naturel (un rang) n_0 tel que pour tous entiers $n \geq n_0$ et $m \geq n_0$ on a $|x_n - x_m| < \varepsilon$.*

Théorème 1.3 *Une suite réelle (x_n) est une suite de Cauchy si et seulement si elle est convergente.*

Théorème 1.4 (de Rolle) *Soit une fonction $f : [a,b] \to \mathbb{R}$ qui est continue sur $[a,b]$ (i.e. $f \in C^0([a,b])$) et différentiable sur $]a,b[$.*

Si, de plus, $f(a) = f(b)$, alors il existe (au moins) un point $\xi \in]a,b[$ tel que $f'(\xi) = 0$.

Théorème 1.5 (des accroissements finis) *Soit une fonction $f : [a,b] \to \mathbb{R}$ qui est continue sur $[a,b]$ (i.e. $f \in C^0([a,b])$) et différentiable sur $]a,b[$.*

Alors, il existe (au moins) un point $\xi \in]a,b[$ tel que $f'(\xi) = \dfrac{f(b) - f(a)}{b - a}$.

Théorème 1.6 (Formule de Taylor) *Soit un intervalle ouvert $I \subset \mathbb{R}$, un nombre réel $x_0 \in I$ et une fonction $f : I \to R$ qui est $(n+1)$ fois différentiable sur I.*

Alors, pour tout $x \in I$, il existe un élément ξ, appartenant à l'intervalle ouvert d'extrémités x_0 et x, tel que la relation suivante soit vraie :

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)(x-x_0) + ... + f^{(k)}(x_0)\frac{(x-x_0)^k}{k!} + ... \\ &\quad + f^{(n)}(x_0)\frac{(x-x_0)^n}{n!} + f^{(n+1)}(\xi)\frac{(x-x_0)^{n+1}}{(n+1)!} \\ &= f(x_0) + \sum_{k=1}^{n} f^{(k)}(x_0)\frac{(x-x_0)^k}{k!} + f^{(n+1)}(\xi)\frac{(x-x_0)^{n+1}}{(n+1)!} \end{aligned} \tag{1.1}$$

1.2 Méthodes numériques itératives

Le calcul scientifique, en général, et celui d'ingénieur, en particulier, exigent couramment la résolution de problèmes mathématiques dont les solutions ne peuvent pas être exprimées seulement à l'aide des fonctions élémentaires (comme les fonctions algébriques, trigonométriques, exponentielles ou logarithmiques). Pour une équation polynomiale, par exemple, il n'y a pas de formules générales explicites donnant ses zéros pour un degré plus grand ou égal à 5.

Dans la plupart de tels cas, on fait appel à des méthodes numériques. Ordinairement, la résolution numérique d'un problème "particulier" repose sur un algorithme approprié qui conduit au (bon) résultat après une succession finie d'opérations (élémentaires). Cet algorithme est implémenté à l'aide d'un langage de programmation et exécuté par un ordinateur.

L'idée d'une méthode numérique itérative est de construire une suite de solutions intermédiaires x_n afin d'approcher de manière de plus en plus précise la solution exacte $\bar{x}$ du problème à résoudre.

Définition 1.2 *On dit que la méthode itérative converge si $\lim_{n\to\infty} x_n = \bar{x}$. Sinon, on dit que la méthode itérative diverge.*

Afin d'amorcer un tel procédé répétitif, il faut d'abord choisir ou estimer une valeur (voire plusieurs valeurs) de départ x_0 "convenable(s)". La valeur de départ s'avère souvent déterminante pour la convergence de la méthode itérative.

En réalité, bien que le calcul numérique se fasse aujourd'hui à l'aide de l'ordinateur, on ne peut pas effectuer un nombre infini d'opérations. Par conséquent, même si la suite x_n converge, on se contente généralement de calculer une solution approchée $\hat{x}$. La différence entre la solution exacte et celle approchée représente, en fait, une erreur de calcul e "inévitable" qui est due :

- d'une part, à la troncature imposée par le nombre fini d'opérations effectuées ;

- d'autre part, aux arrondis liés à la représentation des nombres réels dans la mémoire de l'ordinateur.

Définition 1.3 *On définit l'erreur de calcul absolue e_{abs} comme la différence entre la solution exacte et la solution approchée (cette différence étant calculée à l'aide de la valeur absolue, du module, de la norme ou d'une autre mesure, selon le cas) : $e_{abs} = |\bar{x} - \hat{x}|$.*

Définition 1.4 *En outre, si $\bar{x} \neq 0$, on définit l'erreur de calcul relative e_{rel} par la relation* $e_{rel} = \dfrac{e_{abs}}{|\bar{x}|} = \dfrac{|\bar{x} - \hat{x}|}{|\bar{x}|}$.

Dans la pratique, même pour les cas où on peut rendre l'erreur de calcul aussi petite que l'on veut, on se contente souvent d'une valeur approchée "raisonnable" à cause des coûts de calcul induits (à la fois au niveau du temps d'exécution et de la taille de la mémoire vive, voire de la mémoire physique ou externe, utilisée).

Si on s'intéresse à la variation de l'erreur de calcul entre une étape n quelconque et l'étape suivante $n+1$ d'une méthode itérative convergente, on peut distinguer plusieurs type de convergence.

Définition 1.5 *Une méthode est dite convergente d'ordre p, $p \in \mathbb{N}^*$, si (à partir d'un rang $n_0 \in \mathbb{N}^*$, pour tous entiers $n \geq n_0$) $\exists$ une constante (finie non nulle) C t.q.*

$$|\bar{x} - x_{n+1}| \leq C \ |\bar{x} - x_n|^p \tag{1.2}$$

De plus :

- si $p = 1$ et $C < 1$, la convergence est dite linéaire ;
- si $p = 1$ et $C = C_n$ avec $\lim_{n\to\infty} C_n = 0$, la convergence est dite surlinéaire ;
- si $p = 2$, la convergence est dite quadratique ;
- si $p = 3$, la convergence est dite cubique.

La méthode converge d'autant plus rapidement que l'ordre de convergence est élevé.

1.3 Méthodes de dichotomie

Les méthodes de dichotomie sont des méthodes numériques utilisées pour le calcul des zéros des fonctions réelles continues à une variable. Le principe de cette classe de méthodes sera détaillé pour la méthode de bissection. Ensuite, la méthode des parties proportionnelles sera brièvement présentée.

Soit :

- $f : \mathbb{R} \to \mathbb{R}$ une fonction continue $f \in C^0$;
- $\bar{x}$ un zéro de la fonction f, c'est-à-dire une valeur réelle pour laquelle $f(\bar{x}) = 0$ et qui est, donc, solution de l'équation (en général, non linéaire)

$$f(x) = 0 \tag{1.3}$$

Souvent, il est très difficile, voire impossible, de trouver la valeur exacte d'un tel zéro $\bar{x}$. Les méthodes de dichotomie calculent des valeurs approchées du zéro $\bar{x}$, en se basant sur le théorème de Bolzano qui est en fait un corollaire (un cas particulier) du théorème de la valeur intermédiaire.

1.3.1 Méthode de la bissection

La méthode de la bissection est une méthode itérative de recherche des zéros d'une **fonction continue** et elle est basée sur la dichotomie (l'opposition) entre deux sous-intervalles: un sous-intervalle dont on sait qu'il contient au moins un zéro de la fonction et l'autre sous-intervalle dont on ne sait pas dire s'il contient ou pas de zéro de la fonction. Plus précisément, on suit l'approche suivante (voir la Figure 1.1) :

- on considère un "bon" intervalle de départ $[a_0, b_0]$ t.q.

$$f(a_0) \cdot f(b_0) < 0 \tag{1.4}$$

  et la condition ci-dessus (appelée aussi la condition de Bolzano) garantit (grâce à la continuité de la fonction f) l'existence d'au moins un zéro $\bar{x} \in]a_0, b_0[$;

- on construit une suite d'intervalles $[a_n, b_n]$, $n \in \mathbb{N}^*$, contenant le zéro cherché, dont les longueurs deviennent de plus en plus petites $\lim_{n\to\infty} |b_n - a_n| = 0$;
- à l'étape n, si le zéro n'est pas le milieu de l'intervalle $]a_n, b_n[$, on divise cet intervalle en deux sous-intervalles de même longueur et on choisit comme intervalle $[a_{n+1}, b_{n+1}]$ celui où la fonction f change de signe, c'est-à-dire :
  - on calcule la valeur approchée x_n du zéro cherché par la relation $x_n = \dfrac{a_n + b_n}{2}$;
  - on calcule la valeur $f(x_n)$;
  - si $f(x_n) = 0$, alors $\bar{x} = x_n$ et on s'arrête (car la valeur exacte du zéro a été trouvée) ;
  - autrement, si $f(x_n) \neq 0$ et $f(a_n) \cdot f(x_n) < 0$, on pose $a_{n+1} = a_n$ et $b_{n+1} = x_n$;
  - autrement (c'est-à-dire si $f(x_n) \neq 0$ et $f(x_n) \cdot f(b_n) < 0$), on pose $a_{n+1} = x_n$ et $b_{n+1} = b_n$;

- dans les deux derniers cas, on passe à l'étape suivante $n+1$ pour l'intervalle $[a_{n+1}, b_{n+1}]$ qui contient forcément le zéro $\bar{x}$;
- le calcul itératif s'arrête soit quand la valeur exacte du zéro $\bar{x}$ est trouvée, soit quand une certaine condition d'arrêt (qui sera précisée plus tard) est remplie.

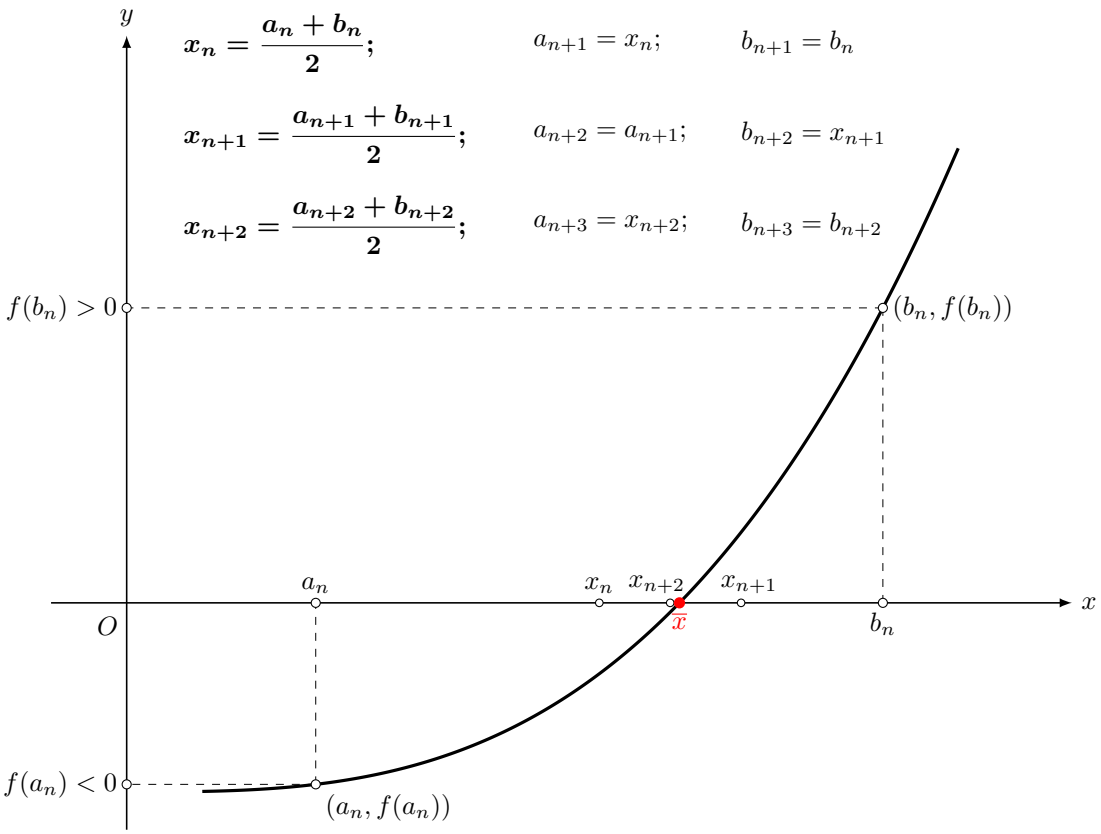

Figure 1.1: Méthode de bissection

Algorithme de calcul (Algorithme de Bolzano)

Notations recommandées :

- a et b variables réelles qui correspondent initialement aux bornes de l'intervalle $]a; b[$ contenant (au moins) un zéro de la fonction continue f ;
- *xbar* variable réelle qui correspond initialement au milieu de l'intervalle $[a; b]$;
- *eps* constante réelle qui représente la valeur d'une tolérance (positive) donnée (qui doit tenir compte de la précision de la machine).

Proposition 1.1 *Si l'intervalle de départ $[a_0, b_0]$ respecte la condition (1.4), la méthode de la bissection est convergente.*

Dém. A chaque étape n, la longueur de l'intervalle $[a_n, b_n]$ est $|b_n - a_n| = \dfrac{|b_0 - a_0|}{2^n}$.

Ainsi, $\lim\limits_{n\to\infty} |b_n - a_n| = \lim\limits_{n\to\infty} \dfrac{|b_0 - a_0|}{2^n} = 0$.
Mais, $\forall n \in \mathbb{N}$, $x_n \in]a_n, b_n[$ et $\bar{x} \in]a_n, b_n[$.
Par conséquent, $\lim\limits_{n\to\infty} x_n = \bar{x}$ et, donc, la méthode de bissection est convergente. ■

En outre, à chaque étape n, l'erreur absolue commise $e_n = |\bar{x} - x_n|$ peut être majorée par la relation

$$e_n = |\bar{x} - x_n| < \frac{|b_n - a_n|}{2} = \frac{|b_0 - a_0|}{2^{n+1}} \tag{1.5}$$

Ceci nous permet de calculer un nombre minimal d'itérations à effectuer $n_{\min}$ afin de s'assurer que le zéro cherché est approché avec une tolérance demandée ε. En imposant $e_n < \varepsilon$, on obtient :

$$n_{\min} > \log_2 \left(\frac{|b_0 - a_0|}{\varepsilon} \right) - 1 \tag{1.6}$$

Etant donnée que $\dfrac{1}{2^4} < \dfrac{1}{10} < \dfrac{1}{2^3}$, on peut estimer qu'on a besoin de 3 à 4 itérations supplémentaires afin d'obtenir une décimale de plus dans la valeur approchée du zéro cherché par la méthode de la bissection.

Remarque

Si la fonction f possède plusieurs zéros sur l'intervalle $]a, b[$, la méthode n'en calcule qu'un seul. Afin de palier cet inconvénient, on divise l'intervalle d'étude $]a, b[$ en plusieurs sous-intervalles pour lesquels on applique individuellement la méthode de la bissection (démarche qui est appelée la **méthode de la bissection par intervalles**).

Avantages de la méthode de bissection

- La méthode n'impose aucune hypothèse supplémentaire concernant la fonction f (à part sa continuité).

- La méthode converge "toujours", à condition que l'intervalle de départ $[a, b]$ respecte la condition $f(a) \cdot f(b) < 0$.

Désavantages de la méthode de bissection

- La convergence de la méthode est relativement lente.

- Le méthode ne peut pas être généralisée pour $\mathbb{R}^n$.

1.3.2 *Méthode des parties proportionnelles (méthode de la sécante)

Comme pour la méthode de bissection, dans la méthode des parties proportionnelles, on construit une suite d'intervalles qui contiennent le zéro cherché et dont les longueurs tendent vers 0. Le "bon" 'intervalle de départ vérifie toujours la condition de Bolzano.

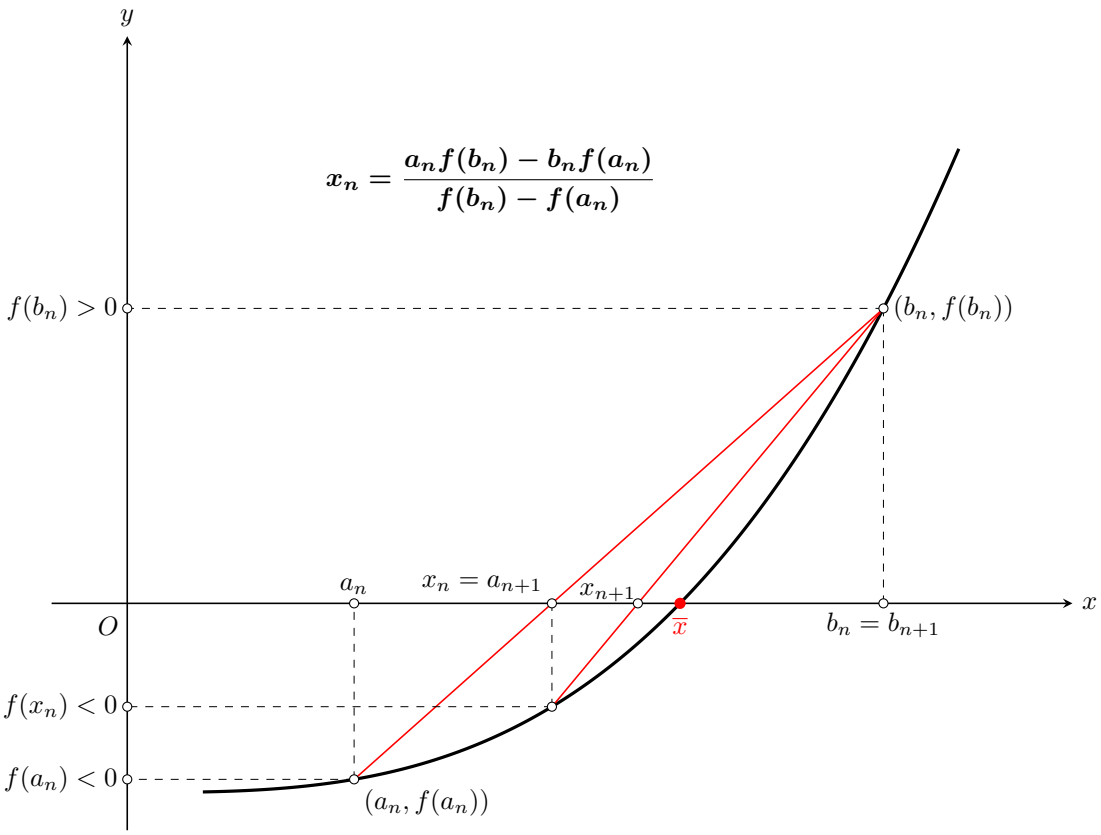

Figure 1.2: Méthode des parties proportionnelles

Cependant (voir la Figure 1.2), à l'étape courante n, la valeur approchée x_n n'est plus (forcément) le milieu de l'intervalle $]a_n, b_n[$ car, cette fois, on choisit x_n comme étant le point d'intersection entre l'axe des abscisses et la corde passant par les points $(a_n, f(a_n))$ et $(b_n, f(b_n))$. Par conséquent, grâce à la relation de proportionnalité $\frac{f(b_n)}{b_n - x_n} = \frac{f(a_n)}{a_n - x_n}$, on obtient :

$$x_n = \frac{a_n f(b_n) - b_n f(a_n)}{f(b_n) - f(a_n)} \tag{1.7}$$

Par la suite, le calcul de la valeur approchée du zéro $\bar{x}$ continue comme dans la méthode de la bissection.

La méthode des parties proportionnelles converge, souvent, plus vite que la méthode de la bissection mais le comportement des deux méthodes reste assez semblable.

1.4 Méthodes de point fixe

A part les méthodes de dichotomie, les méthodes de point fixe représentent une autre classe de méthodes numériques utilisées pour le calcul des zéros d'une fonction (non linéaire).

1.4.1 Approche générale - Méthode de Picard

Définition 1.6 *Soit une fonction $g : \mathbb{R} \to \mathbb{R}$ et $\tilde{x} \in \mathbb{R}$ t.q. $g(\tilde{x}) = \tilde{x}$. Le nombre réel $\tilde{x}$ est appelé un point fixe de la fonction g.*

Il convient de remarquer que l'image d'un point fixe d'une fonction par cette fonction est le point fixe lui-même.

Soit la fonction (non linéaire) $f : \mathbb{R} \to \mathbb{R}$ et l'équation

$$f(x) = 0 \tag{1.8}$$

Afin d'utiliser une méthode de point fixe pour résoudre cette équation (1.8), il faut la transformer d'abord en une équation équivalente (i.e. ayant les mêmes solutions) de la forme :

$$x = g(x) \tag{1.9}$$

où la fonction g est appelée **fonction d'itération**.

Sans être unique, une telle transformation est toujours envisageable (et, en fait, on peut trouver une infinité de fonctions g convenables). Un choix immédiat peut être, par exemple :

$$g(x) = x + \lambda \cdot f(x) \tag{1.10}$$

où λ est un coefficient réel non nul (voire une fonction réelle de x qui ne s'annule pas).

Le but de la transformation (1.9) est de remplacer la recherche d'une solution $\bar{x}$ de l'équation (1.8) (qui est donc un zéro de la fonction f) par la recherche du point fixe équivalent $\tilde{x} = \bar{x}$ de la fonction g (et, par la suite, le point fixe de g sera noté, directement, $\bar{x}$). Plus précisément, une méthode de point fixe est une méthode itérative basée sur l'approche suivante :

- on considère une "bonne" approximation de départ x_0 du point fixe $\bar{x}$ de g ;
- on construit une suite récurrente (x_n), $n \in \mathbb{N}$, (censée s'approcher de plus en plus du point fixe $\bar{x}$) définie de la manière suivante :

$$x_{n+1} = g(x_n) \tag{1.11}$$

- on arrête le calcul itératif quand une certaine condition d'arrêt (qui sera précisée plus tard) est remplie (ou quand la valeur exacte du point fixe $\bar{x}$ est trouvée).

Malheureusement, la convergence de cette approche (appelée couramment la **méthode de Picard**) n'est pas, en général, garantie et elle dépend à la fois du choix de la fonction d'itération g et de la valeur de départ x_0 (voir les Figures 1.3, 1.4, 1.5).

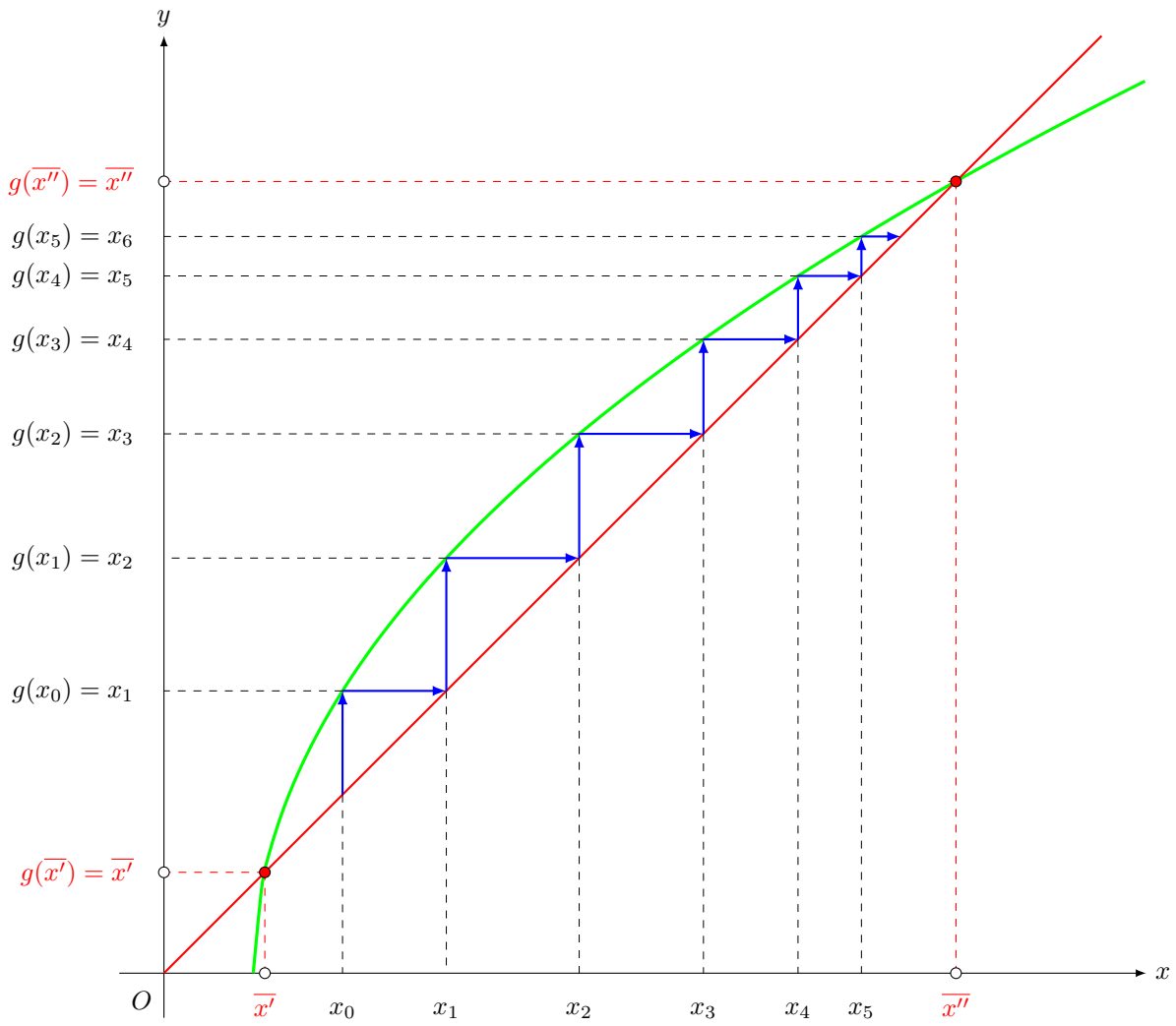

Figure 1.3: Point fixe attracteur (suite monotone et convergente)

Cependant, dans certaines conditions (comme celles précisées dans la proposition suivante, par exemple), la convergence de la méthode de Picard est assurée.

Proposition 1.2 *Si, dans la relation (1.11), la fonction g est continue ($g \in C^0$) et la suite $(x_n)_{n=0}^{\infty}$ est convergente vers une limite l, alors la méthode de Picard converge et l est forcément un point fixe de g.*

Dém. $l = \lim\limits_{n\to\infty} x_n = \lim\limits_{n\to\infty} x_{n+1} \overset{(1.11)}{=} \lim\limits_{n\to\infty} g(x_n) \overset{continuité}{=} g(\lim\limits_{n\to\infty} x_n) = g(l)$ ■

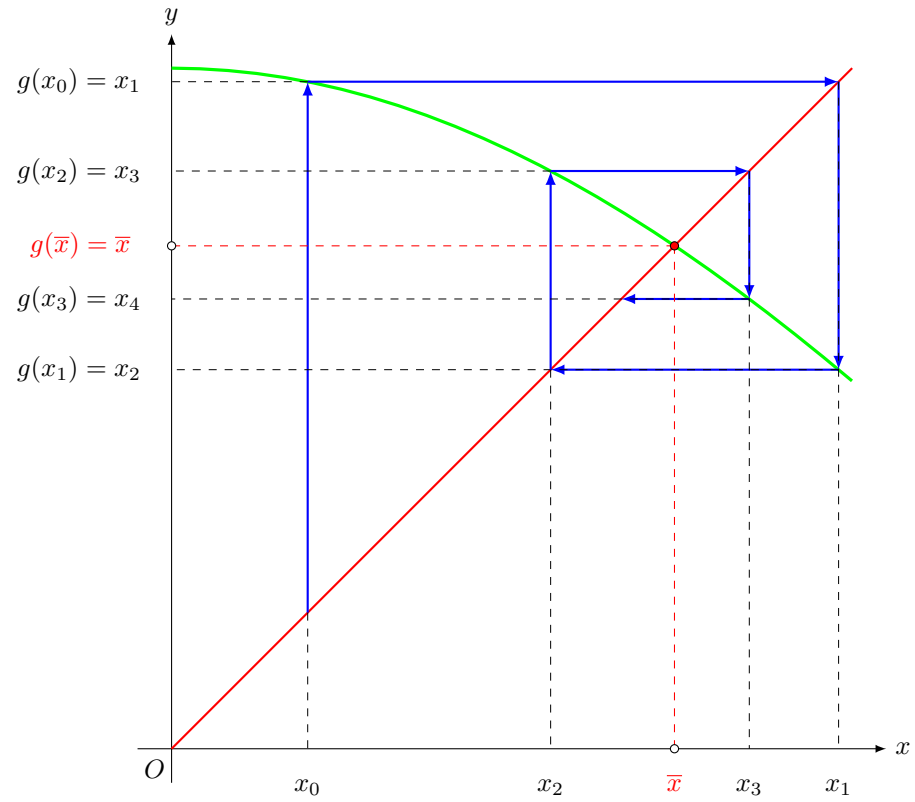

Figure 1.4: Point fixe attracteur (suite alternée et convergente)

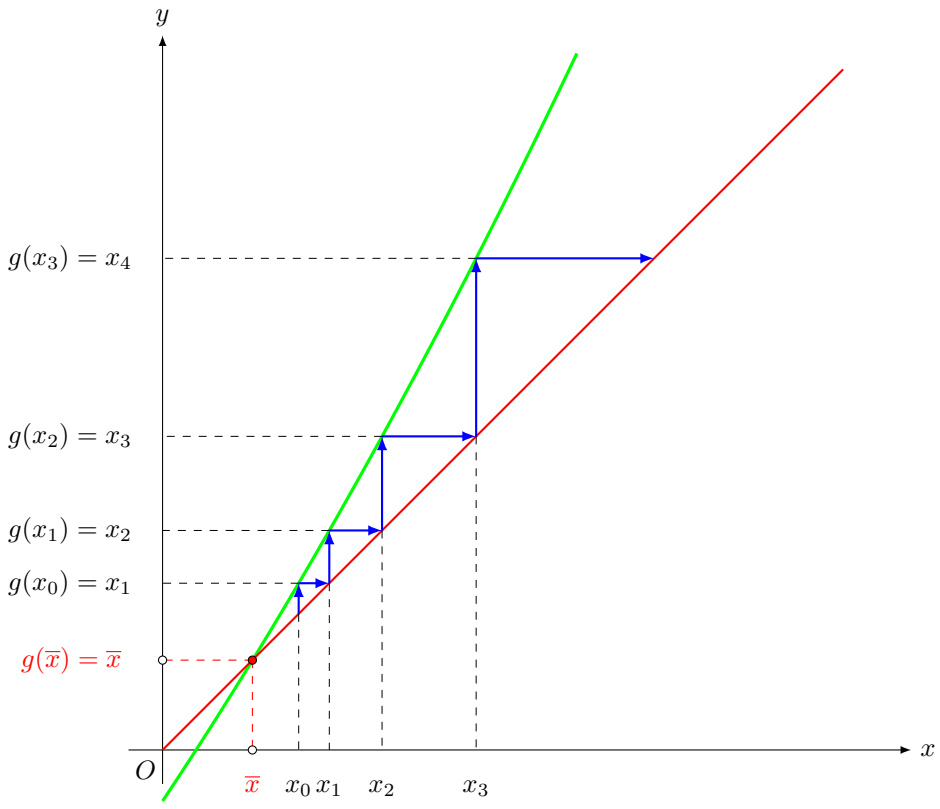

Figure 1.5: Point fixe répulsif (suite monotone et divergente)

En général, on préfère savoir d'abord si un point fixe existe et entamer ensuite sa recherche par une méthode itérative. Dans ce sens, on présente ci-dessous une condition suffisante pour qu'une fonction continue ait un point fixe.

Proposition 1.3 *Soit* $g : [a,b] \subset \mathbb{R} \to [a,b]$ *une fonction continue* ($g \in C^0([a,b])$). *Alors, la fonction* g *admet au moins un point fixe* $\bar{x} \in [a,b]$.

Dém. On considère la fonction auxiliaire $h : [a,b] \to \mathbb{R}$ définie par $h(x) = g(x) - x$.

Ainsi, $h(a) \cdot h(b) = (g(a) - a) \cdot (g(b) - b) \leq 0$.

Comme cette dernière relation est justement la condition de Bolzano pour la fonction continue h, on peut en déduire (grâce au théorème de Bolzano) qu'il existe au moins un nombre réel $\bar{x} \in [a,b]$ tel que $h(\bar{x}) = 0$.

Il s'ensuit que $g(\bar{x}) - \bar{x} = 0$ ou, encore, $g(\bar{x}) = \bar{x}$. ■

Par la suite, on présente d'autres situations où la convergence des méthodes de point fixe peut être démontrée.

Définition 1.7 *Soit un nombre réel positif* $K \in \mathbb{R}_+$ *et une fonction réelle* $g : I \subset \mathbb{R} \to \mathbb{R}$. *La fonction* g *est dite* K*-lipschitzienne (ou lipschitzienne dans le rapport* K*) sur l'intervalle* I *si pour tout couple d'éléments* $\forall x, y \in I$: $|g(x) - g(y)| \leq K\,|x - y|$.

Définition 1.8 *Une fonction lipschitzienne dans un rapport* $K < 1$ *est dite* K*-contractante (ou contractante dans le rapport* $K < 1$*).*

Proposition 1.4 *Si une fonction* $g : I \subset \mathbb{R} \to \mathbb{R}$ *est* K*-contractante sur son domaine de définition, alors elle est (forcément) continue.*

Dém. On considère un point quelconque x du domaine de définition de g. Pour toute suite (x_n) d'éléments de I qui converge vers x, on a (grâce à la propriété de contraction) que $|g(x) - g(x_n)| \leq K\,|x - x_n|$.

Mais $\lim_{n\to\infty} x_n = x$ et donc $\lim_{n\to\infty} |\ g(x) - g(x_n)\ | = 0$ ou encore $\lim_{n\to\infty} g(x_n) = g(x)$.

Ainsi, vu la définition de la continuité des fonctions, on vient de prouver le résultat énoncé. ■

Théorème 1.7 (Théorème du point fixe de Banach) *Soit* $g : \mathbb{R} \to \mathbb{R}$ *une fonction* K*-contractante. Alors* g *a un et un seul point fixe* $\bar{x}$ *dans* $\mathbb{R}$.

Remarque 1.1 *La preuve du théorème du point fixe de Banach est similaire à celle détaillée pour le théorème ci-dessous.*

Théorème 1.8 *Soit une fonction d'itération* $g : [a,b] \subset \mathbb{R} \to \mathbb{R}$ *qui est* K*-contractante.*

Si, de plus, $g(\ [a,b]\ ) \subset [a,b]$ *(i.e.* $g(x) \in [a,b]\ \forall x \in [a,b]$*), alors :*

1. *g a un et un seul point fixe $\bar{x} \in [a, b]$;*

2. *$\forall x_0 \in [a, b]$, la suite récurrente (x_n), $n \in \mathbb{N}$, (correspondant à la méthode de Picard) donnée par la relation :*

$$x_{n+1} = g(x_n) \tag{1.12}$$

est convergente vers l'unique point fixe de g, i.e.

$$\lim_{n\to\infty} x_n = \bar{x} \tag{1.13}$$

3. *la convergence de la suite (x_n) est (au moins) linéaire.*

Dém. On considère un point quelconque du domaine de définition $x_0 \in [a, b]$ et une suite récurrente de nombres réels $(x_n)_{n=0}^{\infty}$ définie par la relation (1.12).

Comme, par hypothèse, l'image d'un point du domaine de définition $[a, b]$ reste dans $[a, b]$, il s'ensuit que tous les termes de la suite (x_n) sont bien dans l'intervalle $[a, b]$.

Par conséquent, on peut utiliser (une première fois) la propriété de contraction de g en écrivant, pour $n \geq 1$:

$$| \, x_{n+1} - x_n \, | = | \, g(x_n) - g(x_{n-1}) \, | \leq K \, | \, x_n - x_{n-1} \, |$$

ou, en itérant encore $(n-1)$ fois :

$$| \, x_{n+1} - x_n \, | \leq K^n \, | \, g(x_0) - x_0 \, |$$

Ainsi, pour tout couple $n, m \in \mathbb{N}$ avec $m > n$, on a :

$$\begin{aligned} | \, x_m - x_n \, | &\leq | \, x_m - x_{m-1} \, | + | \, x_{m-1} - x_{m-2} \, | + ... + | \, x_{n+1} - x_n \, | \\ &\leq K^{m-1} \, | \, g(x_0) - x_0 \, | + ... + K^n \, | \, g(x_0) - x_0 \, | \\ &= K^n (1 + ... + K^{m-n-1}) \, | \, g(x_0) - x_0 \, | \\ &\leq \frac{K^n}{1-K} \, | \, g(x_0) - x_0 \, | \end{aligned}$$

Or, cette dernière inégalité prouve que la suite (x_n) est une suite de Cauchy (car, par hypothèse, $K < 1$) et, donc, elle est convergente vers une limite que l'on note (en anticipant la suite de la démonstration) $\bar{x}$.

Etant donné que l'intervalle $[a, b]$ est fermé, on a que $\bar{x} \in [a, b]$.

Par la suite, on utilise la continuité de la fonction g, assurée par la proposition (1.4), et on écrit :

$$\bar{x} = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} g(x_n) = g(\lim_{n\to\infty} x_n) = g(\bar{x})$$

On vient ainsi de prouver que la fonction d'itération admet un point fixe qui est, de plus, la limite de la suite convergente (x_n).

Une fois l'existence du point fixe démontrée, on va prouver maintenant son unicité, en supposant que g admet un autre point fixe $\bar{y}$. Dans ces conditions, on calcule :

$$| \bar{x} - \bar{y} | = | g(\bar{x}) - g(\bar{y}) | \leq K | \bar{x} - \bar{y} |$$

Mais, comme $K < 1$, cette dernière inégalité est satisfaite ssi $\bar{y} = \bar{x}$, ce qui prouve l'unicité du point fixe $\bar{x}$.

Il nous reste à étudier la nature de la convergence de la suite (x_n) et on calcule :

$$| \bar{x} - x_{n+1} | = | g(\bar{x}) - g(x_n) | \leq K | \bar{x} - x_n |$$

Conformément à la définition (1.5), cette dernière relation montre que la convergence de la suite (x_n) (et donc de la méthode de Picard) est linéaire. ■

Dans le théorème ci-dessous, on considère que la fonction d'itération admet un point fixe et on précise des conditions (suffisantes) pour que la méthode de Picard qui lui est associée soit convergente.

Théorème 1.9 *Soit une fonction d'itération $g : \mathbb{R} \to \mathbb{R}$ qui est (une fois) continûment différentiable ($g \in C^1$) et qui admet un point fixe $\bar{x}$ (i.e. $g(\bar{x}) = \bar{x}$).*

Si, de plus, $| g'(\bar{x}) | < 1$, alors :

1. *$\exists \varepsilon > 0$ tel que si on choisit $x_0 \in [\bar{x} - \varepsilon, \bar{x} + \varepsilon]$, alors la suite récurrente (x_n), $n \in \mathbb{N}$, (correspondant à la méthode de Picard) donnée par la relation $x_{n+1} = g(x_n)$ est convergente vers le point fixe $\bar{x}$ (i.e. $\lim_{n\to\infty} x_n = \bar{x}$) ;*
2. *la convergence de la suite (x_n) est (au moins) linéaire.*

Dém. En d'autres termes, ce théorème affirme que si la fonction d'itération g satisfait les hypothèses précisées, alors il existe un voisinage V du point fixe $\bar{x}$ tel que pour toute valeur de départ $x_0 \in V$, la suite récurrente (x_n) est convergente vers le point fixe.

Vu que, par hypothèse, la fonction dérivée g' est continue et que $| g'(\bar{x}) | < 1$, il existe deux nombres réels $\varepsilon > 0$ et $0 < K < 1$ tels que, $\forall x \in V = [\, \bar{x} - \varepsilon, \bar{x} + \varepsilon]$, on a :

$$| g'(x) | \leq K$$

De plus, grâce au théorème des accroissements finis, on sait que pour tout couple de nombres réels $x, y \in V$ avec $y < x$, il existe $\xi \in]\, y, x \,[$ tels que :

$$g(x) - g(y) = g'(\xi)\,(x - y\ )$$

Ceci nous permet d'introduire les majorations suivantes :

$$| g(x) - g(y)| \leq \max_{\zeta \in V} |g'(\zeta)| \, | x - y \, | \leq K \, | x - y \, |$$

Cette dernière inégalité montre, en fait, que la fonction g est K-contractante sur le voisinage V du point fixe $\bar{x}$.

En outre, si on pose $y = \bar{x}$, on a (grâce à la dernière inégalité) :

$$| g(x) - g(\bar{x}) \, | = | g(x) - \bar{x} \, | \leq K \, | x - \bar{x} \, | \leq | x - \bar{x} \, | \leq \varepsilon$$

On vient de montrer ainsi que $\forall x \in V : g(x) \in V$ et, donc, $g(V) \subset V$.

On peut maintenant constater qu'il existe (toujours) un voisinage V du point fixe $\bar{x}$ pour lequel la fonction d'itération g respecte bien les hypothèses du théorème (1.8).

Ceci nous permet de conclure immédiatement que la suite récurrente (correspondant à la méthode de Picard) définie avec une valeur de départ choisie dans ce voisinage V du point fixe est convergente vers le point fixe et que la convergence est (au moins) linéaire. ■

1.4.2 Condition d'arrêt

On considère une méthode de Picard convergente vers un point fixe $\bar{x}$ d'une fonction d'itération g suffisamment régulière (par exemple, $g \in C^1$ dans un voisinage V du point fixe). Afin d'estimer l'erreur commise à l'étape n de la méthode de point fixe, on calcule (en utilisant le théorème des accroissements finis) :

$$\bar{x} - x_{n+1} = g(\bar{x}) - g(x_n) = g'(\xi_n) \, (\bar{x} - x_n) \tag{1.14}$$

où ξ_n appartient à l'intervalle (ouvert) d'extrémités $\bar{x}$ et x_n.

Etant donné l'identité

$$\bar{x} - x_n = (\bar{x} - x_{n+1}) + (x_{n+1} - x_n)$$

la relation (1.14) nous permet d'obtenir une expression pour l'erreur (absolue) à l'étape n :

$$\bar{x} - x_n = \frac{1}{1 - g'(\xi_n)} \, (x_{n+1} - x_n) \tag{1.15}$$

Cette dernière relation explique pourquoi, en général, on arrête une méthode de Picard après un nombre minimal d'itérations $n_{\min}$ quand la valeur absolue de la différence entre deux approximations successives du point fixe devient inférieure à une tolérance donnée ε :

$$|x_{n_{\min}} - x_{n_{\min}-1}| < \varepsilon \tag{1.16}$$

Cependant, il convient de remarquer le fait qu'une telle condition d'arrêt est bien adéquate si $g'(\xi_{n_{\min}-1}) \simeq 0$ (comme dans le cas de la méthode de Newton), mais elle est d'autant moins appropriée que $g'(\xi_{n_{\min}-1})$ se trouve proche de 1.

Les considérations présentées jusqu'ici portent sur l'ensemble des méthodes de point fixe. Par la suite, on passe en revue quelques méthodes qui utilisent une approche géométrique pour la résolution des équations non linéaires et qui s'avèrent appartenir à la classe des méthodes de point fixe.

1.4.3 Méthode de Newton (Méthode de Newton-Raphson ou de la tangente)

On considère une fonction $f(x) : \mathbb{R} \to \mathbb{R}$ qui est (une fois) continûment différentiable ($f \in C^1$) et qui admet un **zéro simple** $\bar{x}$ (i.e. $f(\bar{x}) = 0$ et $f'(\bar{x}) \neq 0$).

Afin de trouver ce zéro de f, la **méthode de Newton** (appelée aussi la **méthode de Newton-Raphson** ou la **méthode de la tangente**) propose la démarche itérative suivante :

- on considère une "bonne" approximation de départ x_0 du zéro simple cherché $\bar{x}$;
- à l'étape n, $n \in \mathbb{N}$, on calcule la nouvelle valeur approchée x_{n+1} comme étant l'abscisse du point d'intersection de la tangente au graphique de la fonction f passant par le point $(x_n, f(x_n))$ avec l'axe des abscisses ;
- on arrête le calcul itératif quand une certaine condition d'arrêt (qui sera précisée plus tard) est remplie (ou quand la valeur exacte du zéro $\bar{x}$ est trouvée).

Proposition 1.5 *La méthode de Newton est une méthode de point fixe.*

Dém. Etant donné la construction graphique indiquée dans la Figure 1.6, on peut écrire à l'étape n :

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

ce qui nous donne tout de suite :

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{1.17}$$

Or, cette dernière relation est valable pour $n \in \mathbb{N}$ et définit une suite récurrente de la forme

$$x_{n+1} = g(x_n) \tag{1.18}$$

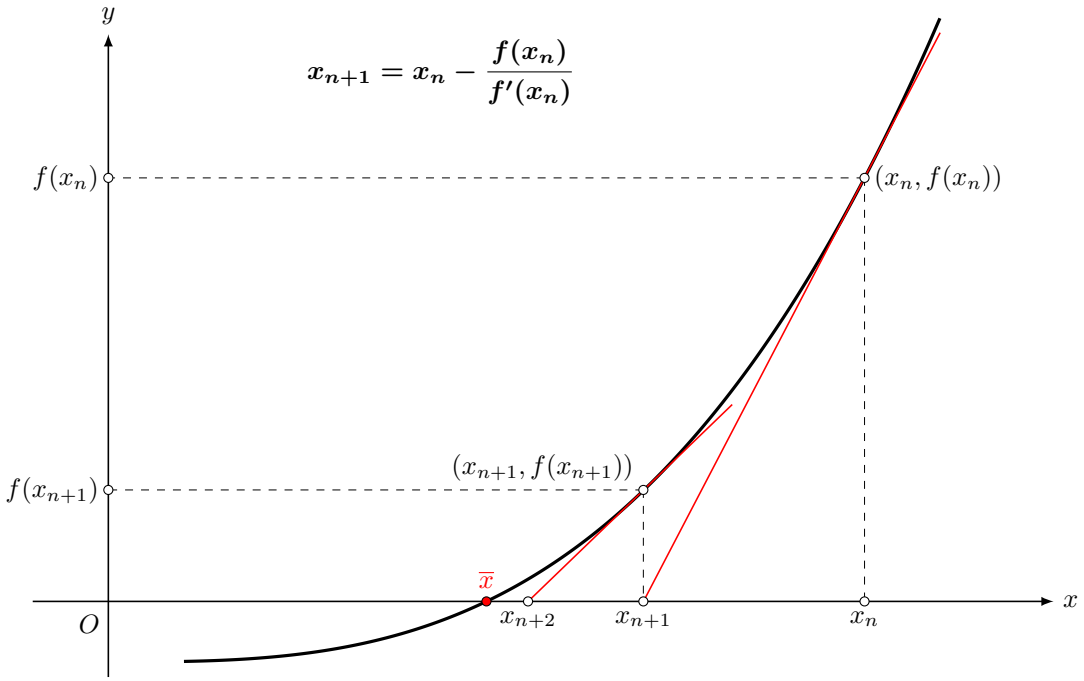

Figure 1.6: Méthode de Newton

pour la fonction d'itération :

$$g(x) = x - \frac{f(x)}{f'(x)} \tag{1.19}$$

En considérant un voisinage V d'un zéro (simple) $\bar{x}$ tel que $f'(x) \neq 0 \; \forall x \in V$, on peut donc affirmer qu'appliquer la méthode de Newton pour trouver le zéro $\bar{x}$ de la fonction f revient à appliquer la méthode de Picard pour trouver le point fixe équivalent $\bar{x}$ de la fonction d'itération g donnée par la relation (1.19). ■

La proposition (1.5) nous permet maintenant d'utiliser le théorème (1.9) dans le cas particulier de la méthode de Newton. On obtient ainsi le résultat suivant :

Théorème 1.10 *Soit une fonction $f : \mathbb{R} \to \mathbb{R}$ qui est deux fois continûment différentiable ($f \in C^2$).*

Si f admet un zéro simple $\bar{x}$ (i.e. $f(\bar{x}) = 0$ et $f'(\bar{x}) \neq 0$), alors :

1. *$\exists \varepsilon > 0$ tel que si on choisit $x_0 \in [\bar{x} - \varepsilon, \bar{x} + \varepsilon]$, alors la suite récurrente (x_n), $n \in \mathbb{N}$, correspondant à la méthode de Newton et donnée par la relation (1.17), est convergente vers le zéro $\bar{x}$ (i.e. $\lim_{n\to\infty} x_n = \bar{x}$) ;*
2. *la convergence de la suite (x_n) est quadratique.*

Dém. A partir de la relation (1.19), on calcule la dérivée g' de la fonction d'itération :

$$g'(x) = 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} \tag{1.20}$$

Au point $\bar{x}$ (qui est un point fixe de g et un zéro simple de f), on a $f(\bar{x}) = 0$ et $f'(\bar{x}) \neq 0$. Ainsi :

$$g'(\bar{x}) = 0 \tag{1.21}$$

Cette dernière condition implique que $|g'(\bar{x})| < 1$.

De plus, la fonction g est bien continûment différentiable (au moins dans un voisinage V du zéro simple $\bar{x}$ pour lequel $f'(x) \neq 0$).

Ainsi, toutes les hypothèses du théorème (1.9) sont remplies et on peut conclure que la suite récurrente (1.17) obtenue par la méthode de Newton est bien convergente vers le zéro simple $\bar{x}$, avec une convergence au moins linéaire.

En outre, on va montrer que cette convergence est en fait quadratique. Pour cela, on écrit (à l'étape n de la méthode de Newton) le développement de la fonction $f \in C^2$ autour du point x_n :

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{f''(\xi_n)}{2}(x - x_n)^2$$

où ξ_n appartient à l'intervalle (ouvert) d'extrémités x et x_n.

En imposant $x = \bar{x}$ dans cette dernière égalité et en tenant compte que $f(\bar{x}) = 0$, on obtient :

$$0 = f(x_n) + f'(x_n)(\bar{x} - x_n) + \frac{f''(\xi_n)}{2}(\bar{x} - x_n)^2$$

Etant donné qu'il existe un voisinage V du point fixe $\bar{x}$ tel que $f'(x_n) \neq 0$, on divise l'égalité ci-dessus par $f'(x_n)$:

$$0 = \frac{f(x_n)}{f'(x_n)} + \bar{x} - x_n + \frac{f''(\xi_n)}{2f'(x_n)}(\bar{x} - x_n)^2$$

et on met en évidence le terme suivant de la suite de Newton, x_{n+1}, grâce à la relation (1.17) :

$$0 = \bar{x} - x_{n+1} + \frac{f''(\xi_n)}{2f'(x_n)}(\bar{x} - x_n)^2$$

Il s'ensuit que :

$$|\bar{x} - x_{n+1}| = \frac{|f''(\xi_n)|}{2\,|f'(x_n)|}\,|\bar{x} - x_n|^2$$

Vu que $|\bar{x} - x_{n+1}|$ représente l'erreur absolue à l'étape $n+1$ et $|\bar{x} - x_n|$ l'erreur absolue à l'étape précédente n, on note :

$$C = \frac{\max\limits_{x \in V} |f''(x)|}{2 \min\limits_{x \in V} |f'(x)|} \tag{1.22}$$

et on obtient :

$$|\bar{x} - x_{n+1}| \leq C\,|\bar{x} - x_n|^2 \tag{1.23}$$

Or, conformément à la définition (1.5), cette dernière inégalité prouve que la convergence de la suite récurrente correspondant à la méthode de Newton est bien quadratique. ■

Dans la démonstration que l'on vient de finir, l'hypothèse que le zéro $\bar{x}$ soit simple, i.e. $f'(\bar{x}) \neq 0$, n'intervient, d'une manière essentielle, que dans la partie concernant la vitesse de convergence quadratique. Par conséquent, on peut prouver que si $f \in C^2$ et $f'(x) \neq 0$ dans un voisinage V d'un zéro $\bar{x}$, alors la méthode de Newton reste encore convergente même si $f'(\bar{x}) = 0$. Par contre, la convergence est seulement linéaire.

D'une manière plus générale, la convergence de la méthode de Newton est seulement linéaire pour des zéros de multiplicité $m > 1$. Même dans ces cas, on peut passer à une convergence quadratique en utilisant une variante de la méthode de Newton dite méthode de Newton adaptative.

Point de départ et condition d'arrêt

Comme toutes les méthodes de point fixe, la méthode de Newton est sensible au choix de la valeur de départ x_0 et, d'une manière générale, x_0 doit être dans un voisinage suffisamment petit du zéro cherché (ce qui assure, sous certaines conditions, comme celles précisées dans le théorème (1.10), une vitesse de convergence quadratique). En pratique, on peut commencer la recherche d'un zéro d'une fonction avec une méthode moins rapide, d'habitude une méthode de dichotomie, qu'on remplace par la suite avec la méthode de Newton, dès qu'on considère être arrivé suffisamment proche du zéro cherché.

Vu que la méthode de Newton est une méthode de point fixe, on peut arrêter le calcul itératif après un nombre minimal d'itérations $n_{\min}$, dès que la relation (1.16) est satisfaite:

$$|x_{n_{\min}} - x_{n_{\min}-1}| < \varepsilon \tag{1.24}$$

où ε est une certaine tolérance exigée. De plus, si on cherche un zéro simple $\bar{x}$ de f, cette condition d'arrêt est bien appropriée car la dérivée de la fonction d'itération correspondante $g'(x)$ reste presque nulle dans un voisinage V du $\bar{x}$.

Cependant, on peut imaginer aussi une condition d'arrêt tenant compte du résidu r_n à l'étape n. Etant donné qu'on cherche un zéro $\bar{x}$, le résidu à l'étape n est donné simplement par la valeur $f(x_n)$, car $f(\bar{x}) = 0$. Ainsi, on peut arrêter le calcul itératif après un nombre minimal d'itérations $n_{\min}$, dès que la condition suivante est remplie :

$$|r_{n_{\min}}| = |f(x_{n_{\min}})| < \varepsilon \tag{1.25}$$

où ε est une tolérance donnée.

Pourtant, la condition d'arrêt basée sur le résidu s'avère adéquate seulement si $|f'(x)| \simeq 1$ dans un voisinage V du zéro cherché. Pour les cas où $|f'(x)| \gg 1$, on surestime l'erreur,

tandis que pour les cas où $|f'(x)| \ll 1$, on sous-estime l'erreur.

Avantages de la méthode de Newton

- La vitesse de convergence est élevée, à condition que la valeur de départ x_0 soit suffisamment proche du zéro cherché $\bar{x}$.
- La méthode peut être généralisée pour des systèmes d'équations non linéaires.

Limitations de la méthode de Newton

- La fonction f doit être assez régulière (en général $f \in C^2$).
- Afin d'assurer une convergence quadratique, le zéro cherché $\bar{x}$ doit être simple (i.e. $f(\bar{x}) = 0$ et $f'(\bar{x}) \neq 0$).
- A chaque étape n, il faut calculer la dérivée au point courant $f'(x_n)$: soit grâce au calcul (préalable) de la fonction dérivée si on connaît l'expression (analytique) de la fonction f, soit numériquement si la fonction f est connue de manière discrète sur un ensemble fini des points (grâce à des valeurs obtenues expérimentalement ou suite à la modélisation discrète d'un problème physique donné, par exemple).

1.4.4 **Méthode de la corde (Méthode de Newton-corde)*

On considère une fonction $f(x) : \mathbb{R} \to \mathbb{R}$ qui est (une fois) continûment différentiable ($f \in C^1$) et qui admet un zéro simple $\bar{x}$ (i.e. $f(\bar{x}) = 0$ et $f'(\bar{x}) \neq 0$).

Afin de trouver ce zéro de f, la **méthode de la corde**, appelée aussi la **méthode de Newton-corde**, propose la démarche itérative suivante :

- on considère une "bonne" approximation de départ x_0 du zéro simple cherché $\bar{x}$ et on calcule la pente $f'(x_0) \neq 0$ de la tangente au graphique de la fonction f passant par le point $(x_0, f(x_0))$;
- à l'étape n, $n \in \mathbb{N}$, on calcule la nouvelle valeur approchée x_{n+1} comme étant l'abscisse du point d'intersection de la droite de pente $f'(x_0)$ passant par le point $(x_n, f(x_n))$ avec l'axe des abscisses ;
- on arrête le calcul itératif quand une certaine condition d'arrêt (qui sera précisée plus tard) est remplie (ou quand la valeur exacte du zéro $\bar{x}$ est trouvée).

On peut constater que la méthode de la corde est très similaire à la méthode de Newton, tout en évitant le calcul de la dérivée $f'(x_n)$ à chaque étape n.

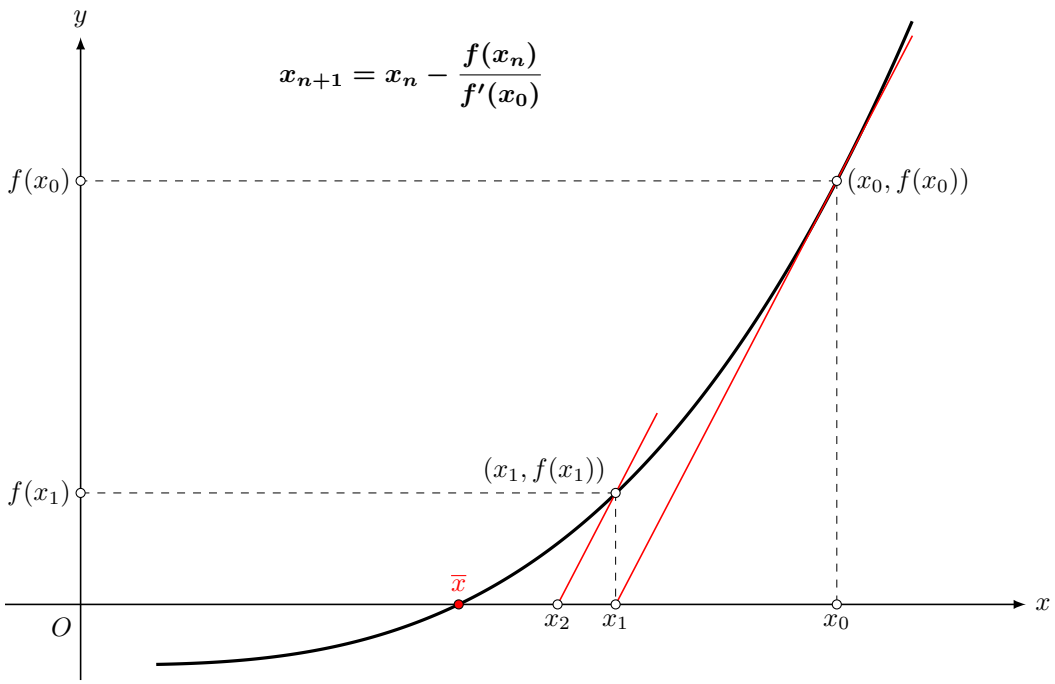

Figure 1.7: Méthode de la corde

Cependant, il y a un prix à payer, car la vitesse de convergence de cette méthode dépend encore plus fortement du choix de la valeur initiale x_0.

Proposition 1.6 *La méthode de la corde est une méthode de point fixe.*

Dém. Etant donné la construction graphique indiquée dans la Figure 1.7, on peut écrire à l'étape n :

$$f'(x_0) = \frac{f(x_n)}{x_n - x_{n+1}}$$

ce qui nous donne tout de suite :

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)} \tag{1.26}$$

Or, cette dernière relation est valable pour $n \in \mathbb{N}$ et définit une suite récurrente de la forme :

$$x_{n+1} = g(x_n) \tag{1.27}$$

pour la fonction d'itération :

$$g(x) = x - \frac{f(x)}{f'(x_0)} \tag{1.28}$$

Il convient de remarquer le fait que la fonction d'itération g dépend cette fois du choix de la valeur de départ x_0.

En considérant un voisinage V d'un zéro (simple) $\bar{x}$ tel que $f'(x) \neq 0 \; \forall x \in V$, on peut donc affirmer qu'appliquer la méthode de la corde pour trouver le zéro $\bar{x}$ de la fonction

f revient à appliquer la méthode de Picard pour trouver le point fixe équivalent $\bar{x}$ de la fonction d'itération g donnée par la relation (1.28). ■

Cette dernière Proposition nous permet de prouver le théorème suivant :

Théorème 1.11 *Soit une fonction $f : \mathbb{R} \to \mathbb{R}$ qui est (une fois) continûment différentiable ($f \in C^1$).*

Si f admet un zéro simple $\bar{x}$ (i.e. $f(\bar{x}) = 0$ et $f'(\bar{x}) \neq 0$), alors :

1. *$\exists \varepsilon > 0$ tel que si on choisit $x_0 \in [\bar{x} - \varepsilon, \bar{x} + \varepsilon]$, alors la suite récurrente (x_n), $n \in \mathbb{N}$, correspondant à la méthode de la corde et donnée par la relation (1.26), est convergente vers le zéro $\bar{x}$ (i.e. $\lim_{n\to\infty} x_n = \bar{x}$) ;*
2. *la convergence de la suite (x_n) est linéaire.*

Dém. A partir de la relation (1.28), on calcule la dérivée g' de la fonction d'itération :

$$g'(x) = 1 - \frac{f'(x)}{f'(x_0)} \tag{1.29}$$

Au point $\bar{x}$ (qui est un point fixe de g et un zéro simple de f), on a $f(\bar{x}) = 0$ et $f'(\bar{x}) \neq 0$. Ainsi :

$$g'(\bar{x}) = 1 - \frac{f'(\bar{x})}{f'(x_0)} \tag{1.30}$$

Etant donné la continuité de la fonction dérivée $f'(x)$, cette dernière relation implique que :

$$|g'(\bar{x})| = \left|1 - \frac{f'(\bar{x})}{f'(x_0)}\right| < 1 \tag{1.31}$$

De plus, la fonction g est bien continûment différentiable.

Ainsi, toutes les hypothèses du théorème (1.9) sont remplies et on peut conclure que la suite récurrente (1.26) obtenue par la méthode de la corde est bien convergente vers le zéro simple $\bar{x}$, avec une convergence (au moins) linéaire. ■

Les considérations faite pour la méthode de Newton concernant la valeur de départ et la condition d'arrêt restent valables aussi pour la méthode de la corde.

En généralisant l'approche géométrique utilisée par la méthode de Newton et la méthode de la corde, on peut imaginer une méthode itérative qui, à l'étape n, $n \in \mathbb{N}$, calcule la nouvelle valeur approchée x_{n+1} comme étant l'abscisse du point d'intersection d'une droite de pente arbitraire constante $\lambda \neq 0$ passant par le point $(x_n, f(x_n))$ avec l'axe des abscisses (voir la Figure 1.8). Une telle méthode est appelée **méthode de la**

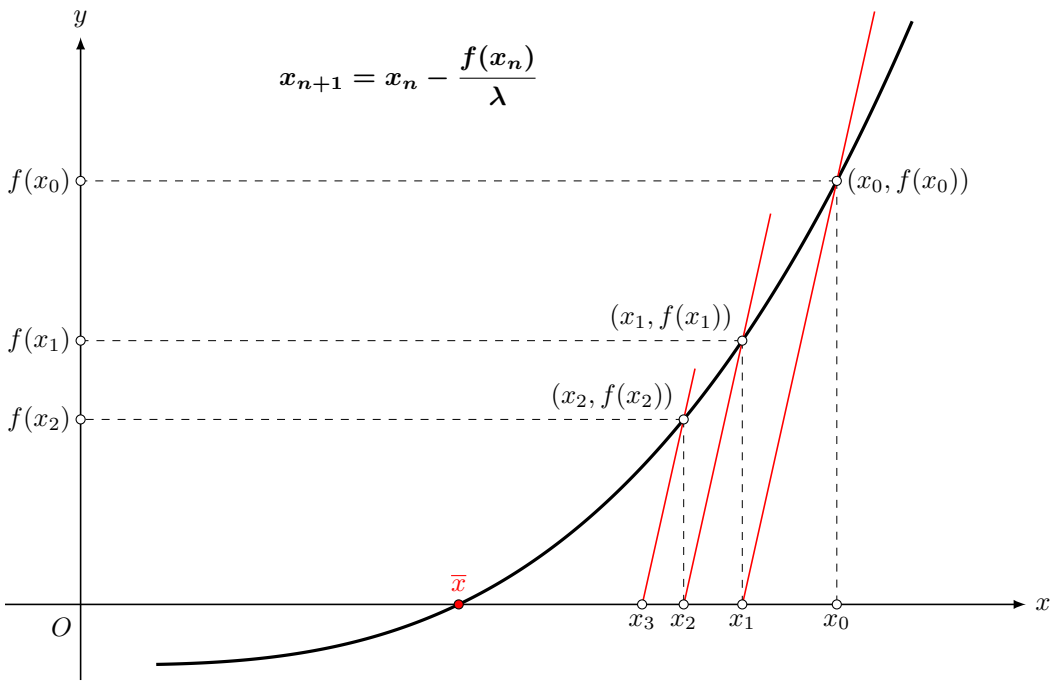

Figure 1.8: Méthode de la parallèle

parallèle et elle ne nécessite plus le calcul de la dérivée f'. La méthode de la parallèle est toujours une méthode de point fixe qui définit une suite récurrente de la forme :

$$x_{n+1} = x_n - \frac{f(x_n)}{\lambda} \tag{1.32}$$

En cas de convergence (pour $0 < \dfrac{f'(\bar{x})}{\lambda} < 2$), la convergence de la méthode de la parallèle est linéaire.

On peut aussi utiliser une **méthode** itérative dite **de la sécante**, mais qui n'est plus une méthode de point fixe (voir la Figure 1.9). Comme la méthode de la parallèle, elle n'a pas besoin de la dérivée f' mais tient mieux compte du comportement de la fonction f au voisinage du zéro cherché. La suite récurrente correspondant à la méthode de la sécante nécessite deux valeur de départ x_0 et x_1 et son terme général est :

$$x_{n+1} = x_n - \frac{f(x_n)(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})} \tag{1.33}$$

En outre, la convergence de la méthode de la sécante est d'ordre $\dfrac{1+\sqrt{5}}{2}$.

1.5 Remarques finales

Il convient de souligner que les méthodes numériques présentées dans ce chapitre ne servent pas à faire l'étude complète des fonctions non linéaires, mais à calculer numériquement,

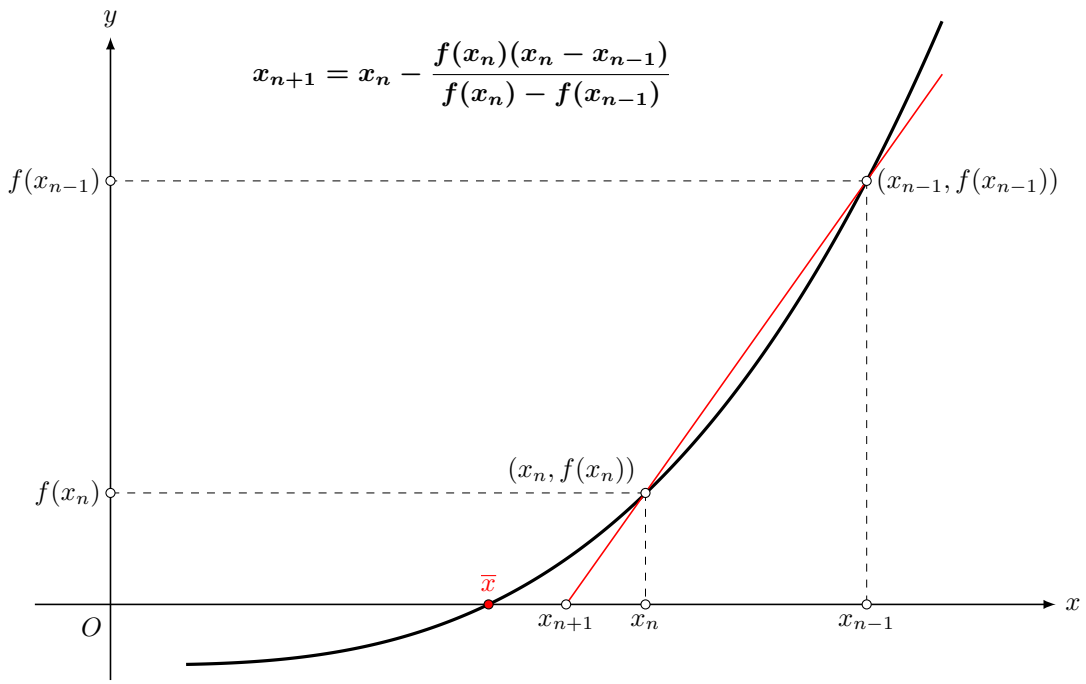

Figure 1.9: Méthode de la sécante

avec une précision voulue, leurs zéros (ceux qui ne peuvent pas être obtenus par un calcul exact, "à la main").

Etant donné leur faible vitesse de convergence, les méthodes de dichotomie sont utilisées ordinairement pour le calcul préliminaire des zéros des fonctions continues non linéaires. Ensuite, une telle approximation "grossière" d'un zéro est considérée comme valeur de départ pour une méthode numérique plus performante, par exemple, une méthode de point fixe (à condition qu'une telle méthode puisse être employée). Typiquement, une méthode de Newton convergente est plus rapide qu'une méthode de dichotomie, mais elle a besoin, normalement, d'une valeur de départ suffisamment proche du zéro cherché. Or, cette valeur de départ peut être bien obtenue, par exemple, à l'aide de la méthode de la bissection.

En outre, quand le problème mathématique à résoudre provient de la modélisation d'un problème physique, le choix d'une bonne condition de départ peut être fait pour des raisons d'ordre physique.

D'une manière générale, une méthode numérique itérative n'est utile que si elle converge vers une "vraie" solution du problème étudié. L'approche numérique ne se réduit en aucun cas à l'emploi de la "force brute" des machines à calculer. Par conséquent, on procède normalement de la manière suivante :

- on prouve d'abord que le problème à résoudre a (au moins) une solution ;
- on choisit ensuite une méthode numérique appropriée et on vérifie qu'elle est bien convergente pour le cas étudié ;

- on accepte une certaine précision pour la solution numérique à obtenir ;
- on écrit l'algorithme associé à la méthode itérative choisie, en indiquant une "bonne" condition de départ (suggérée souvent par des considérations physiques) ainsi qu'une condition d'arrêt convenable (liée d'habitude à la précision acceptée et/ou aux coûts de calcul engendrés) ;
- on écrit le programme qui implémente l'algorithme à l'aide d'un langage de programmation ;
- on met à l'épreuve le programme obtenu sur (au moins) un cas test (pour lequel on connaît les solutions) ;
- on exécute le programme pour le problème à résoudre et on obtient finalement la solution numérique ;
- on vérifie que la solution calculée est bien une solution du problème à résoudre et on discute, éventuellement, l'opportunité de la recherche d'autres solutions.

Chapitre 2

Calcul intégral

2.1 Rappel concernant l'intégrale définie d'une fonction continue

Soit un intervalle réel, borné et fermé $[a, b]$, d'extrémités $a < b$.

Définition 2.1 *Une partition (appelée aussi subdivision) de $[a, b]$ est un sous-ensemble fini et ordonné $\sigma = \{x_0, x_1, ..., x_i, ..., x_n\}$ tel que :*

$$a = x_0 < x_1 < x_2 < ... < x_i < ... < x_n = b$$

Définition 2.2 *Une partition régulière (appelée aussi subdivision régulière) d'ordre n de $[a, b]$ est une partition particulière de $[a, b]$ telle que $x_i = a + i\dfrac{b-a}{n}$ pour $i = 0, 1, .., n$.*

Définition 2.3 *Le pas d'une partition σ de $[a, b]$ est le nombre réel positif $h(\sigma) = \max\limits_{\substack{i\in\mathbb{N} \\ 0\leq i\leq n-1}} |x_{i+1} - x_i|$.*

Le pas d'une partition est en fait une mesure de la finesse de celle-ci. De plus, pour une partition régulière, le pas h devient d'autant plus petit que l'ordre n est grand.

Soit une fonction $f : [a, b] \to \mathbb{R}$ continue sur $[a, b]$ ($f \in C^0([a, b])$) et une subdivision σ de $[a, b]$.

Définition 2.4 *La somme de Darboux inférieure de la fonction f relativement à la partition σ est le nombre réel :*

$$\underline{S}_\sigma(f) = \sum_{i=0}^{n-1} m_{i+1}(x_{i+1} - x_i) \text{ avec } m_{i+1} = \min_{x\in[x_i, x_{i+1}]} f(x) \tag{2.1}$$

Définition 2.5 *La somme de Darboux supérieure de la fonction f relativement à la*

partition σ est le nombre réel :

$$\bar{S}_\sigma(f) = \sum_{i=0}^{n-1} M_{i+1}(x_{i+1} - x_i) \text{ avec } M_{i+1} = \max_{x\in[x_i,x_{i+1}]} f(x) \tag{2.2}$$

Définition 2.6 *Une somme de Riemann de la fonction f relativement à la partition σ est un nombre réel :*

$$S_\sigma(f) = \sum_{i=0}^{n-1} f(\xi_{i+1})(x_{i+1} - x_i) \text{ avec } \xi_{i+1} \in [x_i, x_{i+1}] \tag{2.3}$$

Il convient de remarquer que les sommes de Darboux inférieure et supérieure sont des sommes de Riemann particulières.

De plus, pour toute partition σ de $[a, b]$:

$$m(b-a) \leq \underline{S}_\sigma(f) \leq S_\sigma(f) \leq \bar{S}_\sigma(f) \leq M(b-a) \tag{2.4}$$

où $m = \min_{x\in[a,b]} f(x)$ et $M = \max_{x\in[a,b]} f(x)$.

Par conséquent :

- l'ensemble réel borné $\{\underline{S}_\sigma(f) : \sigma \text{ partition de } [a, b]\}$ admet bien une borne supérieure qui est son plus petit majorant que l'on note $\underline{S}(f)$:

$$\underline{S}(f) = Sup\{\underline{S}_\sigma(f) : \sigma \text{ partition de } [a, b]\} \tag{2.5}$$

- l'ensemble réel borné $\left\{\bar{S}_\sigma(f) : \sigma \text{ partition de } [a, b]\right\}$ admet bien une borne inférieure qui est son plus grand minorant que l'on note $\bar{S}(f)$:

$$\bar{S}(f) = Inf\left\{\bar{S}_\sigma(f) : \sigma \text{ partition de } [a, b]\right\} \tag{2.6}$$

Proposition 2.1 *Si la fonction $f : [a, b] \to \mathbb{R}$ est continue sur $[a, b]$ ($f \in C^0([a, b])$), alors $\underline{S}(f) = \bar{S}(f)$.*

Définition 2.7 *Le nombre réel $\underline{S}(f) = \bar{S}(f)$ est appelé l'intégrale définie de la fonction continue f sur $[a, b]$ et il est noté $\int_{x=a}^{x=b} f(x)dx$ ou encore $\int_a^b f(x)dx$.*

Proposition 2.2 *Soit une fonction $f : [a, b] \to \mathbb{R}$ continue sur $[a, b]$ ($f \in C^0([a, b])$) et (σ_n) une suite de partitions de $[a, b]$ telle que $\lim_{n\to\infty} h(\sigma_n) = 0$.*

Alors,

$$\int_a^b f(x)dx = \lim_{n\to\infty} \underline{S}_{\sigma_n}(f) = \lim_{n\to\infty} S_{\sigma_n}(f) = \lim_{n\to\infty} \bar{S}_{\sigma_n}(f) \tag{2.7}$$

2.2 Considérations générales concernant les méthodes d'intégration numérique

Toute fonction continue est intégrable mais le calcul exact de son intégrale s'avère souvent très difficile, voire impossible. De plus, dans beaucoup de situations pratiques, on ne connaît que les valeurs de la fonction à intégrer sur un ensemble fini de points. Par exemple, ces valeurs peuvent être obtenues :

- par des mesures en laboratoire ou dans des installations industrielles, en tant que données expérimentales ;
- suite à la discrétisation d'un domaine d'étude continu, dans la modélisation de certains problèmes physiques.

Heureusement, l'utilisation des méthodes d'intégration numérique est très peu contraignante et valable, en général, pour toute fonction continue (suffisamment régulière, précisée soit explicitement, soit par ses valeurs en des points imposés ou choisis convenablement).

D'une manière générale, le travail numérique s'effectue à deux niveaux :

- d'abord, on s'intéresse à un sous-intervalle quelconque $[x_i,x_{i+1}]$ de l'intervalle d'intégration $[a,b]$ et on tâche de trouver la "meilleure" méthode pour approcher l'intégrale définie $\int_{x_i}^{x_{i+1}} f(x)dx$;
- ensuite, en utilisant le résultat retenu au premier niveau, on met en oeuvre une stratégie pour diviser l'intervalle d'intégration $[a,b]$ afin de trouver la partition σ qui conduit à la "meilleure" approximation de l'intégrale définie $\int_a^b f(x)dx$ (le plus souvent, en fonction d'une précision imposée).

Concernant le premier niveau, le but est donc de trouver des formules de **quadrature** dites **non composites** qui réalisent un bon compromis entre une précision élevée et une complexité raisonnable des algorithmes induits. D'habitude, on peut adopter :

- soit une approche de nature géométrique qui consiste à remplacer le calcul de l'intégrale par l'évaluation d'une certaine aire convenablement choisie et qui peut être déterminée sans difficulté (et c'est précisément cette démarche qui a donné le nom de quadrature numérique à l'intégration numérique) ;

- soit une approche basée sur l'approximation des fonctions qui consiste à remplacer une fonction f difficilement intégrable par une fonction φ convenablement choisie (le plus souvent, une fonction d'interpolation polynomiale) et dont l'intégrale peut être calculée facilement.

Concernant les deuxième niveau, le but est de faire la "meilleure" intégration par arcs ou par morceaux, appelée **quadrature composite**, en utilisant la formule non composite choisie au premier niveau. De ce point de vue, on peut grouper les méthodes composites d'intégration numérique en deux grandes catégories :

- les **méthodes à pas fixe**, qui utilisent des partitions régulières ;
- les **méthodes à pas variable**, appelées aussi **méthodes (composites) adaptatives**, qui modifient les longueurs des sous-intervalles des partitions σ de $[a, b]$ afin d'augmenter la précision (et la vitesse) du calcul de l'intégrale $\int_a^b f(x)dx$.

A part des cas particuliers, il y aura d'habitude une différence entre la valeur calculée numériquement et la valeur exacte de l'intégrale définie, et une méthode sera d'autant meilleure que cette erreur reste petite et le temps de calcul réduit.

2.3 Interpolation de Lagrange

Soit un ensemble fini de $m+1$ points distincts $t_0, t_1, ...t_j, ..., t_m$ auxquels sont associées, respectivement, les valeurs réelles $p_0, p_1, ..., p_j, ..., p_m$.

On veut trouver un polynôme $p \in \mathbb{P}_m$, $m \in \mathbb{N}$, tel que $\forall j \in \{0, 1, 2, ..., m-1, m\}$:

$$p(t_j) = p_j \tag{2.8}$$

Afin de résoudre ce problème, on peut utiliser la **méthode d'interpolation de Lagrange** qui permet de construire le polynôme p dont le graphique passe par les points (t_j, p_j) sans connaître explicitement ses coefficients.

Ainsi, on considère d'abord un **entier** $k \in \{0, 1, 2, ..., m-1, m\}$ **fixé** et la fonction :

$$\begin{aligned} \varphi_k(t) &= \frac{(t-t_0)(t-t_1)\cdots(t-t_{k-1})\,(t-t_{k+1})\cdots(t-t_m)}{(t_k-t_0)(t_k-t_1)\cdots(t_k-t_{k-1})\,(t_k-t_{k+1})\cdots(t_k-t_m)} \\ &= \prod_{j=0,\, j\neq k}^{m} \frac{t-t_j}{t_k-t_j} \end{aligned} \tag{2.9}$$

La fonction $\varphi_k(t)$ ainsi précisée jouit des propriétés suivantes :

- φ_k est un polynôme de degré m (i.e. $\varphi_k \in \mathbb{P}_m$) ;

- $\varphi_k(t_k) = 1$;
- $\varphi_k(t_j) = 0$ pour $j \neq k$ et $j \in \{0, 1, 2, ..., m\}$.

Proposition 2.3 *Les polynômes $\varphi_0, \varphi_1, ...\varphi_k, ..., \varphi_m$ forment une base de $\mathbb{P}_m$.*

Dém. On montre d'abord que les $m+1$ polynômes $\varphi_0, \varphi_1, ..., \varphi_m$ sont linéairement indépendants. Pour cela, on considère une combinaison linéaire $\sum_{j=0}^{m} \alpha_j \varphi_j(t)$ avec $\alpha_j \in \mathbb{R}$, telle que $\forall t \in R$:

$$\sum_{j=0}^{m} \alpha_j \varphi_j(t) = 0 \tag{2.10}$$

Si, dans la relation ci-dessus, on impose $t = t_k$ pour chaque entier $k \in \{0, 1, 2, ..., m\}$, on obtient :

$$\sum_{j=0}^{m} \alpha_j \varphi_j(t_k) = \alpha_k = 0 \tag{2.11}$$

Par conséquent, tous les coefficients α_j, $j \in \{0, 1, 2, ..., m\}$, sont identiquement nuls et, donc, les polynômes $\varphi_0, \varphi_1, ..., \varphi_m$ sont bien linéairement indépendants.

De plus, vu que la dimension de l'espace vectoriel $\mathbb{P}_m$ est $m+1$, les $m+1$ polynômes $\varphi_0, \varphi_1, ..., \varphi_m$ linéairement indépendants forment bien une base de $\mathbb{P}_m$. ■

Définition 2.8 *La base de Lagrange de l'espace vectoriel $\mathbb{P}_m$ associée aux points $t_0, t_1, ...t_j, ..., t_{m-1}, t_m$ est $\{\varphi_0, \varphi_1, ..., \varphi_j, ..., \varphi_m\}$.*

Proposition 2.4 *La fonction :*

$$\begin{aligned} p(t) &= p_0\varphi_0(t) + p_1\varphi_1(t) + ...p_j\varphi_j(t) + ... + p_m\varphi_m(t) \\ &= \sum_{j=0}^{m} p_j\varphi_j(t) \end{aligned} \tag{2.12}$$

est un polynôme de l'espace vectoriel $\mathbb{P}_m$ qui satisfait bien les conditions (2.8), i.e. $p(t_j) = p_j$ $\forall j \in \{0, 1, 2, ..., m-1, m\}$.

Ce polynôme est appelé le polynôme de Lagrange associé aux couples (t_j, p_j), $j \in \{0, 1, 2, ..., m-1, m\}$.

Dém. Remarquons d'abord que $p(t)$ est une combinaison linéaire de $m+1$ polynômes de degré m. Par conséquent, $p(t)$ est lui-même un polynôme de degré m.

Il suffit de calculer ensuite, pour chaque entier $k \in \{0, 1, 2, ..., m-1, m\}$:

$$p(t_k) = \sum_{j=0}^{m} p_j\varphi_j(t_k) = p_k \tag{2.13}$$

■

On a prouvé donc, **par construction**, que le problème (2.8) a toujours une solution donnée explicitement par (2.12).

De plus, cette solution est unique.

2.4 Formules de quadrature non composites

Soit $[x_i, x_{i+1}]$ un sous-intervalle quelconque de $[a, b]$ et f une fonction à intégrer qui est définie et continue sur $[a, b]$ (i.e. $f \in C^0([a, b])$).

Sans restreindre la généralité, on peut remplacer le sous-intervalle $[x_i, x_{i+1}]$ par **l'intervalle canonique** $[-1, 1]$. Plus précisément, on fait le changement de variable :

$$x = \frac{x_i + x_{i+1}}{2} + \frac{x_{i+1} - x_i}{2} t \tag{2.14}$$

et on note :

$$f\left(\frac{x_i + x_{i+1}}{2} + \frac{x_{i+1} - x_i}{2} t\right) = g(t) \tag{2.15}$$

ce qui conduit à l'égalité :

$$\int_{x_i}^{x_{i+1}} f(x)dx = \frac{x_{i+1} - x_i}{2} \int_{-1}^{1} g(t)dt \tag{2.16}$$

où g est bien une fonction continue définie sur $[-1, 1]$ (i.e. $g \in C^0([-1, 1])$).

Définition 2.9 *Une formule de quadrature interpolatoire non composite pour la fonction continue g définie sur $[-1, 1]$ est une approximation de l'intégrale définie $\int_{-1}^{1} g(t)dt$, donnée comme une combinaison linéaire :*

$$J(g) = \sum_{j=0}^{m} \mu_j g(t_j) \tag{2.17}$$

où les points $-1 \leq t_0 < t_1 < ... < t_{m-1} < t_m \leq 1$ sont appelés les noeuds de quadrature et les coefficients réels μ_j sont appelés les poids de quadrature.

Définition 2.10 *La formule de quadrature interpolatoire non composite :*

$$J(g) = \sum_{j=0}^{m} \mu_j g(t_j) \tag{2.18}$$

est dite exacte pour les polynômes de degré r (avec $r \in \mathbb{N}$) si, $\forall p \in \mathbb{P}_r$ (où $\mathbb{P}_r$ est l'espace vectoriel des polynômes de degré inférieur ou égal à r), on a :

$$J(p) = \sum_{j=0}^{m} \mu_j p(t_j) = \int_{-1}^{1} p(t)dt \tag{2.19}$$

Définition 2.11 *Le degré d'exactitude d'une formule de quadrature interpolatoire de la forme (2.17) est le degré le plus élevé (le degré maximal) des polynômes intégrables*

exactement par cette formule.

En général, on veut qu'une formule de quadrature intègre exactement au moins les fonctions constantes. Ceci implique que, pour l'intervalle d'intégration canonique $[-1, 1]$:

$$\sum_{j=0}^{m} \mu_j = 2 \tag{2.20}$$

Considérons le cas où soit on connaît $m+1$ noeuds de quadrature soit on choisit ces derniers et on veut déterminer les poids correspondants pour que la formule de quadrature interpolatoire (2.17) ait le degré d'exactitude m.

Théorème 2.1 *Soit $m+1$ noeuds de quadrature distincts $-1 \leq t_0 < t_1 < ... < t_{m-1} < t_m \leq 1$ et $\{\varphi_0, \varphi_1, ..., \varphi_j, ..., \varphi_m\}$ la base de Lagrange de l'espace vectoriel P_m associée à ces points de quadrature.*

Alors, le degré d'exactitude de la formule de quadrature :

$$J(g) = \sum_{j=0}^{m} \mu_j g(t_j) \tag{2.21}$$

est (au moins) m si et seulement si, $\forall j \in \{0, 1, 2, ..., m-1, m\}$:

$$\mu_j = \int_{-1}^{1} \varphi_j(t) dt \tag{2.22}$$

Dém. Pour prouver une première des deux implications de l'équivalence, on suppose que le degré d'exactitude est bien m. Alors, pour tout polynôme $p \in \mathbb{P}_m$:

$$J(p) = \sum_{j=0}^{m} \mu_j p(t_j) = \int_{-1}^{1} p(t) dt \tag{2.23}$$

En particulier, pour un **entier** $k \in \{0, 1, 2, ..., m-1, m\}$ **fixé**, on peut choisir $p = \varphi_k$ et on a :

$$J(\varphi_k) = \sum_{j=0}^{m} \mu_j \varphi_k(t_j) = \int_{-1}^{1} \varphi_k(t) dt \tag{2.24}$$

Mais :

$$\sum_{j=0}^{m} \mu_j \varphi_k(t_j) = \mu_k \tag{2.25}$$

et donc :

$$\mu_k = \int_{-1}^{1} \varphi_k(t) dt \tag{2.26}$$

Pour prouver la deuxième implication de l'équivalence, on suppose que les poids de

quadrature sont bien donnés par les formules :

$$\mu_j = \int_{-1}^{1} \varphi_j(t)dt \tag{2.27}$$

On considère maintenant un polynôme quelconque $p \in P_m$. Etant donné que la base de Lagrange est une base de P_m, on peut écrire l'expression du polynôme p dans cette base :

$$p\left(t\right) = \sum\nolimits_{j=0}^{m} p\left(t_j\right) \varphi_j\left(t\right) \tag{2.28}$$

et on calcule :

$$\begin{aligned} \int_{-1}^{1} p(t)dt &= \int_{-1}^{1} \left(\sum\nolimits_{j=0}^{m} p\left(t_j\right) \varphi_j\left(t\right) \right) dt \\ &= \sum\nolimits_{j=0}^{m} p\left(t_j\right) \int_{-1}^{1} \varphi_j\left(t\right) dt \\ &= \sum\nolimits_{j=0}^{m} p\left(t_j\right) \mu_j = J(p) \end{aligned} \tag{2.29}$$

■

Il convient de remarquer que les formules (2.22) nous permettent de **calculer les poids de quadrature μ_j en fonction des polynômes φ_j (de la base de Lagrange)** qui peuvent être écrits explicitement à partir des noeuds de quadrature.

Si on a $m+1$ noeuds de quadrature fixés où on connaît les valeurs de la fonction à intégrer g, le théorème (2.1) nous donne une condition nécessaire et suffisante pour obtenir une formule de quadrature avec le degré d'exactitude d'au moins m. Par contre, si les noeuds de quadrature peuvent être choisis librement dans l'intervalle d'intégration, les formules (2.22) peuvent conduire à des formules de quadrature de degré d'exactitude assuré plus grand que m.

Le choix des $m+1$ noeuds de quadrature est optimal (dans le sens où il assure le plus grand degré d'exactitude possible) pour les **formules** de quadrature dites **gaussiennes**, à savoir :

- les **formules de Gauss-Legendre** qui, en plaçant les noeuds de quadrature dans l'intervalle ouvert $]-1,1[$ selon les zéros réels du polynôme de Legendre L_{m+1}, assurent un degré d'exactitude optimal de $2m+1$;

- les **formules de Gauss-Legendre-Lobatto** qui, en prenant comme noeuds de quadrature aussi les extrémités de l'intervalle fermé $[-1,1]$, assurent un degré d'exactitude optimal de $2m-1$.

2.5 Formules de quadrature non composites de Newton-Cotes

Les formules dites de Newton-Cotes sont des formules de quadrature interpolatoires non composites basées sur le théorème (2.1). Ce théorème assure un degré d'exactitude d'au moins m pour la formule de quadrature utilisant $m+1$ poids calculés en fonction de la base de Lagrange correspondant aux $m+1$ noeuds de quadrature. Pour déduire les **formules de Newton-Cotes**, on considère les $m+1$ noeuds de quadratures **équidistants**, en prenant en compte **aussi les limites** de l'intervalle d'intégration.

On présente ci-dessous des formules de Newton-Cotes pour les cas particuliers où $m \in \{0, 1, 2, 3\}$. On traite au départ une fonction continue g sur l'intervalle d'intégration canonique $[-1, 1]$ et on détermine $J(g)$, mais on revient ensuite, pour chaque formule obtenue, à la fonction continue f sur un sous-intervalle d'intégration $[x_i, x_{i+1}]$, pour écrire la formule de quadrature correspondante $J_i(f)$.

2.5.1 Formule non composite du point milieu ($m = 0$)

Le cas $m = 0$ est un cas "spécial" car on a un seul point de quadrature t_0 qui peut être choisi arbitrairement dans l'intervalle fermé $[-1, 1]$, vu que le problème du positionnement équidistant des points ne se pose pas.

La base de Lagrange de $\mathbb{P}_0$ est formée d'un seul polynôme de degré 0, à savoir $\varphi_0 = 1$. Par la formule (2.22), on calcule le seul poids de quadrature :

$$\mu_0 = \int_{-1}^{1} \varphi_0(t)dt = \int_{-1}^{1} dt = 2 \tag{2.30}$$

Ainsi, la formule de quadrature interpolatoire non composite (2.21) devient :

$$\begin{aligned} J(g) &= \mu_0 g(t_0) \\ &= 2g(t_0) \end{aligned} \tag{2.31}$$

Cette formule de quadrature a un degré d'exactitude d'au moins 0, c'est-à-dire elle intègre exactement au moins les fonctions constantes. En particulier, on peut choisir $t_0 = -1$ ou $t_0 = 1$.

Si on veut que la formule de quadrature ait le degré d'exactitude 1, il suffit de choisir le seul point de quadrature comme étant le point milieu de l'intervalle $[-1, 1]$, i.e. $t_0 = 0$. Dans ce cas, la formule ci-dessus devient :

$$J^{PM}(g) = 2g(0) \tag{2.32}$$

où l'exposant "PM" précise qu'il s'agit de la formule de quadrature non composite dite "du point milieu".

En effet, pour un polynôme quelconque de degré 1, $p \in \mathbb{P}_1$ de la forme $p = c_1 t + c_0$ avec $c_0, c_1 \in \mathbb{R}$, on a, d'un côté :

$$\int_{-1}^{1} p(t)\, dt = \int_{-1}^{1} (c_1 t + c_0)\, dt = 2c_0 \tag{2.33}$$

et, de l'autre côté :

$$J^{PM}(p) = 2p(0) = 2c_0 \tag{2.34}$$

Si on revient maintenant à la fonction $f(x)$ qui correspond à la fonction $g(t)$, la formule (2.32) devient :

$$J_i^{PM}(f) = (x_{i+1} - x_i)\, f\left(\frac{x_i + x_{i+1}}{2}\right) \tag{2.35}$$

Interprétation géométrique

La formule de quadrature non composite (2.35) est appelée la **formule du point milieu** ou la **formule du rectangle**.

Cette formule donne une approximation de l'intégrale définie $\int_{x_i}^{x_{i+1}} f(x)dx$ par l'aire "analytique" d'un rectangle de base égale à la longueur de l'intervalle d'intégration et de hauteur égale à la valeur de la fonction à intégrer au point milieu de l'intervalle d'intégration.

2.5.2 Formule non composite du trapèze ($m = 1$)

Le cas $m = 1$ correspond à deux points de quadrature : $t_0 = -1$ et $t_1 = 1$ (car il faut prendre en compte les limites de l'intervalle d'intégration).

La base de Lagrange de $\mathbb{P}_1$ est formée de deux polynômes de degré 1, à savoir :

$$\varphi_0(t) = \frac{(t - t_1)}{(t_0 - t_1)} = \frac{t - 1}{-2} \tag{2.36}$$

$$\varphi_1(t) = \frac{(t - t_0)}{(t_1 - t_0)} = \frac{t + 1}{2} \tag{2.37}$$

Par la formule (2.22), on calcule les deux poids de quadrature :

$$\mu_0 = \int_{-1}^{1} \varphi_0(t)dt = \int_{-1}^{1} \frac{t-1}{-2}dt = 1 \tag{2.38}$$

$$\mu_1 = \int_{-1}^{1} \varphi_1(t)dt = \int_{-1}^{1} \frac{t+1}{2}dt = 1 \tag{2.39}$$

Ainsi, la formule de quadrature interpolatoire non composite (2.21) devient :

$$\begin{aligned} J^{Tr}(g) &= \mu_0 g(t_0) + \mu_1 g(t_1) \\ &= g(-1) + g(1) \end{aligned} \tag{2.40}$$

où l'exposant "Tr" précise qu'il s'agit de la formule de quadrature non composite dite "du trapèze".

Cette formule de quadrature a un degré d'exactitude 1, c'est-à-dire elle intègre exactement les polynômes de degré 1.

Si on revient maintenant à la fonction $f(x)$ qui correspond à la fonction $g(t)$, on obtient :

$$J_i^{Tr}(f) = (x_{i+1} - x_i)\frac{f(x_i) + f(x_{i+1})}{2} \tag{2.41}$$

Interprétation géométrique

La formule de quadrature non composite (2.41) est appelée la **formule du trapèze** ou la **formule de la sécante**.

Cette formule donne une approximation de l'intégrale définie $\int_{x_i}^{x_{i+1}} f(x)dx$ par l'aire "analytique" d'un trapèze de bases égales à la valeur de la fonction à l'extrémité gauche et, respectivement, à l'extrémité droite de l'intervalle d'intégration et de hauteur égale à la longueur de l'intervalle d'intégration.

2.5.3 Formule non composite de Simpson ($m = 2$)

Le cas $m = 2$ correspond à trois points de quadrature équidistants : $t_0 = -1$, $t_1 = 0$ et $t_2 = 1$.

La base de Lagrange de $\mathbb{P}_2$ est formée de trois polynômes de degré 2, à savoir :

$$\varphi_0(t) = \frac{(t-t_1)(t-t_2)}{(t_0-t_1)(t_0-t_2)} = \frac{t(t-1)}{2} = \frac{1}{2}\left(t^2-t\right) \tag{2.42}$$

$$\varphi_1(t) = \frac{(t-t_0)(t-t_2)}{(t_1-t_0)(t_1-t_2)} = \frac{(t+1)(t-1)}{-1} = -t^2+1 \tag{2.43}$$

$$\varphi_2(t) = \frac{(t-t_0)(t-t_1)}{(t_2-t_0)(t_2-t_1)} = \frac{(t+1)t}{2} = \frac{1}{2}\left(t^2+t\right) \tag{2.44}$$

Par la formule (2.22), on calcule les trois poids de quadrature :

$$\mu_0 = \int_{-1}^{1} \varphi_0(t)dt = \int_{-1}^{1} \frac{1}{2}\left(t^2-t\right)dt = \frac{1}{3} \tag{2.45}$$

$$\mu_1 = \int_{-1}^{1} \varphi_1(t)dt = \int_{-1}^{1} \left(-t^2+1\right)dt = \frac{4}{3} \tag{2.46}$$

$$\mu_2 = \int_{-1}^{1} \varphi_2(t)dt = \int_{-1}^{1} \frac{1}{2}\left(t^2+t\right)dt = \frac{1}{3} \tag{2.47}$$

Ainsi, la formule de quadrature interpolatoire non composite (2.21) devient :

$$\begin{aligned} J^S(g) &= \mu_0 g(t_0) + \mu_1 g(t_1) + \mu_2 g(t_2) \\ &= \frac{1}{3}g(-1) + \frac{4}{3}g(0) + \frac{1}{3}g(1) \end{aligned} \tag{2.48}$$

où l'exposant "S" précise qu'il s'agit de la formule de quadrature non composite dite "de Simpson".

Cette formule de quadrature a un degré d'exactitude d'au moins 2, c'est-à-dire elle intègre exactement au moins les polynômes de degré 2.

On peut montrer qu'en réalité le degré d'exactitude de la formule ci-dessus est meilleur, à savoir 3. Il suffit de considérer une fonction polynomiale $p(t) = c_3 t^3$, où $c_3 \in \mathbb{R}$ est un coefficient réel.

Alors, d'un côté, on a :

$$\int_{-1}^{1} p(t)\,dt = \int_{-1}^{1} c_3 t^3 dt = 0 \tag{2.49}$$

et, d'autre côté, on a :

$$\begin{aligned} J^S(p) &= \frac{1}{3}p(-1) + \frac{4}{3}p(0) + \frac{1}{3}p(1) \\ &= -\frac{1}{3}c_3 + \frac{1}{3}c_3 = 0 \end{aligned} \tag{2.50}$$

Si on revient maintenant à la fonction $f(x)$ qui correspond à la fonction $g(t)$, la formule (2.48) devient :

$$J_i^S(f) = (x_{i+1} - x_i)\left(\frac{1}{6}f(x_i) + \frac{4}{6}f\left(\frac{x_i + x_{i+1}}{2}\right) + \frac{1}{6}f(x_{i+1})\right) \tag{2.51}$$

Interprétation géométrique

La formule de quadrature non composite (2.51) est appelée la **formule de Simpson**. Elle peut être regardée comme une moyenne pondérée entre la formule du rectangle (à laquelle on donne le poids $\frac{2}{3}$) et la formule du trapèze (à laquelle on donne le poids $\frac{1}{3}$).

De plus, la formule de Simpson donne une approximation de l'intégrale définie $\int_{x_i}^{x_{i+1}} f(x)dx$ par l'aire "analytique" comprise entre l'axe des abscisses et la parabole qui passe par les trois points du graphique de la fonction à intégrer correspondant à l'extrémité gauche, au point milieu et à l'extrémité droite de l'intervalle d'intégration.

2.5.4 *Formule non composite des $\frac{3}{8}$ de Newton ($m = 3$)

Le cas $m = 3$ correspond à quatre points de quadrature équidistants : $t_0 = -1$, $t_1 = -\frac{1}{3}$, $t_2 = \frac{1}{3}$ et $t_3 = 1$.

La base de Lagrange de $\mathbb{P}_3$ est formée de quatre polynômes de degré 3, à savoir :

$$\begin{aligned}
\varphi_0(t) &= \frac{(t-t_1)(t-t_2)(t-t_3)}{(t_0-t_1)(t_0-t_2)(t_0-t_3)} = \frac{\left(t+\frac{1}{3}\right)\left(t-\frac{1}{3}\right)(t-1)}{\left(-\frac{2}{3}\right)\cdot\left(-\frac{4}{3}\right)\cdot(-2)} \\
&= -\frac{1}{16}\left(9t^3 - 9t^2 + t + 1\right) && (2.52) \\
\varphi_1(t) &= \frac{(t-t_0)(t-t_2)(t-t_3)}{(t_1-t_0)(t_1-t_2)(t_1-t_3)} = \frac{(t+1)\left(t-\frac{1}{3}\right)(t-1)}{\frac{2}{3}\cdot\left(-\frac{2}{3}\right)\cdot\left(-\frac{4}{3}\right)} \\
&= \frac{9}{16}\left(3t^3 - t^2 - 3t + 1\right) && (2.53) \\
\varphi_2(t) &= \frac{(t-t_0)(t-t_1)(t-t_3)}{(t_2-t_0)(t_2-t_1)(t_2-t_3)} = \frac{(t+1)\left(t+\frac{1}{3}\right)(t-1)}{\frac{4}{3}\cdot\frac{2}{3}\cdot\left(-\frac{2}{3}\right)} \\
&= -\frac{9}{16}\left(3t^3 + t^2 - 3t - 1\right) && (2.54) \\
\varphi_3(t) &= \frac{(t-t_0)(t-t_1)(t-t_2)}{(t_3-t_0)(t_3-t_1)(t_3-t_2)} = \frac{(t+1)\left(t+\frac{1}{3}\right)\left(t-\frac{1}{3}\right)}{2\cdot\frac{4}{3}\cdot\frac{2}{3}} \\
&= \frac{1}{16}\left(9t^3 + 9t^2 - t - 1\right) && (2.55)
\end{aligned}$$

Par la formule (2.22), on calcule les quatre poids de quadrature :

$$\mu_0 = \int_{-1}^{1} \varphi_0(t)dt = -\frac{1}{16}\int_{-1}^{1} \frac{1}{2}\left(9t^3 - 9t^2 + t + 1\right) dt = \frac{1}{4} \tag{2.56}$$

$$\mu_1 = \int_{-1}^{1} \varphi_1(t)dt = \frac{9}{16}\int_{-1}^{1} \left(3t^3 - t^2 - 3t + 1\right) dt = \frac{3}{4} \tag{2.57}$$

$$\mu_2 = \int_{-1}^{1} \varphi_2(t)dt = -\frac{9}{16}\int_{-1}^{1} \left(3t^3 + t^2 - 3t - 1\right) dt = \frac{3}{4} \tag{2.58}$$

$$\mu_3 = \int_{-1}^{1} \varphi_3(t)dt = \frac{1}{16}\int_{-1}^{1} \left(9t^3 + 9t^2 - t - 1\right) dt = \frac{1}{4} \tag{2.59}$$

Ainsi, la formule de quadrature interpolatoire non composite (2.21) devient :

$$\begin{aligned} J^N(g) &= \mu_0 g(t_0) + \mu_1 g\left(t_1\right) + \mu_2 g(t_2) + \mu_3 g(t_3) \\ &= \frac{1}{4}g\left(-1\right) + \frac{3}{4}g\left(-\frac{1}{3}\right) + \frac{3}{4}g\left(\frac{1}{3}\right) + \frac{1}{4}g\left(1\right) \end{aligned} \tag{2.60}$$

où l'exposant "N" précise qu'il s'agit de la formule de quadrature non composite dite "des $\frac{3}{8}$ de Newton".

Cette formule de quadrature a un degré d'exactitude 3, c'est-à-dire elle intègre exactement les polynômes de degré 3.

Si on revient maintenant à la fonction $f\left(x\right)$ qui correspond à la fonction $g(t)$, la formule (2.60) devient :

$$\begin{aligned} J_i^N(f) &= \left(x_{i+1} - x_i\right)\left(\frac{1}{8}f\left(x_i\right) + \frac{3}{8}f\left(x_i + \frac{2}{3}\left(x_{i+1} - x_i\right)\right)\right. \\ &\quad \left. + \frac{3}{8}f\left(x_i + \frac{4}{3}\left(x_{i+1} - x_i\right)\right) + \frac{1}{8}f\left(x_{i+1}\right)\right) \end{aligned} \tag{2.61}$$

Interprétation géométrique

La formule de quadrature non composite (2.61) est appelée la **formule des $\frac{3}{8}$ de Newton** (ou de Simpson). Elle donne une approximation de l'intégrale définie $\int_{x_i}^{x_{i+1}} f(x)dx$ par l'aire "analytique" comprise entre l'axe des abscisses et le graphique de la fonction polynomiale de degré 3 qui passe par les quatre points du graphique de la fonction à intégrer correspondant aux quatre abscisses équidistantes (y compris les extrémités de l'intervalle d'intégration $[x_i, x_{i+1}]$).

2.5.5 *Formules non composites pour m grand

D'une manière générale, on peut déduire des formules de Newton-Cotes pour n'importe quel nombre de points de quadrature $m+1$. Cependant, pour $m > 3$, les algorithmes induits peuvent devenir compliqués et instables. Par conséquent, afin de gagner en précision, on préfère utiliser des méthodes composites basées sur des méthodes non composites obtenues pour $m \leq 3$ et assurer, en revanche, une finesse élevée de la partition utilisée à laquelle on ajoute, éventuellement, l'adaptation du pas d'intégration.

On peut mentionner, quand même, la formule pour $m = 4$ appelée la formule non composite de Bode ou de Boole-Villarceau :

$$J_i(f) = (x_{i+1} - x_i)\left(\frac{7}{90}f_0 + \frac{16}{45}f_1 + \frac{6}{45}f_2 + \frac{16}{45}f_3 + \frac{7}{90}f_4\right) \tag{2.62}$$

où f_j, $j \in \{0, 1, ..., 4\}$, sont les valeurs de la fonction à intégrer aux points de quadrature équidistants (y compris les extrémités de l'intervalle d'intégration $[x_i, x_{i+1}]$).

En ce qui concerne le degré d'exactitude, une formule de Newton-Cotes obtenue pour $m+1$ points de quadrature est exacte pour des polynômes de degré maximal m si m est impaire, tandis qu'elle est exacte pour des polynômes de degré maximal $m+1$ si m est paire.

2.6 Formules de quadrature composites

Après avoir choisi une méthode de quadrature non composite à utiliser pour un sous-intervalle $[x_i, x_{i+1}]$, il faut maintenant approcher l'intégrale définie sur l'intervalle complet d'intégration $[a, b]$.

Définition 2.12 *Soit une fonction f définie et continue sur $[a, b]$ (i.e. $f \in C^0([a, b])$), une partition σ de l'intervalle $[a, b]$ avec $a = x_0 < x_1 < x_2 < ... < x_i < ... < x_n = b$ et une formule de quadrature interpolatoire non composite $J_i(f)$, $i = 0, 1, ..., n-1$, qui donne une approximation de l'intégrale définie sur un sous-intervalle $[x_i, x_{i+1}]$, i.e. une valeur approchée de $\int_{x_i}^{x_{i+1}} f(x)dx$.*

Alors, la formule de quadrature composite correspondante est une approximation de l'intégrale définie $\int_a^b f(x)dx$, donnée par :

$$I(f) = \sum_{i=0}^{n-1} J_i(f) \tag{2.63}$$

Ainsi, pour chaque formule de quadrature interpolatoire non composite, on peut écrire

la formule composite correspondante.

La formule de quadrature composite du point milieu ou du rectangle :

$$I^{PM}(f) = \sum_{i=0}^{n-1} (x_{i+1} - x_i) f\left(\frac{x_i + x_{i+1}}{2}\right) \tag{2.64}$$

La formule de quadrature composite du trapèze ou de la sécante :

$$I^{Tr}(f) = \sum_{i=0}^{n-1} (x_{i+1} - x_i) \frac{f(x_i) + f(x_{i+1})}{2} \tag{2.65}$$

La formule de quadrature composite de Simpson :

$$I^{S}(f) = \sum_{i=0}^{n-1} (x_{i+1} - x_i) \left(\frac{1}{6} f(x_i) + \frac{4}{6} f\left(\frac{x_i + x_{i+1}}{2}\right) + \frac{1}{6} f(x_{i+1})\right) \tag{2.66}$$

Dans le cas particulier d'une partition régulière de pas $h = \dfrac{b-a}{n}$, où n est le nombre de sous-intervalles de longueur constante de la partition, les formules ci-dessus deviennent:

$$I_{reg}^{PM}(f) = h \sum_{i=0}^{n-1} f\left(\frac{x_i + x_{i+1}}{2}\right) \tag{2.67}$$

$$I_{reg}^{Tr}(f) = h \sum_{i=0}^{n-1} \frac{f(x_i) + f(x_{i+1})}{2} \tag{2.68}$$

$$I_{reg}^{S}(f) = h \sum_{i=0}^{n-1} \left(\frac{1}{6} f(x_i) + \frac{4}{6} f\left(\frac{x_i + x_{i+1}}{2}\right) + \frac{1}{6} f(x_{i+1})\right) \tag{2.69}$$

où l'indice "reg" précise qu'il s'agit d'une partition régulière.

2.7 Estimation d'erreur

La qualité d'une méthode d'intégration numérique dépend, en grande mesure, de la possibilité d'estimer l'erreur absolue $e_{abs} = \left| \int_a^b f(x)dx - I(f) \right|$ commise en approchant l'intégrale définie $\int_a^b f(x)dx$ par la valeur obtenue à l'aide d'une formule de quadrature composite $I(f)$. Pour une fonction suffisamment régulière, on donne ci-dessous (sans faire la démonstration dans le cas général) un théorème qui précise un majorant pour cette erreur. Ce majorant dépend du degré d'exactitude r de la formule non composite qui est à la base de la formule composite utilisée, ainsi que du pas h de la partition de l'intervalle d'intégration $[a, b]$.

Théorème 2.2 *Soit :*

- *une formule de quadrature non composite qui a le degré d'exactitude r ;*
- *une fonction f qui est définie et $(r+1)$ fois continûment différentiable sur un intervalle $[a,b]$ (i.e. $f \in C^{r+1}([a,b])$) ;*
- *une partition σ de l'intervalle $[a,b]$, qui a le pas h ;*
- *une formule de quadrature composite $I(f)$ basée sur la formule non composite mentionnée ci-dessus.*

Alors, il existe une constante réelle C indépendante du choix des points de la partition σ telle que :

$$\left| \int_a^b f(x)dx - I(f) \right| \leq Ch^{r+1} \tag{2.70}$$

Dém. Dans le cadre de ce théorème, on peut déduire d'abord des relations qui permettent de quantifier l'erreur commise afin de trouver ensuite des majorants appropriés pour chacune des formules de quadrature obtenues auparavant.

- Pour la formule de quadrature non composite du point milieu (qui a le degré d'exactitude $r = 1$) :

$$\int_{x_i}^{x_{i+1}} f(x)\,dx - J_i^{PM}(f) = \frac{(x_{i+1}-x_i)^3}{24} f''(\xi_i) = \frac{h_i^3}{24} f''(\xi_i) \tag{2.71}$$

  où $\xi_i \in]x_i, x_{i+1}[$.

  Pour la formule de quadrature composite du point milieu et pour une partition régulière en n sous-intervalles de pas constant $h = \dfrac{b-a}{n}$:

$$\int_a^b f(x)\,dx - I_{reg}^{PM}(f) = \frac{(b-a)}{24} h^2 f''(\xi) \tag{2.72}$$

  où $\xi \in]a,b[$.

- Pour la formule de quadrature non composite du trapèze (qui a le degré d'exactitude $r = 1$) :

$$\int_{x_i}^{x_{i+1}} f(x)\,dx - J_i^{Tr}(f) = -\frac{(x_{i+1}-x_i)^3}{12} f''(\xi_i) = -\frac{h_i^3}{12} f''(\xi_i) \tag{2.73}$$

où $\xi_i \in]x_i, x_{i+1}[$.

Pour la formule de quadrature composite du trapèze et pour une partition régulière de l'intervalle d'intégration en n sous-intervalles de pas constant $h = \dfrac{b-a}{n}$:

$$\int_a^b f(x)\,dx - I_{reg}^{Tr}(f) = -\frac{(b-a)}{12}h^2 f''(\xi) \tag{2.74}$$

où $\xi \in]a, b[$.

- Pour la formule de quadrature non composite de Simpson (qui a le degré d'exactitude $r = 3$) :

$$\int_{x_i}^{x_{i+1}} f(x)\,dx - J_i^S(f) = -\frac{(x_{i+1}-x_i)^5}{90 \cdot 32} f^{IV}(\xi_i) = -\frac{h_i^5}{90 \cdot 32} f^{IV}(\xi_i) \tag{2.75}$$

où $\xi_i \in]x_i, x_{i+1}[$.

Pour la formule de quadrature composite de Simpson et pour une partition régulière en n sous-intervalles de pas constant $h = \dfrac{b-a}{n}$:

$$\int_a^b f(x)\,dx - I_{reg}^{S}(f) = -\frac{(b-a)}{90 \cdot 32}h^4 f^{IV}(\xi) \tag{2.76}$$

où $\xi \in]a, b[$.

On va maintenant démontrer les deux dernières relations et on commence avec la formules de quadrature non composite de Simpson.

D'un côté, par le théorème fondamental du calcul intégral, on a que :

$$\int_{x_i}^{x_{i+1}} f(x)\,dx = F(x_{i+1}) - F(x_i) \tag{2.77}$$

où la fonction continue $F : [a, b] \to \mathbb{R}$ est une primitive de la fonction f sur $[a, b]$ et donc, par définition :

$$F'(x) = f(x) \tag{2.78}$$

$\forall x \in]a, b[$.

De plus, en supposant que la fonction à intégrer f soit 4 fois continûment différentiable sur $[x_i, x_{i+1}]$ (i.e. $f \in C^4([x_i, x_{i+1}])$), on peut écrire, au voisinage du point milieu :

$$x_{PM_i} = \frac{x_i + x_{i+1}}{2} \tag{2.79}$$

la formule de Taylor pour la primitive F, à savoir :

$$\begin{aligned} F(x_{i+1}) &= F(x_{PM_i}) + \left(\frac{h_i}{2}\right) F'(x_{PM_i}) + \frac{1}{2!}\left(\frac{h_i}{2}\right)^2 F''(x_{PM_i}) + \frac{1}{3!}\left(\frac{h_i}{2}\right)^3 F'''(x_{PM_i}) \\ &\quad + \frac{1}{4!}\left(\frac{h_i}{2}\right)^4 F^{IV}(x_{PM_i}) + \frac{1}{5!}\left(\frac{h_i}{2}\right)^5 F^{V}(\xi_{i_1}) \\ &= F(x_{PM_i}) + \left(\frac{h_i}{2}\right) f(x_{PM_i}) + \frac{1}{2!}\left(\frac{h_i}{2}\right)^2 f'(x_{PM_i}) + \frac{1}{3!}\left(\frac{h_i}{2}\right)^3 f''(x_{PM_i}) \\ &\quad + \frac{1}{4!}\left(\frac{h_i}{2}\right)^4 f'''(x_{PM_i}) + \frac{1}{5!}\left(\frac{h_i}{2}\right)^5 f^{IV}(\xi_{i_1}) \end{aligned} \tag{2.80}$$

où $\xi_{i_1} \in]x_{PM_i}, x_{i+1}[$.

De même :

$$\begin{aligned} F(x_i) &= F(x_{PM_i}) - \left(\frac{h_i}{2}\right) F'(x_{PM_i}) + \frac{1}{2!}\left(\frac{h_i}{2}\right)^2 F''(x_{PM_i}) - \frac{1}{3!}\left(\frac{h_i}{2}\right)^3 F'''(x_{PM_i}) \\ &\quad + \frac{1}{4!}\left(\frac{h_i}{2}\right)^4 F^{IV}(x_{PM_i}) - \frac{1}{5!}\left(\frac{h_i}{2}\right)^5 F^{V}(\xi_{i_2}) \\ &= F(x_{PM_i}) - \left(\frac{h_i}{2}\right) f(x_{PM_i}) + \frac{1}{2!}\left(\frac{h_i}{2}\right)^2 f'(x_{PM_i}) - \frac{1}{3!}\left(\frac{h_i}{2}\right)^3 f''(x_{PM_i}) \\ &\quad + \frac{1}{4!}\left(\frac{h_i}{2}\right)^4 f'''(x_{PM_i}) - \frac{1}{5!}\left(\frac{h_i}{2}\right)^5 f^{IV}(\xi_{i_2}) \end{aligned} \tag{2.81}$$

où $\xi_{i_2} \in]x_i, x_{PM_i}[$.

Si on remplace les relations (2.80) et (2.81) dans (2.77), on obtient :

$$\int_{x_i}^{x_{i+1}} f(x)\,dx = 2\left(\frac{h_i}{2}\right) f(x_{PM_i}) + \frac{1}{3}\left(\frac{h_i}{2}\right)^3 f''(x_{PM_i}) + \frac{1}{60}\left(\frac{h_i}{2}\right)^5 f^{IV}(\xi_i) \tag{2.82}$$

où $\xi_i \in]x_i, x_{i+1}[$ (car, grâce à la continuité de la quatrième dérivée de f, il existe ξ_i tel que $f^{IV}(\xi_{i_1}) + f^{IV}(\xi_{i_2}) = 2f^{IV}(\xi_i)$).

De l'autre côté, la formule non composite de Simpson (2.51) s'écrit :

$$J_i^S(f) = \frac{h_i}{6}\left(f(x_i) + 4f(x_{PM_i}) + f(x_{i+1})\right) \tag{2.83}$$

Comme pour F, on peut écrire aussi pour f la formule de Taylor au voisinage du point milieu :

$$\begin{aligned} f(x_{i+1}) &= f(x_{PM_i}) + \left(\frac{h_i}{2}\right) f'(x_{PM_i}) + \frac{1}{2!}\left(\frac{h_i}{2}\right)^2 f''(x_{PM_i}) \\ &\quad + \frac{1}{3!}\left(\frac{h_i}{2}\right)^3 f'''(x_{PM_i}) + \frac{1}{4!}\left(\frac{h_i}{2}\right)^4 f^{IV}(\xi_{i_1}) \end{aligned} \tag{2.84}$$

De même :

$$\begin{aligned} f(x_i) &= f(x_{PM_i}) - \left(\frac{h_i}{2}\right) f'(x_{PM_i}) + \frac{1}{2!}\left(\frac{h_i}{2}\right)^2 f''(x_{PM_i}) \\ &\quad -\frac{1}{3!}\left(\frac{h_i}{2}\right)^3 f'''(x_{PM_i}) + \frac{1}{4!}\left(\frac{h_i}{2}\right)^4 f^{IV}(\xi_{i_2}) \end{aligned} \tag{2.85}$$

Si on remplace les relations (2.84) et (2.85) dans (2.83), on obtient :

$$\begin{aligned} J_i^S(f) &= \frac{h_i}{6}\left(6f(x_{PM_i}) + \left(\frac{h_i}{2}\right)^2 f''(x_{PM_i}) + \frac{1}{12}\left(\frac{h_i}{2}\right)^4 f^{IV}(\xi_i)\right) \\ &= 2\left(\frac{h_i}{2}\right) f(x_{PM_i}) + \frac{1}{3}\left(\frac{h_i}{2}\right)^3 f''(x_{PM_i}) + \frac{1}{36}\left(\frac{h_i}{2}\right)^5 f^{IV}(\xi_i) \end{aligned} \tag{2.86}$$

Grâce aux relations (2.82) et (2.86), on obtient bien la formule (2.75) :

$$\int_{x_i}^{x_{i+1}} f(x)\,dx - J_i^S(f) = -\frac{1}{90}\left(\frac{h_i}{2}\right)^5 f^{IV}(\xi_i) = -\frac{h_i^5}{90\cdot 32} f^{IV}(\xi_i) \tag{2.87}$$

valable pour la quadrature non composite de Simpson.

Si on considère maintenant une partition régulière de l'intervalle d'intégration $[a,b]$ en n sous-intervalles de pas constant $h = \dfrac{b-a}{n}$, on obtient bien la formule (2.76) :

$$\begin{aligned} \int_a^b f(x)\,dx - I_{reg}^S(f) &= \sum_{i=0}^{n-1}\left(\int_{x_i}^{x_{i+1}} f(x)\,dx - J_i^S(f)\right) \\ &= -\frac{h^5}{90\cdot 32}\sum_{i=0}^{n-1} f^{IV}(\xi_i) \\ &= -\frac{h^5}{90\cdot 32}\; n\; f^{IV}(\xi) \\ &= -\frac{(b-a)}{90\cdot 32} h^4 f^{IV}(\xi) \end{aligned} \tag{2.88}$$

car, vu la continuité de la quatrième dérivée de f, on peut écrire $\sum_{i=0}^{n-1} f^{IV}(\xi_i) = n\; f^{IV}(\xi)$, où $\xi \in]a,b[$, et $nh = b-a$.

Finalement, grâce à la relation (2.88), on peut exhiber un majorant pour l'erreur absolue commise en utilisant la formule de quadrature composite de Simpson :

$$e_{abs} = \left| \int_a^b f(x)dx - I_{reg}^S(f) \right| \leq C_{reg}^S h^4 \tag{2.89}$$

où :

$$C_{reg}^{S} = \frac{(b-a)}{90 \cdot 32} \max_{\xi \in]a,b[} \left| f^{IV}(\xi) \right| \tag{2.90}$$

■

La relation (2.70) nous permet de connaître l'ordre de précision d'une formule de quadrature composite par rapport au pas h de la partition σ de l'intervalle d'intégration $[a, b]$.

Il convient de remarquer aussi que le théorème (2.2) montre que, pour diminuer l'erreur absolue $\left| \int_a^b f(x)dx - I(f) \right|$, il faut :

- d'une part, utiliser une partition aussi fine que possible, ce qui assure une valeur petite du pas h ;
- d'autre part, baser la formule composite sur une formule non composite avec un degré d'exactitude élevé.

On considère, par exemple, une formule de quadrature composite basée sur une formule non composite de degré d'exactitude 3 (comme la méthode de Simpson). Si on passe d'une partition régulière avec n sous-intervalles (de même longueur $\frac{b-a}{n}$) à une partition régulière avec $2n$ sous-intervalles (de même longueur $\frac{b-a}{2n}$), alors l'erreur sera divisée par 16.

2.8 *Pas fixe versus pas variable*

Afin de calculer numériquement une intégrale définie $\int_a^b f(x)dx$ avec une précision imposée ε, on choisit d'abord une formule de **quadrature non composite**. Ensuite, il faut se décider pour une méthode de **quadrature composite associée** qui peut utiliser soit un pas fixe soit un pas variable.

Les méthode de quadrature composites à **pas fixe** sont des méthodes itératives qui travaillent avec des **partitions régulières** de plus en plus fines. Plus précisément :

- pour commencer, à l'étape 0, on utilise la formule de quadrature non composite qui correspond, en fait, à une partition formée seulement de deux points (les extrémités de l'intervalle complet d'intégration $[a, b]$) ; autrement dit, au début, il n'y a qu'un seul "sous-intervalle" ($n_0 = 1$) de pas initial $h_0 = b - a$;
- à chaque nouvelle itération, on double le nombre de sous-intervalles de la partition régulière antérieure et on applique la formule de quadrature composite ; ainsi, à

l'étape k, $k > 1$, l'intervalle $[a, b]$ est partagé en $n_k = 2n_{k-1} = 2^k$ sous-intervalles de pas fixe $h_k = \dfrac{b-a}{2^k}$;

- finalement, on arrête le calcul itératif à l'étape k_{fin} quand l'erreur absolue commise devient plus petite ou égale à la tolérance imposée :

$$(e_{abs})_{k_{fin}} \leq \varepsilon \tag{2.91}$$

(ou quand un nombre maximal d'itérations $k_{\max}$ est atteint ou dépassé $k_{fin} \geq k_{\max}$).

En pratique, afin de satisfaire la condition d'arrêt (2.91), il faut trouver un estimateur d'erreur convenable qui dépend de la méthode de quadrature utilisée. Cependant, d'une manière générale, on se contente souvent avec la **condition d'arrêt** :

$$\left|(I_{reg})_{k_{fin}} - (I_{reg})_{k_{fin}-1}\right| \leq \varepsilon \tag{2.92}$$

pour une estimation de l'erreur absolue, ou encore :

$$\left|(I_{reg})_{k_{fin}} - (I_{reg})_{k_{fin}-1}\right| \leq \varepsilon \left|(I_{reg})_{k_{fin}}\right| \tag{2.93}$$

pour une estimation de l'erreur relative.

De plus, selon une technique proposée initialement par Romberg pour la méthode de quadrature du trapèze, on peut "**doper**" une méthode de quadrature afin d'obtenir un ordre de précision (par rapport au pas) plus élevé.

Dans le cas de la formule non composite de Simpson, on note J_i^S la valeur approchée de $\displaystyle\int_{x_i}^{x_{i+1}} f(x)\,dx$ et $\left(J_i^S\right)^+$ la somme des valeurs approchées de $\displaystyle\int_{x_i}^{x_{PM_i}} f(x)\,dx$ et $\displaystyle\int_{x_{PM_i}}^{x_{i+1}} f(x)\,dx$. Alors, la formule de quadrature non composite dopée de Simpson est donnée par :

$$\left(J_i^S\right)^{++} = \frac{16\left(J_i^S\right)^+ - J_i^S}{15} \tag{2.94}$$

Dans le cas de la formule composite de Simpson, on peut "doper" le résultat final avec une formule similaire :

$$\left(I_{reg}^S\right)^{++} = \frac{16\left(I_{reg}^S\right)_{k_{fin}} - \left(I_{reg}^S\right)_{k_{fin}-1}}{15} \tag{2.95}$$

et l'ordre de précision par rapport à $h_{k_{fin}}$ devient 6.

Les méthodes de quadrature composites à **pas variable**, appelées aussi **méthodes adaptatives**, sont des méthodes itératives qui travaillent avec des **partitions non régulières**. Plus précisément, les sous-intervalles d'intégration ont des longueurs

variables qui s'adaptent automatiquement selon la fonction à intégrer. Par exemple, pour la formule de quadrature composite de Simpson, à partir des relations (2.89), (2.90) et (2.91), on obtient :

$$\frac{(b-a)}{90 \cdot 32} \max_{\xi \in]a,b[} \left| f^{IV}(\xi) \right| h^4 \leq \varepsilon \tag{2.96}$$

Cette dernière inégalité montre que, pour satisfaire la tolérance imposée ε, le pas fixe h d'une partition régulière doit devenir (globalement) très petit, même si la valeur absolue de la quatrième dérivée de la fonction à intégrer est grande seulement dans certaines parties de l'intervalle d'intégration. Dans une telle situation, l'utilisation d'une partition non régulière permet de diminuer, de manière ciblée, seulement les longueurs des sous-intervalles où la quatrième dérivée varie fortement.

Le but recherché par les méthodes adaptatives est d'assurer la même précision que celle obtenue avec les méthodes à pas fixe, mais avec un nombre de sous-intervalles d'intégration plus petit, ce qui implique moins d'évaluations de la formule non composite et, donc, une vitesse de calcul améliorée. Le prix à payer réside dans la complexité relativement importante des algorithmes associés aux méthodes de quadrature composites à pas variable par rapport aux méthodes à pas fixe.

Chapitre 3

Résolution des équations différentielles ordinaires

3.1 Introduction

Définition 3.1 *Une **équation différentielle** est une équation qui fait intervenir au moins une fonction inconnue et certaines de ses dérivées (éventuellement partielles).*

Définition 3.2 *Une **équation aux dérivées partielles** est une équation différentielle qui fait intervenir des dérivées partielles d'au moins une fonction inconnue (qui dépend de plusieurs variables).*

Définition 3.3 *Une **équation différentielle ordinaire** est une équation différentielle qui ne contient aucune dérivée partielle d'une fonction inconnue.*

Définition 3.4 *L'**ordre** d'une équation différentielle est le plus élevé ordre de dérivation (éventuellement partielle) d'une fonction inconnue apparaissant dans l'équation différentielle.*

Remarque 3.1 *Les équations différentielles sont des équations fonctionnelles dont les inconnues sont des fonctions (et non plus des nombres, comme dans le cas des équations "habituelles").*

Par exemple, une équation différentielle **ordinaire** (**implicite**) **d'ordre** m qui établie une relation entre une seule fonction inconnue (réelle et qui dépend d'une variable réelle) et certaines de ses dérivées (jusqu'à l'ordre m), peut être écrite sous la forme générale suivante:

$$F\left(t, y, y', y'', ..., y^{(m)}\right) = 0 \tag{3.1}$$

où la fonction $F : U \subset \mathbb{R} \times \mathbb{R}^{m+1}$ est continue sur le domaine ouvert U.

Résoudre une équation différentielle revient à trouver (toutes) les fonctions solutions.

Une fonction $\bar{y}(t) : I \subset \mathbb{R} \to \mathbb{R}$ est une **solution** de l'équation différentielle (3.1) si elle est m fois continûment différentiable sur I (i.e. $\bar{y} \in C^m(I)$) et si, pour $\forall t \in I$, on a :

$$F\left(t, \bar{y}, \bar{y}', \bar{y}'', ..., \bar{y}^{(m)}\right) = 0 \tag{3.2}$$

Par exemple, l'équation différentielle :

$$y'' + y = 0 \tag{3.3}$$

admet une infinité de solutions de la forme :

$$\bar{y}(t) = A\cos(t) + B\sin(t) \tag{3.4}$$

où A et B sont des constantes réelles arbitraires (qui peuvent être déterminées en imposant des conditions supplémentaires appelées couramment **conditions initiales**).

Dans les domaines les plus divers (physique, chimie, biologie, sciences de l'ingénieur, économie, etc.), la grande majorité des modèles mathématiques utilisés pour l'étude des phénomènes spécifiques au domaine considéré font intervenir des équations différentielles (ordinaires ou aux dérivées partielles).

Dans ce chapitre, on présente des méthodes de résolution numériques pour les **équations différentielles ordinaires du premier ordre**. D'une manière générale, on peut réduire la résolution d'une équation différentielle ordinaire d'ordre $p > 1$ à la résolution d'un système de p équations différentielles ordinaires d'ordre 1.

D'habitude, une équation différentielle ordinaire admet une infinité de solutions. Afin d'identifier une certaine fonction appartenant à l'ensemble de solutions, on peut fournir une information supplémentaire qui donne, d'habitude, la valeur de la fonction cherchée en un certain point de son domaine de définition. Cette approche nous conduit à résoudre un problème appelé le problème de Cauchy.

Définition 3.5 (Problème de Cauchy) *Trouver une fonction $y(t) : I \subset \mathbb{R} \to \mathbb{R}$ continûment différentiable (i.e. $y \in C^1(I)$) qui satisfait :*

$$\begin{cases} y'(t) = f(t, y(t)) & \forall t \in I \\ y(t_0) = y_0 \end{cases} \tag{3.5}$$

où :

- $I \subset \mathbb{R}$ est un intervalle réel (borné ou pas et, le plus souvent, $I \subset \mathbb{R}_+$) ;
- $t \in I$ est la variable indépendante (d'habitude, une variable temporelle qu'on appellera **temps**) ;

- $y'(t)$ est la première dérivée de la fonction **inconnue** $y(t)$;
- $f : I \times \mathbb{R} \to \mathbb{R}$ est une fonction **donnée** de deux variables ;
- $t_0 \in I$ est un point du domaine de définition de y appelé **point** ou **moment initial** (et, souvent, on considère $t_0 = 0$) ;
- $y_0 \in \mathbb{R}$ est une valeur réelle donnée et appelée **valeur** ou **donnée initiale** (car il s'agit de la valeur de la fonction cherchée y au point initial) ;
- la première équation ci-dessus est l'**équation différentielle** à résoudre ;
- la deuxième équation ci-dessus est appelée une **condition de Cauchy**.

Concernant l'étude de l'existence et de l'unicité des solutions du problème de Cauchy, on rappelle ci-dessous le théorème de Cauchy-Lipschitz.

Théorème 3.1 (Cauchy-Lipschitz) *Si la fonction $f(t, y) : I \times \mathbb{R} \to \mathbb{R}$ est :*

- *continue sur son domaine de définition (par rapport à chacune de ses deux variables), i.e. $f \in C^0(I \times \mathbb{R})$;*
- *lipschitzienne par rapport à la deuxième variable, i.e. il existe une constante réelle (strictement) positive L telle que, pour $\forall t \in I$ et $\forall y_1, y_2 \in \mathbb{R}$, on a :*

$$|f(t, y_1) - f(t, y_2)| \leq L |y_1 - y_2| \tag{3.6}$$

*alors le problème de Cauchy (3.5) admet une **solution unique** $y(t)$ (qui est continûment différentiable, c'est-à-dire $y \in C^1(I)$ et qui est appelée **l'intégrale de l'équation différentielle** précisée dans le problème de Cauchy).*

Remarque 3.2 *Le théorème Cauchy-Lipschitz est valable aussi pour des intervalles I non bornés de la forme $I = [t_0, \infty[$ (où t_0 est un point initial, par exemple, $t_0 = 0$) et on parle alors d'un résultat d'existence et d'unicité **global** (dans le sens où on peut intégrer (3.5) jusqu'à l'infini).*

Remarque 3.3 *Dans la plupart de modèles mathématiques obtenus à partir des problèmes physiques, la fonction f intervenant dans un problème de Cauchy est lipschitzienne pour des raisons physiques.*

3.2 Approche numérique

Même dans les cas où on peut prouver (grâce au théorème Cauchy-Lipschitz) que le problème de Cauchy admet une solution unique, on ne peut que rarement exprimer les solutions sous forme analytique explicite (ou implicite). Par conséquent, le calcul numérique de la solution cherchée est d'autant plus important.

On considère un problème de Cauchy (3.5) :

$$\begin{cases} y'(t) = f(t, y(t)) & \forall t \in I \\ y(t_0) = y_0 \end{cases} \tag{3.7}$$

qui admet une solution unique $y = y(t)$, et une méthode de résolution numérique qui lui est associée.

De plus, on considère (au moins, au début) un intervalle I borné et de la forme $I = [t_0, T]$, où t_0 est le **moment initial** et $T - t_0$ la **durée d'étude**, et on choisit une partition (ou une subdivision) σ de I avec des noeuds :

$$t_0 < t_1 < ... < t_n < t_{n+1} < ... < t_N = T \tag{3.8}$$

où $N \geq 1$ représente le nombre de sous-intervalles de la partition.

Dans le cas usuel d'une partition régulière, le pas $h = \max\limits_{0 \leq n \leq N-1} |t_{n+1} - t_n| = \max\limits_{0 \leq n \leq N-1} h_n$, appelé aussi **pas de discrétisation**, sera donné par la relation :

$$h = \frac{T - t_0}{N} \tag{3.9}$$

car $t_{n+1} - t_n = h_n = constant$ pour $\forall t \in [0, N-1]$.

La **méthode numérique** calcule alors, de manière itérative, pour des partitions de plus en plus fines, une suite d'ensembles discrets de valeurs $\{u_0, u_1, ..., u_n, u_{n+1}, ..., u_N\}$, représentant des approximations de plus en plus précises de la solution exacte $y(t)$ aux points $\{t_0, t_1, ..., t_n, t_{n+1}, ..., t_N\}$.

A chaque itération (c'est-à-dire pour chaque partition σ fixée), au début (c'est-à-dire au t_0), on connaît $u_0 = y_0$. Ensuite, à chaque noeud t_n, $n = 1, 2, ..., N$, on approche la valeur $y_n = y(t_n)$ de la solution exacte par une valeur numérique correspondante u_n. De plus, par la suite, on note $f_n = f(t_n, u_n)$.

Concrètement, sur chaque sous-intervalle $[t_n, t_{n+1}]$, on intègre la première équation du problème de Cauchy (3.7), c'est-à-dire l'équation différentielle à résoudre, en appliquant le théorème fondamental du calcul intégral et on obtient :

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} f(t, y(t))\, dt \tag{3.10}$$

L'intégrale définie intervenant dans la relation ci-dessus peut être maintenant calculée à l'aide d'une formule de quadrature (non composite) convenable qui nous permet d'obtenir la valeur approchée u_{n+1}, c'est-à-dire l'approximation de y_{n+1}, à partir d'une ou de plusieurs valeurs approchées précédentes.

Définition 3.6 *Une méthode numérique est dite* ***convergente*** *si, pour* $\forall n = 0, 1, ..., N$, *l'erreur absolue* $e_n = |y_n - u_n|$ *reste bornée par une valeur* $C(h)$ *qui ne dépend que du pas de discrétisation* h *:*

$$e_n = |y_n - u_n| \leq C(h) \tag{3.11}$$

et qui, de plus, tend vers zéro pour des partitions de plus en plus fines :

$$\lim_{h \to 0} C(h) = 0 \tag{3.12}$$

Définition 3.7 *Une méthode numérique* ***convergente*** *est dite* ***d'ordre*** p, $p > 0$, *si* $C(h) = \mathcal{O}(h^p)$.

Il convient de remarquer que les relations (3.11) et (3.12) assurent que :

$$\lim_{N \to \infty} |y_N - u_N| = \lim_{N \to \infty} |y(T) - u_N| = \lim_{N \to \infty} e_N = 0 \tag{3.13}$$

Les méthodes numériques utilisées pour la résolution des équations différentielles ordinaires du premier ordre peuvent être classifiées selon divers critères. On peut ainsi distinguer :

- des **méthodes à un pas** (appelées aussi méthodes **à pas unique** ou méthodes **à pas séparés**) qui calculent la valeur approchée u_{n+1} au noeud t_{n+1} en utilisant seulement des informations concernant le noeud précédent t_n ;
- des **méthodes multipas** (appelées aussi méthodes **à pas liés**) qui calculent la valeur approchée u_{n+1} au noeud t_{n+1} en utilisant des informations concernant 2 ou plusieurs noeuds précédents ;
- des **méthodes explicites** qui calculent la valeur approchée u_{n+1} au noeud t_{n+1} à l'aide d'une formule explicite (qui ne fait pas intervenir la valeur à calculer u_{n+1} dans le membre de droite) ;
- des **méthodes implicites** qui calculent la valeur approchée u_{n+1} au noeud t_{n+1} à l'aide d'une formule implicite (qui fait intervenir la valeur à calculer u_{n+1} aussi dans le membre de droite).

On présente ci-dessous quelques exemples de formules associées à différents schémas numériques :

- méthode **explicite à un pas** :

$$u_{n+1} = \Phi\left(t_n, u_n, t_{n+1}\right) \tag{3.14}$$

- méthode **implicite à un pas** :

$$u_{n+1} = \Psi\left(t_n, u_n, t_{n+1}, u_{n+1}\right) \tag{3.15}$$

- méthode **explicite** multipas **à** 3 **pas** :

$$u_{n+1} = \Phi\left(t_{n-2}, u_{n-2}, t_{n-1}, u_{n-1}, t_n, u_n, t_{n+1}\right) \tag{3.16}$$

- méthode **implicite** multipas **à** 3 **pas** :

$$u_{n+1} = \Psi\left(t_{n-2}, u_{n-2}, t_{n-1}, u_{n-1}, t_n, u_n, t_{n+1}, u_{n+1}\right) \tag{3.17}$$

où les fonctions Φ et Ψ dépendent à la fois du schéma numérique choisi pour la résolution et de la fonction f donnée mais spécifique à chaque problème (3.7) à résoudre.

Il convient de remarquer que :

- pour entamer une méthode multipas à r pas, il faut connaître, à part la valeur initiale $u_0 = y_0$, encore $r - 1$ valeurs de départ qui sont, d'habitude, calculées numériquement à l'aide d'une méthode adéquate à un pas (ayant le même ordre de précision que la méthode multipas) ;
- l'utilisation d'une méthode implicite nécessite, à chaque noeud t_{n+1}, la résolution d'une équation non linéaire en u_{n+1}, résolution qui est faite d'habitude à l'aide d'une méthode numérique (par exemple, une méthode de point fixe).

Parfois, il est avantageux de remplacer la résolution de l'équation non linéaire évoquée ci-dessus par le calcul explicite d'une estimation $\tilde{u}_{n+1}$ de u_{n+1} qui sera utilisée ensuite dans l'expression de la fonction Ψ du membre de droite des équations (3.15) ou (3.17). On obtient ainsi des **méthodes** dites **prédicteur-correcteur** (ou de **prédiction-correction**) qui combinent un schéma explicite (qui représente la phase de prédiction) avec un schéma (à l'origine) implicite (qui représente la phase de correction).

Une catégorie spéciale de méthodes à un pas regroupe les **méthodes** (explicites ou implicites) dites **de Runge-Kutta**. Pour une partition régulière de pas h, toute méthode de Runge-Kutta correspond à un schéma général de la forme :

$$u_{n+1} = u_n + h \sum\nolimits_{i=1}^{s} b_i K_i \tag{3.18}$$

où s est le nombre d'étapes de la méthode et les s "pentes" K_i sont données, pour chaque i, par :

$$K_i = f\left(t_n + c_i h,\ u_n + h\sum_{j=1}^{s} a_{ij} K_j\right) \tag{3.19}$$

Les coefficients b_i, a_{ij} et $c_i = \sum_{j=1}^{s} a_{ij}$ sont donnés, d'habitude, pour chaque méthode particulière de Runge-Kutta, dans un **tableau** dit **de Butcher** qui a la forme :

$$\begin{array}{c|c} c_i & a_{ij} \\ \hline & b_i \end{array} \tag{3.20}$$

On note souvent une méthode de Runge-Kutta d'ordre p par RKp.

3.3 Méthodes numériques à un pas

Pour un sous-intervalle quelconque $[t_n, t_{n+1}]$, $n = 0, 1, ..., N-1$, d'une partition régulière σ_{reg}, on reprend la relation (3.10) :

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} f\left(t, y\left(t\right)\right) dt \tag{3.21}$$

et on calcule numériquement l'intégrale définie $\int_{t_n}^{t_{n+1}} f\left(t, y\left(t\right)\right) dt$ à l'aide de différentes formules de quadrature (non composites).

3.3.1 Méthode d'Euler progressive

On utilise la formule de **quadrature** (non composite) **du "point de gauche"** ou du "rectangle à gauche". Ainsi, $\int_{t_n}^{t_{n+1}} f\left(t, y\left(t\right)\right) dt$ est approchée par :

$$J^{PG}\left(f\right) = \left(t_{n+1} - t_n\right) f\left(t_n, y\left(t_n\right)\right) = h f\left(t_n, y\left(t_n\right)\right) \tag{3.22}$$

Etant donné que, sauf pour le cas $n = 0$, on ne connaît pas la valeur exacte $y\left(t_n\right)$ au noeud précédent t_n, mais la valeur approchée u_n, en combinant les relations (3.21) et (3.22) on obtient :

$$\begin{cases} u_{n+1} = u_n + h f_n \\ u_0 = y_0 \end{cases} \tag{3.23}$$

Les relations (3.23), valables pour $n = 0, 1, ..., N-1$, correspondent au schéma appelé schéma d'**Euler progressif** ("**forward**" en anglais).

Il s'agit d'une méthode **explicite à un pas** qui peut être regardée aussi comme une méthode de **Runge-Kutta explicite à une étape**.

3.3.2 Méthode d'Euler rétrograde

On utilise la formule de **quadrature** (non composite) **du "point de droite"** ou du "rectangle à droite". Ainsi, $\int_{t_n}^{t_{n+1}} f(t, y(t))\, dt$ est approchée par :

$$J^{PD}(f) = (t_{n+1} - t_n) f(t_{n+1}, y(t_{n+1})) = hf(t_{n+1}, y(t_{n+1})) \tag{3.24}$$

En procédant comme pour le schéma d'Euler progressif, on obtient cette fois :

$$\begin{cases} u_{n+1} = u_n + hf_{n+1} \\ u_0 = y_0 \end{cases} \tag{3.25}$$

Les relations (3.25), valables pour $n = 0, 1, ..., N-1$, correspondent au schéma appelé schéma d'**Euler rétrograde** ("**backward**" en anglais).

Il s'agit toujours d'une méthode **à un pas**, mais **implicite** car $f_{n+1} = f(t_{n+1}, u_{n+1})$.

3.3.3 **Méthode de Crank-Nicolson*

On utilise la formule de **quadrature** (non composite) **du trapèze** ou de la sécante. Ainsi, $\int_{t_n}^{t_{n+1}} f(t, y(t))\, dt$ est approchée par :

$$J^{Tr}(f) = (t_{n+1} - t_n) \frac{f(t_n, y(t_n)) + f(t_{n+1}, y(t_{n+1}))}{2} = h\frac{f(t_n, y(t_n)) + f(t_{n+1}, y(t_{n+1}))}{2} \tag{3.26}$$

En procédant comme pour les schémas d'Euler, on obtient :

$$\begin{cases} u_{n+1} = u_n + \dfrac{h}{2}(f_n + f_{n+1}) \\ u_0 = y_0 \end{cases} \tag{3.27}$$

Les relations (3.27), valables pour $n = 0, 1, ..., N-1$, correspondent au schéma appelé schéma de **Crank-Nicolson** qui représente la moyenne des schémas d'Euler progressif et rétrograde.

Il s'agit d'une méthode **à un pas** qui est **implicite** car $f_{n+1} = f(t_{n+1}, u_{n+1})$.

3.3.4 *Méthode de Heun

On vient de voir que la méthode de Crank-Nicolson est une méthode implicite qui nécessite la résolution numérique de la première équation (3.27) qui est non linéaire. Afin d'éviter ce calcul qui est assez "cher" du point de vue numérique, on peut envisager une **approche** de type **prédicteur-correcteur** :

- Dans la phase de **prédiction**, on calcule, à l'aide de la méthode explicite d'Euler progressive, une première estimation :

$$\tilde{u}_{n+1} = u_n + hf_n \tag{3.28}$$

- Dans la phase de **correction**, on remplace f_{n+1} par $\tilde{f}_{n+1} = f\left(t_{n+1}, \tilde{u}_{n+1}\right)$, ce qui fait que la première équation (3.27) devient explicite :

$$\begin{cases} u_{n+1} = u_n + \dfrac{h}{2}\left(f_n + \tilde{f}_{n+1}\right) \\ u_0 = y_0 \end{cases} \tag{3.29}$$

Les relations (3.29), valables pour $n = 0, 1, ..., N-1$, correspondent au schéma appelé schéma de **Heun** (ou de la **tangente améliorée**).

La méthode de Heun peut être vue aussi comme une méthode de **Runge-Kutta explicite à** 2 **étapes** dont les coefficients s'écrivent sous la forme :

$$K_1 = f\left(t_n, u_n\right) \tag{3.30}$$

$$K_2 = f\left(t_n + h, u_n + hK_1\right) \tag{3.31}$$

Ainsi, le schéma de Heun devient :

$$\begin{cases} u_{n+1} = u_n + \dfrac{h}{2}\left(K_1 + K_2\right) \\ u_0 = y_0 \end{cases} \tag{3.32}$$

3.3.5 *Méthode d'Euler modifiée (ou améliorée)

On utilise la formule de **quadrature** (non composite) **du point milieu** ou du rectangle. Ainsi, l'intégrale $\int_{t_n}^{t_{n+1}} f\left(t, y\left(t\right)\right) dt$ est approchée par :

$$J^{PM}\left(f\right) = h\ f\left(t_n + \frac{h}{2}, y\left(t_n + \frac{h}{2}\right)\right) \tag{3.33}$$

On ne peut plus procéder comme pour les trois schémas précédents, car on ne connaît pas la valeur $u_{n+\frac{1}{2}}$ qui correspond au point milieu $t_n + \frac{h}{2}$.

Cependant, on peut approcher cette valeur par un schéma d'Euler progressif, c'est-à-dire :

$$u_{n+\frac{1}{2}} = u_n + \frac{h}{2} f_n \tag{3.34}$$

On obtient ainsi :

$$\begin{cases} u_{n+1} = u_n + h\ f_{n+\frac{1}{2}} \\ u_0 = y_0 \end{cases} \tag{3.35}$$

où $f_{n+\frac{1}{2}} = f\left(t_n + \frac{h}{2}, u_{n+\frac{1}{2}}\right)$.

Les relations (3.35), valables pour $n = 0, 1, ..., N-1$, correspondent au schéma appelé schéma d'**Euler modifié** ou **amélioré** (ou, encore, schéma de **Runge** ou de **Runge-Kutta à 2 étapes "classique"**).

Il s'agit finalement d'une méthode **explicite à un pas** qui peut être vue comme une variante "simplifiée" d'une méthode de type prédicteur-correcteur. En effet, dans la phase de **prédiction**, on calcule la valeur approchée au point milieu $u_{n+\frac{1}{2}}$ par un schéma explicite et, dans la phase de **correction**, on calcule la valeur approchée "définitive" au point t_{n+1} par un schéma qui était initialement "semi-implicite".

De plus, la méthode d'Euler modifiée est une méthode de **Runge-Kutta explicite à 2 étapes** dont les coefficients s'écrivent sous la forme :

$$K_1 = f(t_n, u_n) \tag{3.36}$$

$$K_2 = f\left(t_n + \frac{1}{2}h, u_n + \frac{h}{2}K_1\right) \tag{3.37}$$

Ainsi, la méthode d'Euler améliorée devient :

$$\begin{cases} u_{n+1} = u_n + hK_2 \\ u_0 = y_0 \end{cases} \tag{3.38}$$

3.3.6 Méthode classique de Runge-Kutta

On utilise la formule de **quadrature** (non composite) de **Simpson**. Ainsi, l'intégrale $\int_{t_n}^{t_{n+1}} f(t, y(t))\, dt$ est approchée par :

$$J^S(f) = \frac{h}{6}\left[f(t_n, y(t_n)) + 4f\left(t_n + \frac{h}{2}, y\left(t_n + \frac{h}{2}\right)\right) + f(t_{n+1}, y(t_{n+1}))\right] \tag{3.39}$$

Comme pour la méthode d'Euler modifiée, on ne connaît pas la valeur $u_{n+\frac{1}{2}}$ qui

correspond au point milieu $t_n + \frac{h}{2}$. De plus, on veut obtenir une **formule explicite** qui ne contient pas $f_{n+1} = f(t_{n+1}, u_{n+1})$.

Dans ces conditions, on peut faire plusieurs choix mais le choix qui correspond à la célébre formule originelle de Runge-Kutta est le suivant :

- le coefficient (la pente) K_1 correspond au "point de gauche" :

$$K_1 = f(t_n, u_n) \tag{3.40}$$

- le coefficient (la pente) K_2 (qui aura dans la formule finale la moitié du poids 4) correspond au point milieu $t_n + \frac{h}{2}$ et sera obtenu(e) à partir du noeud t_n par une **prédiction**, en utilisant le schéma d'Euler progressif avec K_1 :

$$K_2 = f\left(t_n + \frac{h}{2}, u_n + \frac{h}{2}K_1\right) \tag{3.41}$$

- le coefficient (la pente) K_3 (qui aura dans la formule finale l'autre moitié du poids 4) correspond également au point milieu $t_n + \frac{h}{2}$ et sera obtenu(e) à partir du noeud t_n par une **correction**, en utilisant le schéma d'Euler "rétrograde" rendu explicite par l'emploi de (la pente) K_2 :

$$K_3 = f\left(t_n + \frac{h}{2}, u_n + \frac{h}{2}K_2\right) \tag{3.42}$$

- le coefficient (la pente) K_4 correspond au point "d'arrivée" t_{n+1} et sera obtenu(e) à partir du point de départ t_n, en utilisant le schéma d'Euler progressif avec la pente K_3 (qui est plus précise que la pente K_2) :

$$K_4 = f(t_{n+1}, u_n + hK_3) \tag{3.43}$$

Finalement, on obtient une formule de **Runge-Kutta à** 4 **étapes** qui est appelée la **formule classique de Runge-Kutta** et qui s'écrit ainsi :

$$\begin{cases} u_{n+1} = u_n + \frac{h}{6}(K_1 + 2K_2 + 2K_3 + K_4) \\ u_0 = y_0 \end{cases} \tag{3.44}$$

Le **tableau de Butcher** associé à la méthode classique de Runge-Kutta d'ordre 4 est donné ci-dessous :

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array} \tag{3.45}$$

Il convient de remarquer que, pour le **cas particulier** d'une fonction f qui ne dépend que du temps, $K_2 = K_3$ et la formule (3.44) coïncide rigoureusement avec la formule de quadrature non composite de Simpson.

3.4 Estimation d'erreur

La conception des méthodes numériques fiables destinées à résoudre des équations différentielles ordinaires dépend, en grande mesure, de la capacité d'estimer et de contrôler les erreurs.

On considère un problème de Cauchy (3.7) qui admet une solution unique $y = y(t)$ de classe $C^1(I \subset \mathbb{R})$ et une méthode de résolution numérique qui lui est associée. Soit t_n, $n = 1, ...N$, un point courant d'une partition régulière de l'intervalle d'étude **borné** $I = [t_0, T]$ et d_n la différence (positive ou négative) entre la solution exacte et la solution approchée au moment t_n :

$$d_n = y_n - u_n \tag{3.46}$$

La différence d_n est due :

- d'une part, aux **erreurs** dites **d'arrondi** qui sont inévitablement liées à la façon de stocker l'information, en général, et les valeurs numériques, en particulier, dans les ordinateurs ;
- d'autre part, aux **erreurs** dites **de troncature** qui sont dues à la méthode numérique de résolution.

Quoique par la suite on ne prenne pas en compte les **erreurs d'arrondi**, il convient de préciser que ces erreurs limitent la finesse des partitions, dans le sens où un pas de discrétisation trop petit, par exemple inférieur à une valeur minimale $h_{\min}$:

$$h < h_{\min} \tag{3.47}$$

peut entraîner l'augmentation inacceptable, voir l'explosion, de l'erreur de calcul.

Quant aux **erreurs de troncature**, on peut identifier, dans la résolution numériques des équations différentielles (à l'aide des méthodes à un pas), deux contributions distinctes :

- une erreur de troncature dite **locale** qui se produit, suite à une seule itération de la méthode numérique, au passage de la solution exacte y_{n-1} au point t_{n-1} à la solution approchée $\widetilde{u}_n$ au point suivant t_n ;
- une erreur de troncature dite **transportée** qui représente la propagation entre t_{n-1} et t_n de la contribution de toutes les erreurs locales accumulées précédemment, c'est-à-dire produites et transmises entre t_0 et t_{n-1}.

Plus précisément, on peut récrire la différence d_n donnée par (3.46) sous la forme :

$$d_n = (y_n - \widetilde{u}_n) + (\widetilde{u}_n - u_n) \tag{3.48}$$

où :

- $\widetilde{u}_n$ est la solution approchée au point t_n obtenue à partir de la solution exacte au point précédent t_{n-1}, en effectuant une seule itération, ce qui donne, par exemple, pour le cas de la méthode d'Euler explicite :

$$\widetilde{u}_n = y_{n-1} + hf\,(t_{n-1}, y_{n-1}) \tag{3.49}$$

- $(y_n - \widetilde{u}_n)$ est la différence (positive ou négative) entre la valeur exacte au point t_n et la valeur approchée obtenue au même point, par une seule itération, à partir de la solution exacte au point précédent t_{n-1} ;
- $(\widetilde{u}_n - u_n)$ est la différence (positive ou négative) entre la valeur approchée calculée au point t_n, par une seule itération, à partir de la solution exacte au point précédent t_{n-1} et la valeur approchée obtenue au même point t_n en appliquant la méthode numérique, avec toutes les itérations nécessaires, à partir du point initial t_0 où on connaît la valeur initiale y_0.

Autrement dit :

- la valeur absolue $|y_n - \widetilde{u}_n|$ est justement l'erreur de **troncature locale absolue** (obtenue en "injectant" la solution exacte dans le schéma numérique) ;
- la valeur absolue $|\widetilde{u}_n - u_n|$ est justement l'erreur de **troncature transportée absolue**.

Ainsi, si **on néglige l'erreur d'arrondi**, on peut récrire l'expression de l'erreur totale de calcul e_n au point courant t_n :

$$\begin{aligned} e_n &= |y_n - u_n| \\ &= |(y_n - \widetilde{u}_n) + (\widetilde{u}_n - u_n)| \\ &\leq |y_n - \widetilde{u}_n| + |\widetilde{u}_n - u_n| \end{aligned} \tag{3.50}$$

Par conséquent, si, pour $\forall n = 1, ..., N$, les deux erreurs de troncature tendent vers zéro quand le pas de discrétisation h tend vers zéro, alors l'erreur totale de calcul e_n tend forcément vers zéro et la **méthode numérique** est **convergente**.

3.5 Consistance, stabilité, convergence

Afin d'introduire les notions de consistance et stabilité qui sont essentielles pour la compréhension de la convergence des schémas numériques, on considère à nouveau un problème de Cauchy :

$$\begin{cases} y'(t) = f(t, y(t)) & \forall t \in I \\ y(t_0) = y_0 \end{cases} \tag{3.51}$$

qui admet une solution unique.

De plus, on suppose que la fonction $f(t, y)$ est **lipschitzienne** par rapport à y et que la solution $y(t)$ est **suffisamment régulière**, à savoir deux fois continûment différentiable (i.e. $y \in C^2(I)$, où $I = [t_0, T]$).

Afin de résoudre numériquement le problème (3.51), on choisit la méthode d'**Euler progressive** (qui est une méthode explicite à un pas) et on analyse l'erreur absolue e_n au point courant t_n, $n = 1, 2, ..., N$.

En fait, on veut trouver d'abord des estimations pour l'erreur de troncature locale et pour l'erreur de troncature transportée.

3.5.1 Erreur de troncature locale et consistance

Etant donnée que $y \in C^2(I)$, on peut utiliser la **formule de Taylor** pour la solution cherchée au moment t_n :

$$y_n = y_{n-1} + hy'_{n-1} + \frac{1}{2}h^2 y''(\theta_n) \tag{3.52}$$

$$= y_{n-1} + hf(t_{n-1}, y_{n-1}) + \frac{1}{2}h^2 y''(\theta_n) \tag{3.53}$$

$$= \widetilde{u}_n + \frac{1}{2}h^2 y''(\theta_n) \tag{3.54}$$

où $\theta_n \in]t_{n-1}, t_n[$.

Ci-dessus, on a utilisé :

- pour le passage de (3.52) à (3.53), la relation :

$$y'_{n-1} = f(t_{n-1}, y_{n-1}) \tag{3.55}$$

  qui correspond à l'équation différentielle à résoudre ;

- pour le passage de (3.53) à (3.54), la relation :

$$\widetilde{u}_n = y_{n-1} + hf\left(t_{n-1}, y_{n-1}\right) \tag{3.56}$$

  qui correspond au fait que $\widetilde{u}_n$ est la solution approchée au point t_n obtenue numériquement par le schéma d'Euler progressif à partir de la solution exacte au point précédent t_{n-1}.

La relation (3.54) nous permet de calculer l'**erreur de troncature locale** (**absolue**) au point t_n, à savoir :

$$|y_n - \widetilde{u}_n| = \frac{1}{2}h^2 \left|y''\left(\theta_n\right)\right| \tag{3.57}$$

Définition 3.8 *Le quotient :*

$$\tau_n\left(h\right) = \frac{|y_n - \widetilde{u}_n|}{h} \tag{3.58}$$

est appelé ***erreur de troncature locale unitaire****.*

Définition 3.9 *Pour une partition donnée, l'****erreur de troncature locale unitaire maximale*** *(appelée aussi* ***globale****)* $\tau\left(h\right)$ *représente la plus grande erreur de troncature locale unitaire :*

$$\tau\left(h\right) = \max_{n=1,\ldots,N} \tau_n\left(h\right) \tag{3.59}$$

Pour le schéma d'Euler progressif, on obtient :

- l'expression de l'erreur de troncature locale unitaire, en remplaçant (3.57) en (3.58) :

$$\tau_n\left(h\right) = \frac{1}{2}h\left|y''\left(\theta_n\right)\right| \tag{3.60}$$

- l'expression de l'erreur de troncature locale unitaire maximale, en remplaçant (3.60) en (3.59) :

$$\tau\left(h\right) = \frac{1}{2}hM \tag{3.61}$$

  où :

$$M = \max_{t\in]t_0,T[} \left|y''\left(t\right)\right| \tag{3.62}$$

Définition 3.10 *Une méthode numérique de résolution d'équations différentielles ordinaires est dite* ***consistante*** *si :*

$$\lim_{h\to 0} \tau(h) = 0 \tag{3.63}$$

Définition 3.11 *Une méthode numérique consistante est dite* ***consistante d'ordre*** *p, $p \geq 1$, si :*

$$\tau(h) = \mathcal{O}(h^p) \tag{3.64}$$

Il convient de remarquer que la relation (3.61) montre que l'erreur de troncature unitaire maximale tend vers zéro quand le pas de discrétisation tend vers zéro, ce qui revient à dire que la méthode d'**Euler progressive** est consistante et, de plus, cette méthode est **consistante d'ordre 1**.

Remarque 3.4 *La* ***consistance*** *est une* ***condition nécessaire*** *pour la* ***convergence*** *(car l'accumulation des erreurs de troncature locales unitaires qui ne tendent pas vers zéro avec la finesse de la partition, empêcherait l'erreur de calcul maximale de tendre vers zéro quand le pas de discrétisation tend vers zéro et $N = \dfrac{T-t_0}{h} \to \infty$).*

3.5.2 Erreur de troncature transportée

On calcule maintenant la différence (positive ou négative) correspondant à l'**erreur de troncature transportée** pour le schéma d'**Euler progressif** :

$$\begin{aligned} \widetilde{u}_n - u_n &= (y_{n-1} + hf(t_{n-1}, y_{n-1})) - (u_{n-1} + hf(t_{n-1}, u_{n-1})) \\ &= (y_{n-1} - u_{n-1}) + h(f(t_{n-1}, y_{n-1}) - f(t_{n-1}, u_{n-1})) \\ &= d_{n-1} + h(f(t_{n-1}, y_{n-1}) - f(t_{n-1}, u_{n-1})) \end{aligned} \tag{3.65}$$

Vu que la fonction $f(t, y)$ est lipschitzienne par rapport à y, on a que :

$$\begin{aligned} |f(t_{n-1}, y_{n-1}) - f(t_{n-1}, u_{n-1})| &\leq L|y_{n-1} - u_{n-1}| \\ &= Le_{n-1} \end{aligned} \tag{3.66}$$

où L est une constante réelle (strictement) positive.

Grâce aux relations (3.65) et (3.66), on peut trouver un majorant pour l'**erreur de troncature transportée** :

$$\begin{aligned} |\widetilde{u}_n - u_n| &\leq |d_{n-1} + hLe_{n-1}| \\ &\leq |d_{n-1}| + hLe_{n-1} \\ &= (1 + hL)\, e_{n-1} \end{aligned} \tag{3.67}$$

où on a tenu compte que $|d_{n-1}| = e_{n-1}$.

3.5.3 Erreur de calcul

On a vu que, si **on néglige l'erreur d'arrondi**, l'erreur (totale) de calcul e_n au point t_n est due seulement aux deux contributions de l'**erreur de troncature**. De plus, on avait obtenu la relation (3.50) qui, pour le cas du schéma d'**Euler progressif**, peut être complétée ainsi :

$$\begin{aligned}
e_n &\leq |y_n - \widetilde{u}_n| + |\widetilde{u}_n - u_n| \\
&\leq h\tau_n + (1+hL)\, e_{n-1} \\
&\leq h\tau + (1+hL)\, e_{n-1} \\
&\leq h\tau + (1+hL)\,(h\tau + (1+hL)\, e_{n-2}) \\
&= h\tau + h\tau\,(1+hL) + (1+hL)^2\, e_{n-2} \\
&\leq h\tau\left(1 + (1+hL) + ... + (1+hL)^{n-1}\right) \\
&= \tau\frac{(1+hL)^n - 1}{L} \\
&\leq \tau\frac{\exp\left(L\,(t_n - t_0)\right) - 1}{L} \\
&= \frac{1}{2}hM\frac{\exp\left(L\,(t_n - t_0)\right) - 1}{L}
\end{aligned} \tag{3.68}$$

où on a tenu compte que $e_0 = 0$, $\tau = \frac{1}{2}hM$ et on a utilisé les relations :

$$1 + hL \leq \exp(hL) \tag{3.69}$$

$$\begin{aligned}
(1+hL)^n &\leq \exp(hLn) \\
&= \exp\left(hL\frac{(t_n - t_0)}{h}\right) \\
&= \exp\left(L\,(t_n - t_0)\right)
\end{aligned} \tag{3.70}$$

La relation (3.68) montre que la méthode d'**Euler progressive** est **convergente d'ordre 1 en** h.

Remarque 3.5 *Il convient de mentionner que, pour la méthode d'Euler progressive, l'ordre de convergence est le même que l'ordre de consistance (et cette propriété s'avère valable aussi pour d'autres méthodes numériques de résolution d'équations différentielles ordinaires).*

3.5.4 Stabilité et erreur de troncature transportée

L'étude antérieure de l'erreur de troncature transportée a été basée sur le fait que la fonction $f(t, y)$ était lipschitzienne. Mais, dans beaucoup de cas pratiques, la fonction $f(t, y)$ admet une **dérivée partielle selon** y qui est **non positive**, i.e. $\dfrac{\partial f}{\partial y} \leq 0$ pour

$\forall t \in I$ et $\forall y \in \mathbb{R}$.

Dans ces conditions, on peut écrire :

$$\begin{aligned} f\left(t_{n-1}, u_{n-1}\right) &= f\left(t_{n-1}, y_{n-1}\right) + \left(u_{n-1} - y_{n-1}\right)\frac{\partial f}{\partial y}\left(t_{n-1}, \zeta_{n-1}\right) \\ &= f\left(t_{n-1}, y_{n-1}\right) - d_{n-1}\frac{\partial f}{\partial y}\left(t_{n-1}, \zeta_{n-1}\right) \end{aligned} \tag{3.71}$$

où ζ_{n-1} est un point appartenant à l'intervalle ouvert d'extrémités y_{n-1} et u_{n-1}.

Grâce à cette dernière relation, on peut exprimer la différence (positive ou négative) correspondant à l'erreur de troncature transportée (3.65) pour le schéma d'**Euler progressif** sous la forme :

$$\begin{aligned} \widetilde{u}_n - u_n &= d_{n-1} + h\left(f\left(t_{n-1,}y_{n-1}\right) - f\left(t_{n-1}, u_{n-1}\right)\right) \\ &= d_{n-1}\left(1 + h\frac{\partial f}{\partial y}\left(t_{n-1}, \zeta_{n-1}\right)\right) \end{aligned} \tag{3.72}$$

L'erreur de troncature transportée devient alors :

$$\left|\widetilde{u}_n - u_n\right| = e_{n-1}\left|1 + h\frac{\partial f}{\partial y}\left(t_{n-1}, \zeta_{n-1}\right)\right| \tag{3.73}$$

Vu que, par l'hypothèse, $\dfrac{\partial f}{\partial y} \leq 0$, cette dernière expression montre qu'on peut satisfaire l'inégalité :

$$\left|\widetilde{u}_n - u_n\right| \leq e_{n-1} \tag{3.74}$$

si le pas de discrétisation remplit une **condition** dite **de stabilité locale**, à savoir :

$$h \leq \frac{2}{\left|\dfrac{\partial f}{\partial y}\left(t_{n-1}, \zeta_{n-1}\right)\right|} \tag{3.75}$$

En fait, par l'intermédiaire de l'inégalité (3.74), la condition de stabilité locale (3.75) assure que l'erreur de calcul (totale) au point t_{n-1} ne s'amplifie pas quand elle se propage au point suivant t_n.

Pour que la stabilité locale soit garantie pour $\forall n = 1, 2, ..., N$, il suffit d'imposer une **condition de stabilité globale**, à savoir :

$$h \leq \frac{2}{\max\limits_{t \in [t_0, T]}\left|\dfrac{\partial f}{\partial y}\left(t, y\left(t\right)\right)\right|} \tag{3.76}$$

On peut donc affirmer que la méthode d'**Euler progressive** est **conditionnellement stable** (par opposition à des méthodes qui sont, comme la méthode d'Euler rétrograde,

inconditionnellement stables, dans le sens où la stabilité globale est assurée "automatiquement", c'est-à-dire sans aucune limitation du pas de discrétisation).

Si la condition de stabilité globale (3.76) est satisfaite, l'erreur (totale) de calcul pour le schéma d'Euler progressif peut être majorée de la façon suivante :

$$\begin{aligned} e_n &\leq |y_n - \widetilde{u}_n| + |\widetilde{u}_n - u_n| \\ &\leq |y_n - \widetilde{u}_n| + e_{n-1} \\ &= h\tau_n + e_{n-1} \\ &\leq h\tau + e_{n-1} \\ &\leq h\tau + h\tau + e_{n-2} \\ &\leq nh\tau + e_0 \\ &= nh\tau \\ &= nh\frac{1}{2}hM \\ &= (t_n - t_0)\,\frac{1}{2}hM \end{aligned} \tag{3.77}$$

où on a tenu compte que $e_0 = 0$, $\tau = \frac{1}{2}hM$ et $nh = t_n - t_0$.

On retrouve bien le fait que le schéma d'**Euler progressif** est **convergent d'ordre 1 en** h.

3.5.5 **Précisions supplémentaires*

On considère toujours un problème de Cauchy (3.51) défini sur un intervalle **borné** $I = [t_0, T]$.

D'une manière générale, la notion de **consistance** est liée à la **discrétisation** de l'intervalle d'étude à l'aide d'une partition plus ou moins fine. Ainsi, si on utilise une méthode numérique, à un pas ou à multipas, et on part de la solution exacte au point précédent ou, respectivement, aux points précédents, l'itération effectuée produira (en général, de manière inévitable) une **erreur** dite de **troncature locale**.

Par contre, on veut que cette erreur et, surtout, le quotient entre l'erreur de troncature locale et le pas de discrétisation (i.e. l'erreur de troncature locale unitaire) tendent vers zéro partout quand les partitions deviennent de plus en plus fines, i.e. quand le pas de discrétisation $h \to 0$ et le nombre de sous-intervalles $N \to \infty$.

Ceci assure en fait la **consistance** de la méthode numérique qui est une **condition nécessaire pour la convergence**.

Quant à la notion de **stabilité**, elle peut être vue comme la propriété de la **solution** (exacte et/ou numérique) de l'équation différentielle ordinaire de rester **bornée**.

Du point de vue numérique, on fait une distinction entre la **zéro-stabilité**, qui est

spécifique aux intervalles d'étude bornés, et la **stabilité absolue** (ou **A-stabilité**), qui est spécifique aux intervalles d'étude non bornés (pour lesquels le nombre de points d'une partition peut tendre vers l'infini sans que le pas de discrétisation tende nécessairement vers zéro).

La zéro-stabilité assure que des petites perturbations de la condition initiale se propagent à travers les itérations du schéma numérique sans (trop) s'amplifier, de sorte que la solution numérique perturbée reste bornée partout.

Pour une méthode numérique **à un pas** qui est **consistante**, on peut montrer (voir, par exemple, Quarteroni et Saleri) que la **zéro-stabilité** est garantie si la fonction $f(t, y)$ est **lipschitzienne** par rapport à y.

D'une manière plus générale, ajoutée à la condition nécessaire de consistance, la zéro-stabilité s'avère être une condition nécessaire et suffisante pour la convergence de la méthode numérique.

Théorème 3.2 (d'équivalence de Lax-Ritchmyer) *Toute méthode numérique* ***consistante*** *est* ***convergente si et seulement si*** *elle est* ***zéro-stable****.*

La démarche utilisée pour l'étude de la consistance de la méthode d'Euler progressive peut être répétée pour d'autre méthodes numériques. En fait, on effectue un développement de l'équation discrétisée par la formule de Taylor jusqu'à un ordre convenable et on vérifie que l'équation discrétisée tend vers l'équation différentielle ordinaire originale quand le pas de discrétisation tend vers zéro. Ainsi, par exemple :

- pour la méthode d'**Euler rétrograde**, si la solution exacte est deux fois continûment différentiable (i.e. $y \in C^2([t_0, T])$), on trouve pour l'erreur de troncature locale unitaire l'expression :

$$\tau_n(h) = \frac{h}{2} |y''(\theta_n)| \tag{3.78}$$

où $\theta_n \in]t_{n-1}, t_n[$, et pour l'erreur de troncature locale unitaire maximale l'expression :

$$\tau(h) = \frac{h}{2} \max_{t \in]t_0, T[} |y''(t)| \tag{3.79}$$

La méthode d'**Euler rétrograde** est donc **consistante d'ordre 1** et, de plus, aussi **convergente d'ordre 1 en** h.

- pour la méthode de **Crank-Nicolson**, si la solution exacte est trois fois continûment différentiable (i.e. $y \in C^3([t_0, T])$), on trouve pour l'erreur de troncature locale unitaire l'expression :

$$\tau_n(h) = \frac{h^2}{12} |y'''(\theta_n)| \tag{3.80}$$

où $\theta_n \in]t_{n-1}, t_n[$, et pour l'erreur de troncature locale unitaire maximale l'expression :

$$\tau(h) = \frac{h^2}{12} \max_{t\in]t_0,T[} |y'''(t)| \tag{3.81}$$

La méthode de **Crank-Nicolson** est donc **consistante d'ordre 2** et, de plus, aussi **convergente d'ordre 2 en** h.

En ce qui concerne la stabilité, les méthodes d'**Euler rétrograde** et de **Crank-Nicolson** sont des méthodes **implicites inconditionnellement stables**.

Pour les méthodes de type **prédicteur-correcteur**, d'une manière générale, l'ordre de la consistance et de la convergence est donné par la partie correction, tandis que la stabilité est tributaire à la partie prédiction. Ainsi, par exemple :

- pour la méthode de **Heun**, l'ordre de **consistance** et **de convergence** est **2** car il est donné par le correcteur qui est un schéma de Crank-Nicolson correspondant à la formule de quadrature du trapèze, tandis que la méthode est **conditionnellement stable** car la prédiction se fait par un schéma d'Euler explicite ;

- pour la méthode d'**Euler améliorée**, l'ordre de **consistance** et de **convergence** est **2** car il est donné par le correcteur qui correspond à la formule de quadrature du point milieu, tandis que la méthode est **conditionnellement stable** car la prédiction se fait par un schéma d'Euler explicite.

Quant à la méthode classique de **Runge-Kutta à 4 étapes** qui est une méthode explicite à un pas, la **convergence est d'ordre 4 en** h.

D'une manière générale, pour une méthode convergente d'ordre p, $p \geq 1$, l'erreur tend vers zéro comme h^p. Ainsi, si on double la finesse de la partition en passant d'un pas h à un pas $\frac{h}{2}$, l'erreur va diminuer de 2^p.

3.6 **Méthodes numériques multipas*

Les méthodes **multipas**, appelées aussi méthodes **à pas liés**, sont basées sur des schémas qui calculent la valeur approchée numériquement u_{n+1} à un point t_{n+1} en fonction des informations connues en deux ou plusieurs points précédents.

A titre d'exemple, on cite ci-dessous plusieurs méthodes **multipas** qui sont toutes **consistantes** et **zéro-stables**.

3.6.1 Méthode de Nyström

La méthode de Nyström est une méthode multipas **explicite à 2 pas**, précisée par le schéma suivant :

$$u_{n+1} = u_{n-1} + 2hf\left(t_n, u_n\right) \tag{3.82}$$

où $n = 1, ..., N-1$.

Afin d'**amorcer** la méthode de Nyström, il faut calculer d'abord (en fonction de $u_0 = y_0$ connue) la valeur approchée u_1 par une méthode adéquate à un pas.

On peut montrer que la méthode de Nyström est **convergente d'ordre** 2.

Comme interprétation géométrique du schéma de Nyström, la valeur approchée u_{n+1} est l'ordonnée du point d'abscisse t_{n+1} situé sur la droite passant par le point (t_{n-1}, u_{n-1}) et de pente $f\left(t_n, u_n\right)$ (pente qui approche la tangente à la courbe intégrale au point (t_n, y_n)).

3.6.2 Méthodes d'Adams-Bashforth

Les méthodes d'Adams-Bashforth sont des méthodes multipas **explicites**.

Comme cas particulier, la méthode d'Adams-Bashforth **à un pas** est **convergente d'ordre 1** et coïncide, en fait, avec la méthode explicite d'**Euler progressive**.

Les méthodes d'Adams-Bashforth **à 2, 3** et **4 pas** sont **convergentes d'ordre 2, 3** et, respectivement, **4**. Ces méthodes sont appelées couramment AB2, AB3 et, respectivement, AB4 et elles correspondent aux schémas suivants :

$$AB2 \quad : \qquad u_{n+1} = u_n + \frac{h}{2}\left(3f\left(t_n, u_n\right) - f\left(t_{n-1}, u_{n-1}\right)\right) \tag{3.83}$$

$$AB3 \quad : \qquad u_{n+1} = u_n + \frac{h}{12}(23f\left(t_n, u_n\right) - 16f\left(t_{n-1}, u_{n-1}\right) + 5f\left(t_{n-2}, u_{n-2}\right)) \tag{3.84}$$

$$AB4 \quad : \qquad u_{n+1} = u_n + \frac{h}{24}(55f\left(t_n, u_n\right) - 59f\left(t_{n-1}, u_{n-1}\right) + 37f\left(t_{n-2}, u_{n-2}\right) - 9f\left(t_{n-3}, u_{n-3}\right)) \tag{3.85}$$

où n varie de 1, 2 et, respectivement, 3 à $N-1$.

Afin d'**amorcer** une méthode d'Adams-Bashforth AB2, AB3 ou AB4, il faut calculer d'abord (en fonction de $u_0 = y_0$ connue) les valeurs approchées u_1 ou u_1 et u_2 ou, respectivement, u_1, u_2 et u_3 par une (autre) méthode adéquate (le plus souvent, une méthode à un pas qui a, au moins, le même ordre de convergence).

A titre d'exemple, on présente ci-dessous une façon d'obtenir la formule (3.84), c'est-à-dire le schéma numérique correspondant à la méthode d'Adams-Bashforth à 3 pas.

Comme pour les méthodes numériques à un pas, on considère un sous-intervalle quel-

conque $[t_n, t_{n+1}]$, $n = 0, 1, ..., N-1$, d'une partition régulière σ_{reg} de pas h et on reprend la relation (3.10) :

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} f\left(t, y\left(t\right)\right) dt \tag{3.86}$$

Cependant, afin de calculer numériquement cette dernière intégrale définie (à partir de $n = 2$), on n'utilise plus une formule de quadrature non composite, mais on approche l'intégrant $f\left(t, y\left(t\right)\right)$ par un polynôme d'interpolation de degré 2 qui passe par le point "courant" $(t_n, f(t_n, y(t_n)))$ et les deux points précédents $(t_{n-1}, f(t_{n-1}, y(t_{n-1})))$ et $(t_{n-2}, f(t_{n-2}, y(t_{n-2})))$. Vu que les trois points mentionnés sont supposés "connus", cette démarche conduira bien à un schéma explicite.

Tout d'abord, sans perdre de généralité, on remplace la partition σ_{reg} par une partition régulière "canonique" de pas (unitaire) 1. Plus précisément, on fait le changement de variable :

$$t = t_n + h\tau \tag{3.87}$$

Il convient de remarquer que la nouvelle variables (temporelle) τ fait correspondre aux moments successifs $t = t_{n-2}$, $t = t_{n-1}$, $t = t_n$ et $t = t_{n+1}$ les "nouveaux" moments $\tau = -2$, $\tau = -1$, $\tau = 0$ et, respectivement, $\tau = 1$ (voir la Figure 3.1).

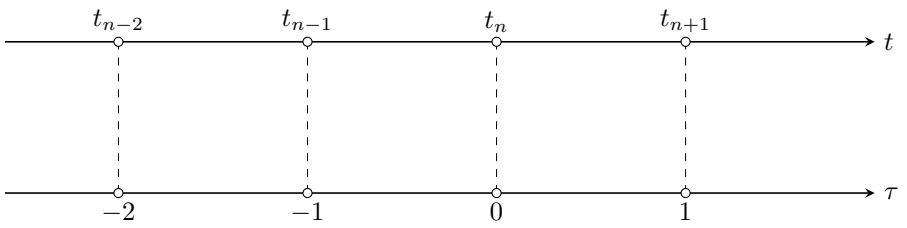

Figure 3.1: AB3 - Changement de variable (temporelle)

En notant :

$$f(t_n + h\tau, y(t_n + h\tau)) = g(\tau) \tag{3.88}$$

on obtient l'égalité :

$$\int_{t_n}^{t_{n+1}} f\left(t, y(t)\right) dt = h \int_0^1 g(\tau) d\tau \tag{3.89}$$

où g est bien une fonction continue (i.e. $g \in C^0$).

Afin de calculer l'intégrale $\int_0^1 g(\tau) d\tau$, on approche la fonction $g(\tau)$ par un polynôme quadratique :

$$p(\tau) = A\tau^2 + B\tau + C \tag{3.90}$$

où les trois coefficients réels A, B et C seront calculés de sorte que la parabole associée au polynôme $p(\tau)$ passe par les points $(-2, g(-2))$, $(-1, g(-1))$ et $(0, g(0))$ (voir la Figure 3.2).

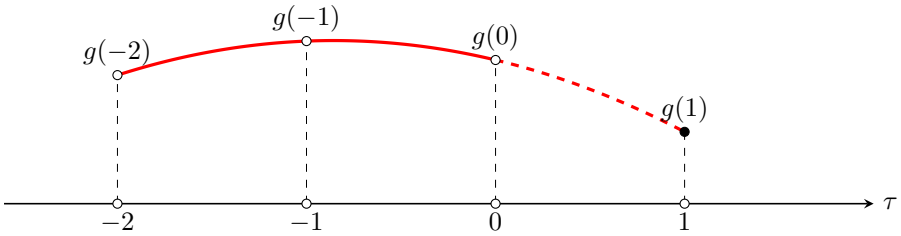

Figure 3.2: AB3 - Interpolation quadratique

Il suffit de résoudre le système :

$$\begin{cases} 4A - 2B + C &= g(-2) \\ A - B + C &= g(-1) \\ C &= g(0) \end{cases} \tag{3.91}$$

ce qui donne :

$$\begin{cases} A = \dfrac{g(-2) - 2g(-1) + g(0)}{2} \\ B = \dfrac{g(-2) - 4g(-1) + 3g(0)}{2} \\ C = g(0) \end{cases} \tag{3.92}$$

On peut maintenant intégrer :

$$\begin{aligned} \int_0^1 g(\tau)d\tau &\approx \int_0^1 p(\tau)d\tau \\ &\overset{(3.90)}{=} \int_0^1 (A\tau^2 + B\tau + C)d\tau \\ &= \frac{A}{3} + \frac{B}{2} + C \end{aligned}$$

et utiliser les expressions (3.92) pour les coefficients A, B et C :

$$\int_0^1 g(\tau)d\tau \approx \frac{1}{12}\left(23g(0) - 16g(-1) + 5g(-2)\right) \tag{3.93}$$

En remplaçant cette dernière relation (3.93) dans l'égalité (3.89) et en tenant compte

que :

$$\left\{\begin{array}{rcl} g(0) & = & f(t_n, y(t_n)) \\ g(-1) & = & f(t_{n-1}, y(t_{n-1})) \\ g(-2) & = & f(t_{n-2}, y(t_{n-2})) \end{array}\right. \tag{3.94}$$

on obtient :

$$\int_{t_n}^{t_{n+1}} f(t)\, dt \approx \frac{h}{12}\left(23f(t_n, y(t_n)) - 16f(t_{n-1}, y(t_{n-1})) + 5f(t_{n-2}, y(t_{n-2}))\right) \tag{3.95}$$

Grâce à cette dernière expression (3.95), la relation (3.86) devient :

$$y_{n+1} \approx y_n + \frac{h}{12}\left(23f(t_n, y(t_n)) - 16f(t_{n-1}, y(t_{n-1})) + 5f(t_{n-2}, y(t_{n-2}))\right) \tag{3.96}$$

En utilisant les valeurs approchées numériquement $u_n \approx y_n$, $n = 0, 1, ..., N$, on trouve finalement le schéma d'Adams-Bashforth à 3 pas :

$$u_{n+1} = u_n + \frac{h}{12}(23f(t_n, u_n) - 16f(t_{n-1}, u_{n-1}) + 5f(t_{n-2}, u_{n-2})) \tag{3.97}$$

3.6.3 Méthodes d'Adams-Moulton

Les méthodes d'Adams-Moulton sont des méthodes **multipas implicites**.

Vues comme des cas particuliers, la méthode implicite d'**Euler rétrograde** peut être considérée comme une méthode d'Adams-Moulton "**à zéro pas**" (et elle est **convergente d'ordre 1)**, tandis que la méthode implicite de **Crank-Nicolson** est la méthode d'Adams-Moulton **à un pas** (et sa **convergence** est **d'ordre 2)**.

Les méthodes d'Adams-Moulton **à 2** et **3 pas** sont **convergentes d'ordre 3** et, respectivement, **4**. Ces méthodes sont appelées couramment AM3 et, respectivement, AM4 et elles correspondent aux schémas suivants :

$$AM3 \qquad u_{n+1} = u_n + \frac{h}{12}\left(5f(t_{n+1}, u_{n+1}) + 8f(t_n, u_n) - f(t_{n-1}, u_{n-1})\right) \tag{3.98}$$

$$AM4 \qquad u_{n+1} = u_n + \frac{h}{24}(9f(t_{n+1}, u_{n+1}) + 19f(t_n, u_n) - 5f(t_{n-1}, u_{n-1}) + f(t_{n-2}, u_{n-2})) \tag{3.99}$$

où n varie de 1 et, respectivement, 2 à $N-1$.

Afin d'amorcer une méthode d'Adams-Moulton AM3 ou AM4, il faut calculer d'abord (en fonction de $u_0 = y_0$ connue) les valeurs approchées u_1 ou u_1 et u_2 par une (autre) méthode adéquate (le plus souvent, une méthode à un pas qui a, au moins, le même ordre de convergence).

A titre d'exemple, on présente ci-dessous une façon d'obtenir la formule (3.98), c'est-à-dire le schéma numérique correspondant à la méthode d'Adams-Moulton à 3 pas.

Comme pour la méthode d'Adams-Bashforth, on considère un sous-intervalle quelconque $[t_n, t_{n+1}]$, $n = 0, 1, ..., N-1$, d'une partition régulière σ_{reg} de pas h et on reprend la relation (3.10) :

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} f\left(t, y\left(t\right)\right) dt \tag{3.100}$$

Afin de calculer numériquement cette dernière intégrale définie (à partir de $n = 2$), on approche l'intégrant $f\left(t, y\left(t\right)\right)$ par un polynôme d'interpolation de degré 2 qui passe par les deux derniers points supposés "connus", c'est-à-dire le point "courant" $(t_n, f(t_n, y(t_n)))$ et le point précédent $(t_{n-1}, f(t_{n-1}, y(t_{n-1})))$, ainsi que par le point "à calculer" $(t_{n+1}, f(t_{n+1}, y(t_{n+1})))$. Par conséquent, cette démarche conduira bien à un schéma implicite.

A nouveau, sans perdre de généralité, on remplace la partition σ_{reg} par une partition régulière "canonique" de pas (unitaire) 1 en faisant le changement de variable :

$$t = t_n + h\tau \tag{3.101}$$

Il convient de remarquer que la nouvelle variables (temporelle) τ fait correspondre aux moments successifs $t = t_{n-1}$, $t = t_n$ et $t = t_{n+1}$ les "nouveaux" moments $\tau = -1$, $\tau = 0$ et, respectivement, $\tau = 1$ (voir la Figure 3.3).

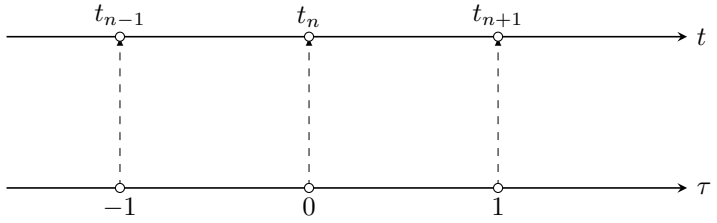

Figure 3.3: AM3 - Changement de variable (temporelle)

Comme dans le paragraphe précédent, on note :

$$f(t_n + h\tau, y(t_n + h\tau)) = g(\tau) \tag{3.102}$$

et on obtient l'égalité :

$$\int_{t_n}^{t_{n+1}} f\left(t, y(t)\right) dt = h \int_0^1 g(\tau) d\tau \tag{3.103}$$

où g est bien une fonction continue (i.e. $g \in C^0$).

Afin de calculer l'intégrale $\int_0^1 g(\tau)d\tau$, on approche la fonction $g(\tau)$ par un polynôme quadratique :

$$p(\tau) = A\tau^2 + B\tau + C \tag{3.104}$$

mais les trois coefficients réels A, B et C seront calculés, cette fois, de sorte que la parabole associée au polynôme $p(\tau)$ passe par les points $(-1, g(-1))$, $(0, g(0))$ et $(1, g(1))$ (voir la Figure 3.4).

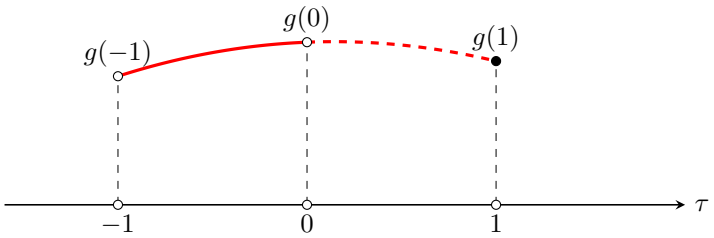

Figure 3.4: AM3 - Interpolation quadratique

Il suffit de résoudre le système :

$$\begin{cases} A - B + C &= g(-1) \\ C &= g(0) \\ A + B + C &= g(1) \end{cases} \tag{3.105}$$

ce qui donne :

$$\begin{cases} A = \dfrac{g(-1) - 2g(0) + g(1)}{2} \\ B = \dfrac{-g(-1) + g(1)}{2} \\ C = g(0) \end{cases} \tag{3.106}$$

On peut maintenant intégrer :

$$\begin{aligned} \int_0^1 g(\tau)d\tau &\approx \int_0^1 p(\tau)d\tau \\ &\overset{(3.104)}{=} \int_0^1 (A\tau^2 + B\tau + C)d\tau \\ &= \frac{A}{3} + \frac{B}{2} + C \end{aligned}$$

et utiliser les expressions (3.106) pour les coefficients A, B et C:

$$\int_0^1 g(\tau)d\tau \approx \frac{1}{12}\left(5g(1)+8g(0)-g(-1)\right) \tag{3.107}$$

En remplaçant cette dernière relation (3.107) dans l'égalité (3.103) et en tenant compte que :

$$\left\{\begin{array}{rcl} g(1) & = & f(t_{n+1},y(t_{n+1})) \\ g(0) & = & f(t_n,y(t_n)) \\ g(-1) & = & f(t_{n-1},y(t_{n-1})) \end{array}\right. \tag{3.108}$$

on obtient :

$$\int_{t_n}^{t_{n+1}} f\left(t\right)dt \approx \frac{h}{12}\left(5f(t_{n+1},y(t_{n+1}))+8f(t_n,y(t_n))-f(t_{n-1},y(t_{n-1}))\right) \tag{3.109}$$

Grâce à cette dernière expression (3.109), la relation (3.100) devient :

$$y_{n+1} \approx y_n + \frac{h}{12}\left(5f(t_{n+1},y(t_{n+1}))+8f(t_n,y(t_n))-f(t_{n-1},y(t_{n-1}))\right) \tag{3.110}$$

En utilisant les valeurs approchées numériquement $u_n \approx y_n$, $n=0,1,...,N$, on trouve finalement le schéma d'Adams-Moulton à 3 pas :

$$u_{n+1} = u_n + \frac{h}{12}\left(5f(t_{n+1},u_{n+1})+8f(t_n,u_n)-f(t_{n-1},u_{n-1})\right) \tag{3.111}$$

3.6.4 Méthodes d'Adams-Bashforth-Moulton

Les méthodes d'Adams-Bashforth-Moulton sont des méthodes **multipas explicites** de type **prédicteur-correcteur**.

Afin d'éviter des schémas implicites comme ceux d'Adams-Moulton :

- on utilise d'abord, dans la phase de **prédiction**, un schéma d'Adams-Bashforth, par exemple AB4, qui donne une première valeur approchée $\widetilde{u}_{n+1}$:

$$\begin{array}{rcl} \widetilde{u}_{n+1} & = & u_n + \dfrac{h}{24}(55f\left(t_n,u_n\right)-59f\left(t_{n-1},u_{n-1}\right) \\ & & +37f\left(t_{n-2},u_{n-2}\right)-9f\left(t_{n-3},u_{n-3}\right)) \end{array} \tag{3.112}$$

- on utilise ensuite, dans la phase de **correction**, un schéma d'Adams-Moulton, par exemple AM4, qui calcule une nouvelle valeur approchée u_{n+1}, plus précise :

$$u_{n+1} = u_n + \frac{h}{24}(9f\left(t_{n+1}, \widetilde{u}_{n+1}\right) + 19f\left(t_n, u_n\right) - 5f\left(t_{n-1}, u_{n-1}\right) + f\left(t_{n-2}, u_{n-2}\right)) \quad (3.113)$$

Dans l'exemple donné ci-dessus, on vient d'obtenir la méthode d'Adams-Bashforth-Moulton **à 4 pas** qui est **convergente d'ordre 4** et appelée ABM4.

Afin d'amorcer la méthode ABM4, il faut calculer d'abord (en fonction de $u_0 = y_0$ connue) les valeurs approchées u_1, u_2 et u_3 par une (autre) méthode adéquate (le plus souvent, une méthode à un pas qui a, au moins, le même ordre de convergence, par exemple la méthode RK4 classique).

D'une manière générale, la **précision** des méthodes multipas de type prédicteur-correcteur est similaire à la précision des méthodes à un pas ayant le même ordre de convergence. Cependant, le **nombre d'évaluations** de la fonction $f(t, y)$ pour une même itération est plus important dans le cas des méthodes à un pas. Par exemple, la méthode de Runge-Kutta classique est d'ordre 4 et nécessite 4 nouvelles évaluations de $f(t, y)$ pour chaque itération (afin de calculer les 4 coefficients K_1, K_2, K_3 et K_4), tandis que la méthode ABM4 est aussi d'ordre 4 mais ne nécessite que 2 nouvelles évaluations de $f(t, y)$ pour chaque itération. Par conséquent, le "**coût numérique**" des méthodes multipas de type prédicteur-correcteur est réduit, donc plus avantageux par rapport aux méthodes à un pas de même ordre.