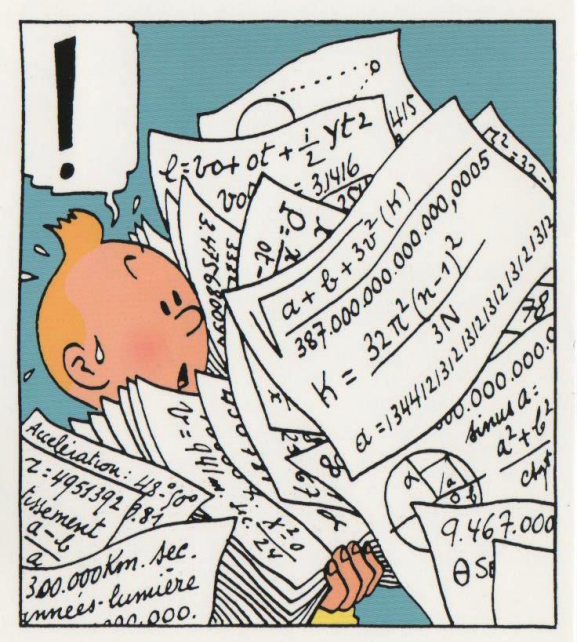
Université Claude Bernard, Lyon I
43, boulevard 11 novembre 1918
69622 Villeurbanne cedex, France

Licence Sciences, Technologies & Santé
Spécialité Mathématiques
L. Pujo-Menjouet
pujo@math.univ-lyon1.fr

Introduction à l'analyse numérique

Deuxième année de licence



Université Claude Bernard Lyon 1

Préambule

L'analyse numérique a commencé bien avant la conception des ordinateurs et leur utilisation quotidienne que nous connaissons aujourd'hui. Les premières méthodes ont été développées pour essayer de trouver des moyens rapides et efficaces de s'attaquer à des problèmes soit fastidieux à résoudre à cause de leur grande dimension (systèmes à plusieurs dizaines d'équations par exemple), soit parce qu'il n'existe pas solutions explicites connues même pour certaines équations assez simples en apparence.
Dès que les premiers ordinateurs sont apparus, ce domaine des mathématiques a pris son envol et continue encore à se développer de façon très soutenue.
Les applications extraordinairement nombreuses sont entrées dans notre vie quotidienne directement ou indirectement. Nous les utilisons désormais sans nous en rendre compte mais surtout en ignorant la plupart du temps toute la théorie, l'expertise, le développement des compétences et l'ingéniosité des chercheurs pour en arriver là. Nous pouvons téléphoner, communiquer par satellite, faire des recherches sur internet, regarder des films où plus rien n'est réel sur l'écran, améliorer la sécurité des voitures, des trains, des avions, connaître le temps qu'il fera une semaine à l'avance,...et ce n'est qu'une infime partie de ce que l'on peut faire.
Le but de ce cours et s'initier aux bases de l'analyse numérique en espérant qu'elles éveilleront de l'intérêt, de la curiosité et pourquoi pas une vocation.

FIGURE 1 – *Entre le Tintin dessiné à la main dans les années 6 par Hergé et celui mis à l'écran par Spielberg, un monde numérique les sépare. Un peu comme les premiers développements à la main des pionniers du numérique et les effets dernier cri des plus puissants des ordinateurs.*

Table des matières

Liste des figures

Chapitre 1

Interpolation polynomiale

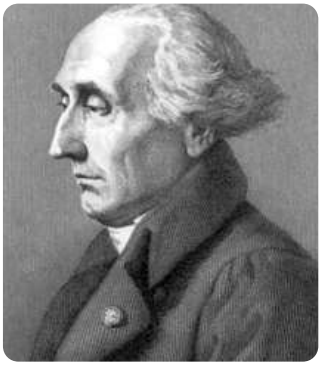
L'ordinateur a l'intelligence de celui qui s'en sert.

Anonyme

Sommaire

(a) Carl Runge(1856-1927), mathématicien et physicien allemand, montre quel'interpolation polynomiale de Lagrange peut diverger, même avec des fonctions présentant toutes les conditions de régularité, c'est ce qu'on appelle le phénomène de Runge.

(b) Charles Hermite (1822-1901), mathématicien français qui a beaucoup travaillé sur les polynômes (entre autres), on lui doit l'interpolation qui porte son nom.



(c) Joseph Louis, comte de Lagrange (1736-1813), mathématicien, on lui doit entre autres choses la notation f' pour désigner la dérivée d'une fonction.

FIGURE 1.1 – Quelques mathématiciens célèbres liés à l'étude de l'interpolation polynomiale.

1.1 Introduction

Dans ce chapitre, nous nous donnons une fonction f supposée continue, définie sur un intervalle de $\mathbb{R}$, la plupart du temps nous noterons $[a, b]$ cet intervalle. Nous cherchons à approcher, dans un sens à préciser, cette fonction par un polynôme.
Attention, il ne faudra pas confondre deux approches différentes. Suivant le problème, nous choisissons plutôt l'une ou l'autre de ces approches qui sont :

1. l'**interpolation** : nous cherchons un polynôme qui coïncide avec f en un certain nombre de points,
2. l'**approximation** (au sens des moindres carrés) : nous nous donnons une distance "entre fonctions" et nous cherchons un polynôme proche de f au point de vue de cette distance sans toutefois passer nécessairement par des points de f.

Pour être plus clair, dans le cas de l'interpolation polynomiale, on cherche un polynôme dont le graphe passe par tous les points donnés (ou connus) de f tandis que dans le cas de l'approximation, on essaie d'"épouser" le mieux possible la courbe de f par un polynôme.
Les méthodes d'interpolation et d'approximation sont utilisées par les logiciels du type Matlab (ou Scilab en version gratuite) pour tracer des graphes continus à partir de valeurs discrètes.
Dans ce cours nous ne nous intéresserons qu'à l'interpolation polynômiale. L'approximation au sens des moindres carrés se verra en 3ème année de licence dans le cours d'analyse

matricielle.
Il est possible de trouver des cours et des exercices dans de nombreux ouvrages disponibles à la bibliothèque. Ceux que je suggère pour ce chapitre sont ceux de Burden et Faires [2], Filbet [4], Quarteroni *et al*. [5] et Schatzman [6].
Avant de commencer, rappelons quelques résultats connus sur les polynômes.

1.2 Rappels sur les polynômes

Soit $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. Nous notons $\mathscr{P}(\mathbb{K})$l'ensemble des fonctions polynômes. Autrement dit, pour tout $P \in \mathscr{P}(\mathbb{K})$, il existe un entier $n \in \mathbb{N}$ et des coefficients a_i, $i = 0, ..., n$ dans $\mathbb{K}$ tels que

$$P(x) = \sum_{i=0}^{n} a_i x^i = a_0 + a_1 x^1 + a_2 x^2 + ... + a_n x^n,$$

pour tout $x \in \mathbb{K}$.

Définition 1 (Degré de polynôme - Monôme)

- Le **degré** d'un polynôme non nul est l'indice maximum d'un coefficient non nul.
Par convention, le **degré d'un polynôme nul** (dont tous les coefficients sont nuls) est $-\infty$.
- Un **monôme** est un polynôme de la forme $a_n x^n$, pour $n \in \mathbb{N}$, a_n et $x \in \mathbb{K}$.

Remarque

1. *Notons que l'ensemble $\mathscr{P}(\mathbb{K})$ est un $\mathbb{K}$-espace vectoriel. Nous ne le démontrerons pas.*
2. *Pour tout $n \in \mathbb{N}^*$, nous noterons $\mathscr{P}_n(\mathbb{K})$ le sous-espace vectoriel de $\mathscr{P}(\mathbb{K})$ des polynômes de degré inférieur ou égal à n dont $\{1, x, ..., x^n\}$ est une base appelée **base canonique**.*
3. *La dimension de $\mathscr{P}_n(\mathbb{K})$ est $n + 1$.*

Nous avons alors la proposition suivante.

Proposition 1 (Base de $\mathscr{P}(\mathbb{K})$)

Toute famille de polynômes de degré deux à deux distincts est libre.
Par conséquent, si $(P_i)_{0 \leq i \leq n}$ est une suite de polynômes telle que le degré de P_i soit égal à i, alors $\{P_0, P_1, ..., P_n\}$ est une base de $\mathscr{P}_n(\mathbb{K})$.

Rappelons quelques résultats de première année de licence.

Définition 2 (Divisibilité)

Soient P et $Q \in \mathscr{P}(\mathbb{K})$ deux polynômes. Nous disons que P divise Q (ou que Q est un multiple de P) s'il existe un polynôme $R \in \mathscr{P}(\mathbb{K})$ tel que $Q(x) = P(x)R(x)$ pour tout $x \in \mathbb{K}$.

Définition 3 (Racine)

Soient $P \in \mathscr{P}(\mathbb{K})$ et $x_0 \in \mathbb{K}$.

1. x_0 est une **racine** (on dit aussi zéro) de P lorsque $P(x_0) = 0$,
2. x_0 est une **racine d'ordre** k de P si $(x-x_0)^k$ est un diviseur de P tandis que $(x-x_0)^{k+1}$ ne l'est pas,
3. un polynôme de degré n est **scindé** lorsqu'il est le produit de n polynômes de degré 1. Autrement dit, il admet n racines distinctes,
4. un polynôme est dit unitaire (ou normalisé) quand son coefficient dominant est 1.

Proposition 2 (Polynômes et racines)

Soit $P \in \mathscr{P}(\mathbb{K})$.

1. Si le polynôme P est non nul, alors il possède un nombre fini de racines distinctes inférieur ou égal à son degré.
2. Si le polynôme P s'annule en un nombre infini de points, alors il est nul.
3. Si le polynôme P est de degré n et s'annule en $n+1$ points alors il est nul.
4. Soient P et Q deux polynômes , alors $d^\circ(P+Q) \leq \max(d^\circ P, d^\circ Q)$.
5. Soient P et Q deux polynômes , alors $d^\circ(PQ) = d^\circ P + d^\circ Q$.

Définition 4 (Dérivée)

Le polynôme dérivé du polynôme $P(x) = \sum_{k=0}^{n} a_k x^k$ est le polynôme $P'(x) = \sum_{k=0}^{n} k a_k x^{k-1}$

Proposition 3 (Formule de Taylor)

Soit $P \in \mathscr{P}(\mathbb{K})$, qui s'écrit $P(x) = \sum_{k=0}^{n} a_k x^k$ alors,

1. $P(x) = \sum_{k=0}^{n} \frac{P^{(k)}(x_0)}{k!}(x - x_0)^k$, où $P^{(k)}$ désigne la dérivée k-ième de P, et $x_0 \in \mathbb{K}$.
   De cette façon un polynôme de degré n est entièrement déterminé par la connaissance de $P(x_0)$, ...,$P^{(n)}(x_0)$.
2. x_0 est une racine d'ordre k de P si $P(x_0) = ... = P^{(k-1)}(x_0) = 0$ mais $P^{(k)}(x_0) \neq 0$.

Intéressons nous maintenant à l'interpolation polynomiale à proprement parlé. Il existe plusieurs méthodes. Nous allons décrire ici les plus "incontournables" ou "classiques" selon les points de vue. Rappelons cependant qu'il y a un point commun pour chacune des méthodes : il faut suffisamment d'informations pour avoir une "bonne" interpolation" : à la fois sur la régularité et sur le nombre de points.
Commençons par la plus célèbre, l'interpolation de Lagrange.

1.3 Interpolation de Lagrange

Soit $f : [a, b] \to \mathbb{R}$ continue. Nous nous donnons $n + 1$ points distincts dans $[a, b]$,

$$a \leq x_0 < x_1 < ... < x_n \leq b,$$

appelés nœuds.
Pour plus de simplicité, quand ce n'est pas précisé nous noterons $\mathscr{P}_n$ au lieu de $\mathscr{P}_n(\mathbb{R})$ l'espace des polynômes de degré inférieur ou égal à n à coefficients réels. Rappelons que cet espace est de dimension $n + 1$.

1.3.1 Avec les polynômes de Lagrange

Théorème 1 (POLYNÔMES DE LAGRANGE)

Pour tout choix de nœuds $x_0, x_1, ..., x_n$ dans $[a, b]$, il existe un unique polynôme P_n de degré inférieur ou égal à n qui coïncide avec f aux points $x_0, x_1, ..., x_n$ (*i. e.* $P(x_j) = f(x_j)$ pour tout $j = 0, ..., n$).
Ce polynôme s'écrit

$$P_n(x) = \sum_{j=0}^{n} f(x_j) L_j(x), \tag{1.1}$$

où

$$L_j(x) = \prod_{\substack{k=0,\\ k\neq j}}^{n} \frac{x - x_k}{x_j - x_k}.$$

pour tout $j = 0, ..., n$.

Remarque

1. *Les polynômes de Lagrange sont tels que*

$$L_j(x_k) = \delta_{jk} = \begin{cases} 1, & si \ \ j = k, \\ 0, & sinon. \end{cases}$$

   on rappelle que δ_{jk} est appelé symbole de Kronecker.

2. *L'écriture (1.1) n'est pas utilisée en pratique. On ne peut pas calculer facilement le polynôme d'interpolation de f aux point $x_0, x_1, ..., x_n$ à partir du polynôme d'interpolation aux nœuds $x_0, x_1, ..., x_n$ étant donné que chacun des L_j dépend de tous les nœuds.*

Il existe une autre forme, plus pratique à utiliser : la forme de Newton.

1.3.2 Forme de Newton

Construisons la méthode de Newton pas après pas. En commençant avec 1 nœud, puis 2 nœuds, puis n nœuds

1. ***Avec un seul nœud x_0*** :
   P_0 de degré 0 et $P_0(x_0) = f(x_0)$, donc

$$P_0(x) = f(x_0).$$

2. ***Avec 2 nœuds x_0 et x_1*** :
   P_1 de degré inférieur ou égal à 1, nous avons

$$P_1(x_0) = f(x_0), P_1(x_1) = f(x_1).$$

   Nous écrivons alors

$$P_1(x) = P_0(x) + R_1(x), \text{ où } \deg(R_1) \leq 1.$$

Or $P_1(x_0) = f(x_0) = P_0(x_0)$ donc $R_1(x_0) = 0$.
D'où $R_1(x) = a_1(x - x_0)$.
De plus $P_1(x_1) = f(x_1) = f(x_0) + a_1(x_1 - x_0)$, et donc

$$a_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

Nous notons alors $f[x_0, x_1] := \dfrac{f(x_1) - f(x_0)}{x_1 - x_0}$ le coefficient a_1. Et donc,

$$P_1(x) = f(x_0) + f[x_0, x_1](x - x_0).$$

3. ***Avec 3 nœuds x_0, x_1 et x_2*** :
   P_2 de degré inférieur ou égal à 2, nous avons

$$P_2(x_0) = f(x_0),\ P_2(x_1) = f(x_1) \text{ et } P_2(x_2) = f(x_2).$$

Nous écrivons alors

$$P_2(x) = P_1(x) + R_2(x), \text{ où } \deg(R_2) \leq 2.$$

Or

$$\begin{array}{lllllllll} P_2(x_0) & = & f(x_0) & = & P_1(x_0) & \text{donc} & R_1(x_0) & = & 0, \ et \\ P_2(x_1) & = & f(x_1) & = & P_1(x_1) & \text{donc} & R_2(x_1) & = & 0. \end{array}$$

D'où $R_2(x) = a_2(x - x_0)(x - x_1)$.
De plus $P_2(x_2) = f(x_2) = f(x_0) + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1)$.
En soustrayant l'expression de $P_2(x_2)$ à celle de $P_2(x_1)$, nous obtenons

$$a_2 = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0},$$

où $f[x_1, x_2] = \dfrac{f(x_2) - f(x_1)}{x_2 - x_1}$ et $f[x_0, x_1]$ est défini précédemment comme le coefficient a_1.
Nous notons alors $f[x_0, x_1, x_2] := \dfrac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$ le coefficient a_2. Et donc,

$$P_2(x) = f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1).$$

De façon générale nous pouvons construire le cas à n nœuds. Pour le calcul des coefficients : nous introduisons la notation des différences divisées

$$f[x_0] = f(x_0) \text{ pour tout réel } x_0,$$

et pour tout $n \in \mathbb{N}^*$, pour tous nœuds réel x_0, x_1,...,x_n,

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

Théorème 2 (MÉTHODE DE NEWTON)

Pour tout $n \in \mathbb{N}^*$, pour tous nœuds $x_0, x_1, ..., x_n$ dans $[a, b]$, il existe un unique polynôme P_n de degré inférieur ou égal à n qui coïncide avec f aux points $x_0, x_1, ..., x_n$ (*i.e.* $P(x_j) = f(x_j)$ pour tout $j = 1, ..., n$).
Ce polynôme s'écrit

$$P_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + ... + f[x_0, ..., x_n] \prod_{k=0}^{n-1} (x - x_k).$$

Pour construire les coefficients de Newton nous procédons de la façon suivante :

$$\begin{aligned}
a_0 &= f(x_0) \\
a_1 &= f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \\
a_2 &= f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = \frac{\dfrac{f(x_2) - f(x_1)}{x_2 - x_1} - \dfrac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0} \\
&\vdots \\
a_n &= f[x_0, x_1, \cdots, x_n] = \frac{f[x_1, x_2, \cdots, x_n] - f[x_0, x_1, x_2, \cdots, x_{n-1}]}{x_n - x_0}
\end{aligned}$$

Remarque
Cette fois-ci, la forme de Newton est bien adaptée à un algorithme, il suffit de calculer les coefficients à l'aide de la définition des différences divisées par récurrence.

1.3.3 Erreur d'interpolation

Théorème 3 (ERREUR D'INTERPOLATION)

Soit $f : [a, b] \to \mathbb{R}$ de classe $\mathscr{C}^{n+1}$. Notons P_n son polynôme d'interpolation aux nœuds $x_0, x_1, ..., x_n$ dans $[a, b]$. Alors pour tout $x \in [a, b]$

$$|f(x) - P_n(x)| \leq \frac{M_{n+1}}{(n+1)!} |\pi_n(x)|,$$

où

$$\pi_n(x) = \prod_{k=0}^{n} (x - x_k) \text{ et } M_{n+1} = \sup_{x \in [a,b]} |f^{(n+1)}(x)|,$$

Remarque

1. *Notons que M_{n+1} est fini car $f^{(n+1)}$ est continue sur le compact $[a, b]$.*
2. *D'autre part, ce résultat repose sur le lemme suivant*

Lemme 1 (LIEN DIFFÉRENCES DIVISÉES ET DÉRIVÉES)

Si f est de classe $\mathscr{C}^p$ alors pour tous nœuds $x_0, x_1, ..., x_p$ dans $[a, b]$ il existe $\xi \in [a, b]$ tel que

$$f[x_0, ..., x_p] = \frac{f^{(p)}(\xi)}{p!}.$$

Remarque

Ce résultat n'entraîne pas la convergence uniforme de P_n vers f quand $n \to +\infty$.
En effet, le polynôme π_n peut beaucoup osciller quand on augmente son degré et la norme

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

peut être très grande. L'erreur peut donc dépendre à la fois de la taille du segment $[a, b]$ et de la répartition des nœuds sur cet intervalle. Un exemple classique lorsque la répartition est équidistante est appelé phénomène de Runge.
Pour y remédier, nous pouvons chercher une répartition des points $x_0, ..., x_n$ qui rendent la norme $\|\pi_n\|_\infty$ minimale. C'est le cas pour les polynômes de Tchebychev (voir en TD). Mais même ça n'évite pas nécessairement le phénomène d'oscillations.
En pratique, un autre moyen est d'utiliser l'interpolation composée.

1.3.4 Interpolation composée

Afin de se prémunir des instabilités numériques comme le phénomène de Runge qui peuvent apparaître sur des grands intervalles ou des degrés de polynômes trop grands, nous utilisons une interpolation polynomiale avec des polynômes de degré peu élevé (pour se préserver des instabilités) sur des sous intervalles de $[a, b]$ (et donc des intervalles de petites tailles).
Nous considérons les nœuds $a = a_0 < a_1 < ... < a_n = b$ qui subdivisent $[a, b]$.
D'autre part, dans chaque intervalle $[a_j, a_{j+1}]$ nous choisissons $m + 1$ nœuds $x_0^1, ..., x_m^j$ et nous construisons un polynôme de Lagrange sur chacun de ces sous-intervalles aux nœuds choisis. Il faut quand même que l'interpolation soit continue aux extrémités de chacun des sous-intervalles. On s'assure de ça en prenant $x_0^i = a_i$ et $x_m^i = a_{i+1}$. Nous avons alors le résultat suivant.

Théorème 4 (INTERPOLATION COMPOSÉE)

Soit $f : [a, b] \to \mathbb{R}$ continue.

1. (Existence) Il existe,
2. (Unicité) une unique fonction $f_{m,n} : [a, b] \to \mathbb{R}$
3. (Continuité) continue sur $[a, b]$ telle que

$$f_{m,n}|_{[a_j, a_{j+1}]} \text{ polynôme de degré inférieur à } m,$$

4. D'autre part,

$$f_{m,n}(x_k) = f(x_k) \text{ pour tout } 0 \leq k \leq m,\ 0 \leq j \leq n.$$

5. De plus, si f est de classe $\mathscr{C}^{m+1}$ alors

$$\|f - f_{m,n}\|_\infty \leq \frac{h^{m+1}}{(m+1)!} \|f^{(m+1)}\|_\infty,$$

où $h = \max_{0 \leq j \leq n-1} |a_{j+1} - a_j|$.

Remarque

Ici à m fixé, quand $h \to 0$ nous avons bien la convergence uniforme de $f_{m,n}$ vers f sur $[a, b]$.

Remarque

Dans la pratique, on constate qu'une très bonne approximation n'est pas forcément effectuée avec polynôme de très haut degré, mais plutôt avec une successions de plusieurs polynômes de degré peu élevés répartis sur des sous intervalles beaucoup plus petite que $[a, b]$.

Cependant, nous avons besoin quelques fois non seulement que l'interpolation se fasse non seulement pour les fonctions f aux nœuds, mais également pour la dérivée f' en ces mêmes nœuds. C'est ce qu'on appelle un mélange de polynôme interpolateur (dont la valeur en un nœud donné est celle de f en ce nœud) et osculateur (dont la valeur en un nœud donné est celle de f en ce nœud). Noter qu'osculateur vient du latin *osculare* qui signifie "embrasser". Autrement dit, on souhaite vraiment qu'en ces nœuds, le polynôme "embrasse" le graphe de f. Cette méthode est une alternative permettant d'éviter les phénomènes d'interpolation du type Runge.

1.4 Interpolation d'Hermite

Nous construisons un polynôme d'interpolation en utilisant les valeurs de f et de sa dérivée. Nous supposerons f de classe $\mathscr{C}^1$ sur $[a, b]$.
Nous cherchons maintenant un polynôme P tel que pour tout $0 \leq j \leq n$

$$\left\{ \begin{array}{rcl} f(x_j) & = & P(x_j), \\ f'(x_j) & = & P'(x_j). \end{array} \right.$$

Le polynôme P est de degré $2n+1$. Pour ça, nous utilisons les polynômes H_k et $\hat{H}_k$ tels que

$$H_k(x_j) = \delta_{kj}, \quad H'_k(x_j) = 0, \, \hat{H}_k(x_j) = 0 \text{ et } \quad \hat{H}'_k(x_j) = \delta_{jk}$$

pour tous $0 \leq k, j \leq n$. Avec la base de Lagrange , nous pouvons écrire

$$\left\{ \begin{array}{rcl} H_k(x) & = & L_k(x)^2 \left(1 - 2L'_k(x_k)(x - x_k)\right), \\ \hat{H}_k(x) & = & L_k(x)^2 (x - x_k). \end{array} \right.$$

Théorème 5 (INTERPOLATION D'HERMITE)

Soient x_0, ..., x_n des nœuds de $[a, b]$, $f : [a, b] \to \mathbb{R}$de classe $\mathscr{C}^1$. Il existe un unique polynôme P de degré inférieur ou égal à $2m + 1$ tel que pour tout $0 \leq j \leq n$,

$$P(x_j) = f(x_j), \text{ et } \quad P'(x_j) = f'(x_j).$$

Il s'écrit,

$$P(x) = \sum_{k=0}^{n} \Big(f(x_k) H_k(x) + f'(x_k) \hat{H}_k(x) \Big).$$

De plus, si f est de classe $\mathscr{C}^{2n+1}$ alors

$$|f(x) - P(x)| \leq \frac{\|f^{(2n+2)}\|_\infty}{(2n+2)!} |\pi_n(x)|^2, \text{ pour tout } x \in [a, b].$$

Remarque *Comme pour l'interpolation de Lagrange, cette forme n'est pas adaptée à un algorithme. Quand on rajoute un nœud il faut tout recalculer. Il existe aussi pour les polynômes d'interpolation d'Hermite une écriture à l'aide des différences divisées que nous n'aborderons pas ici.*

1.5 Splines

Nous avons vu qu'il était possible d'éviter les instabilités en divisant l'intervalle $[a, b]$ en sous intervalles et en utilisant les interpolations composées qui sont des interpolations par morceaux avec des polynômes de degrés assez bas. Ceci permet ainsi q'obtenir des résultats corrects mais avec peu de régularité sur les extrémités de chaque morceaux. On ne demandait en effet que de la continuité pour relier les différents polynômes entre eux.
L'idée ici est de combiner cette méthode avec celle utilisée pour les polynômes d'Hermite afin de régulariser à souhait chacune des liaisons entre les morceaux de polynômes de faible degrés de l'interpolation composée.
Cette méthode est la méthode d'interpolation par les spline (cerce en français). Le mot spline (prononcer "splaïne" pour ne pas confondre avec le spleen (ce qui n'a rien à voir), même si

communément en France, le mot est prononcé “spleen” justement). Ce mot spline désignait à l’origine une latte de métal ou de bois très fine. C’est devenu en 1895 un règle flexible pour dessiner non pas des lignes droites mais des courbes. Ces splines ont été utilisée pour la construction d’avions et de bateaux.
On utilise désormais les splines “numériques” pour le dessin industriel de tout ce qui est courbe (les avions, les bateaux, les voitures,...). Ce n’est donc pas un hasard si les personnes qui ont créé l’interpolation par les splines viennent du design industriel dans les années 50-60 comme de Casteljau chez Citroën, Pierre Bézier chez Renault, et Birkhoff, Garabedian, de Boor chez General Motors.

Définition 5 (SPLINE)

Soient x_0, ..., x_n, $n+1$ nœuds de $[a,b]$, avec

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

La fonction $sk : [a,b] \to \mathbb{R}$ est une spline de degré k relative aux nœuds x_i si

$$s_{k,i} = s_k \mid_{[x_i,x_{i+1}]} \in \mathscr{P}_k \text{ , pour } i = 0, ..., n \text{ et } s_k \in \mathscr{C}^{k-1}[a,b].$$

tout polynôme de degré k sur $[a,b]$ est donc une spline, mais en pratique on considère

1. les splines constituées de n polynômes différents $s_{k,i}$ sur chaque sous intervalle $[x_i, x_{i+1}]$,
2. et il peut y avoir des discontinuités de la dérivée k-ième aux nœuds internes.

1.5.1 Calcul d’une spline

Le i-ième ($i = 0, ..., n-1$) polynôme $s_{k,i} = s_{k|[x_i,x_{i+1}]}$ composant la spline est d’ordre k et peut s’écrire sous la forme

$$s_{k,i} = \sum_{h=0}^{k} a_{hi}(x - x_i)^h.$$

Calculer une spline équivaut donc à déterminer $(k+1)n$ coefficients a_{hi} avec les conditions suivantes :

1. il faut $k(n-1)$ conditions pour les dérivées

$$s^{(m)}_{k,i-1}(x_i) = s^{(m)}_{k,i}(x_i),$$

   c’est à dire que la dérivée m-ième du polynôme à gauche en x_i est la même que celle du polynôme à droite en x_i pour $i = 0, ..., n-1$ et $m = 0, ..., k-1$.

2. Il y a $n+1$ conditions d’interpolations aux points x_i, c’est à dire

$$s_{k,i}(x_i) = f(x_i) \text{ pour } i = 0, ..., n.$$

3. $k-1$ interpolations additionnelles qui sont données par trois options :

   (a) les splines "périodiques" : c'est à dire $s_k^{(m)}(a) = s_k^{(m)}(b)$ pour $m = 1, ..., k-1$,

   (b) les splines "serrées" (en anglais "clamped") : c'est à dire $s_k^{(m)}(a) = f(a)^{(m)}$ et $s_k^{(m)}(b) = f(b)^{(m)}$ pour $m = 1, ..., k-1$,

   (c) les splines "naturelles" : c'est à dire que pour $k = 2l-1$, avec $l \geq 2$, on a $2(l-1) = k-1$ conditions additionnelles qui sont :

$$s_k^{(l+j)}(a) = 0, \text{ et } s_k^{(l+j)}(b) = 0 \text{ pour } j = 0, 1, ..., l-2.$$

   Exemple *Si $k = 3$, alors $l = 2$ et on a $s_k^{(2)}(a) = 0$ et $s_k^{(2)}(b) = 0$ (2 conditions).*

1.5.2 Splines cubiques

Les splines les plus couramment utilisées sont les splines cubiques (de degré $k = 3$) pour deux raisons principales :

1. ce sont les splines de moindre degré qui permettent une approximation $\mathscr{C}^2$ (autrement dit les dérivées premières et secondes sont continues),
2. elles ont de bonnes propriétés de régularité.

Le calcul d'une spline cubique se fait de la façon suivante : il faut déterminer $(k+1)n = 4n$ coefficients pour déterminer le polynôme s_3 avec les conditions :

1. d'interpolation : $s_3(x_i^-) = f(x_i) = s_3(x_i^+)$ pour $i = 1, ..., n-1$ (ce qui fait $2(n-1)$ conditions),
2. $s_3(x_0) = f(a)$ (ce qui fait une condition de plus),
3. $s_3(x_n) = f(n)$ (ce qui fait une condition de plus),
4. $s_3'(x_i^-) = s_3'(x_i^+)$ pour $i = 1, ..., n-1$ (ce qui fait $(n-1)$ conditions),
5. $s_3''(x_i^-) = s_3''(x_i^-)$ pour $i = 1, ..., n-1$ (ce qui fait $(n-1)$ conditions).

Au total, ça fait $4n-2$ conditions. Les $k-1 = 2$ conditions restantes sont les conditions additionnelles imposées par l'utilisateur (soit les splines périodiques, soit les splines naturelles).
Le calcul d'une spline demande donc la résolution d'un système linéaire de $4n$ équations.

Théorème 6 (SPLINE CUBIQUE)

Soient x_0, ..., x_n, $n+1$ nœuds de $[a,b]$, avec

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

Soit $f : [a,b] \to \mathbb{R}$.

1. Alors il existe une unique spline cubique naturelle (avec la condition additionnelle naturelle $s_k^{(2)}(a) = 0$ et $s_k^{(2)}(b) = 0$) ou périodique (avec la condition périodique ($s_k^{(m)}(a) = s_k^{(m)}(b)$ pour $m = 1, ..., k-1$)) interpolant les $(n+1)$ nœuds.
2. Si de plus f est différentiable en a et b alors il existe une unique spline serrée ($s_k'(a) = f'(a)$ et $s_k'(b) = f'(b)$).

On a alors l'erreur d'interpolation suivante :

Théorème 7 (SPLINE CUBIQUE : ERREUR)

Soit $f \in \mathscr{C}([a,b])$ avec $\max\limits_{a\leq x\leq b} |f^{(4)}(x)| = M$. Si s_3 est l'unique spline cubique serrée interpolant f aux nœuds $a = x_0 < x_1 < ... < x_n = b$, alors

$$\max_{a\leq x\leq b} |f(x) - s_3(x)| \leq \frac{5M}{384} \max_{0\leq j\leq n-1} (x_{j+1} - x_j)^4.$$

Il est à noté que la condition additionnelle naturelle donne en général des résultats moins précis que les conditions serrées aux bords de l'intervalle $[a,b]$ sauf si la fonction f satisfait également $f''(a) = f''(b) = 0$. Une alternative à la condition naturelle qui ne requiert pas de condition particulière sur la dérivée de f est la condition "not-a-knot" (pas un nœeud). Mais par contre il faut que $s_3^{(3)}$ soit continue en x_1 et en x_{n-1}.

Remarque

Dans la pratique, on constate qu'une très bonne approximation n'est pas forcément effectuée avec polynôme de très haut degré, mais plutôt avec une successions de plusieurs polynômes de degré peu élevés répartis sur des sous intervalles beaucoup plus petite que $[a,b]$.

Les splines d'interpolation peuvent toutefois présenter trois inconvénients :

1. la spline peut également devenir oscillante si les dérivées de la fonction à interpoler deviennent trop grandes,
2. la spline dépend du choix du système de coordonnées, donc elle ne possède pas de propriété d'invariance géométrique,
3. ceci peut être génant si la spline est utilisée pour représenter graphiquement une courbe qui n'est pas une fonction mais une fonction paramétrée (une ellipse par exemple). Dans ce cas là, il faut utiliser la spline de manière paramétrique.

Chapitre 2

Calculs approchés d'intégrales

Il ne faut pas uniquement intégrer. il faut aussi désintégrer. C'est ça la vie. C'est ça la philosophie. C'est ça la science. C'est ça le progrès, la civilisation.

Eugène Ionesco, 1950

Sommaire

(a) Roger Cotes(1682 – 1716), mathématicien anglais, proche d'Isaac Newton. Leur collaboration aboutira entre autres à la découverte de la méthode de Newton-Cotes en analyse numérique. Celle-ci est une généralisation des méthodes des trapèzes et lSimpson pour le calcul des intégrales.

(b) Georg Friedrich Bernhard Riemann (1826– 1866) , mathématicien allemand, très prolifique, on lui doit entre autres des travaux sur les intégrales, dans la continuité de ceux de Cauchy, à l'origine de ce qu'on appelle désormais les intégrales de Riemann.

(c) Thomas Simpson (1710-1761), mathématicien anglais généralisa la méthode de Newton au calcul itératif des solutions d'une équation non linéaire, en utilisant les dérivées

FIGURE 2.1 – Mathématiciens célèbres liés à l'étude des intégrales et de leurs approximations.

Il est très difficile, voire souvent impossible de trouver les primitives d'une fonction. Et donc, de calculer l'intégrale d'une fonction sur un intervalle. Pourtant, dans de nombreuses applications en physique, en chimie, en biologie il est nécessaire d'estimer ces intégrales. Par exemple, en dynamique de population, l'intégrale entre l'âge $a = 0$ jusqu'à l'âge maximum a_{max} (qui pourrait être estimé à 130 ans), en un instant t d'une fonction f dépendant de a permet d'évaluer en un temps t, une population totale observée suivant tous les âges. C'est ce qui permet de voir évoluer par exemple la pyramide des âges au cours du temps.
Il est donc essentiel de trouver des moyens pour évaluer ces intégrales sans connaître explicitement les expressions des primitives.
Il est possible de trouver des cours et des exercices dans de nombreux ouvrages disponibles à la bibliothèque. Ceux que je suggère pour ce chapitre sont ceux de Burden et Faires [2], Filbet [4], Quarteroni *et al*. [5] et Schatzman [6].
L'objectif de ce chapitre est donc de donner des méthodes permettant d'approcher les intégrales. C'est ce qu'on appelle en mathématiques, la quadrature. A l'origine, la quadrature d'une surface est la recherche d'un carré ayant même aire que la surface en question. C'est devenu par extension le calcul d'une intégrale.
L'une des méthodes les plus connues, la plus intuitive qui permet de construire les intégrales de Riemann comme vu en première année de licence est la méthode des rectangles.

2.1 Méthode des rectangles

Considérons une fonction $f : [a, b] \to \mathbb{R}$ une fonction continue. On découpe l'intervalle $[a, b]$ en n sous-intervalles $[x_k, x_{k+1}]$ réguliers ($k = 0, ..., n-1$), avec $x_0 = a$ et $x_n = b$, $n \in \mathbb{N}$.
Par construction, la longueur h d'un sous-intervalle est

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

Et donc, chacun des x_k se calcul en fonction de h par

$$x_k = a + k\frac{b-a}{n} = a + kh, \text{ pour } k = 0, ...n.$$

Sur chacun des sous-intervalles , on approche f par une fonction constante égale à la valeur de f en un point $[x_k, x_{k+1}]$ pour $k = 0, .., n$.

1. Si c'est le point x_k on parle de méthode des rectangles à gauche.
2. Si c'est le point x_{k+1} on parle de méthode des rectangles à droite.
3. Si c'est le point $\frac{x_k + x_{k+1}}{2}$, on parle de méthode du point milieu.

Supposons que ce soit un point quelconque $\xi_k \in [x_k, x_{k+1}]$. La somme de Riemann $I_n(f)$ correspondante est donnée par

$$I_n(f) = \frac{b-a}{n}\sum_{k=0}^{n-1} f(\xi_k).$$

C'est la somme des aires des rectangles de hauteur $f(\xi_k)$ et largeur $h = \frac{b-a}{n}$.

On montrerait que $\lim_{n\to+\infty} I_n(f) = \int_a^b f(x)dx$.

Remarque
On ne considèrera ici que les fonctions continues. Des cas plus généraux seront abordés dans le cours d'intégration de troisième année.

De façon générale, qu'appelle-t-on une formule de quadrature ?
Une formule de quadrature est une formule du type

$$\int_a^b f(x)dx \simeq \sum_{i=0}^{n} \omega_i f(x_i),$$

où $x_0, ..., x_n$ sont des point différents dans $[a, b]$, $n \in \mathbb{N}$ et $\omega_0, ..., \omega_n$ sont des réels appelés poids, qui sont à déterminer pour avoir une valeur approchée de l'intégrale de f entre a et b. Le but est donc de trouver les meilleurs poids ω_i pour avoir la meilleure approximation de l'intégrale.

2.2 Méthodes construites à partir des polynômes d'interpolation

Une des méthodes les plus simples consiste à utiliser les polynômes d'interpolation (voir chapitre 1).

2.2.1 Formules simples

L'idée est la suivante : on sait calculer explicitement les intégrales de polynômes puisque l'on connait leurs primitives.
Il ne nous reste plus qu'à trouver des interpolations polynômiales satisfaisante pour le problème que l'on cherche à résoudre.
Soit $f : [a, b] \to \mathbb{R}$. Soit $n \in \mathbb{N}$. On note P_n le polynôme d'interpolation de Lagrange de f aux nœuds $x_0 < ... < x_n$ dans $[a, b]$.
On approche alors l'intégrale de f de la façon suivante

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

On sait d'après le chapitre 1 que

$$P_n(x) = \sum_{j=0}^{n} f(x_j)L_j(x).$$

Par conséquent

$$\int_a^b P_n(x)dx = \sum_{j=0}^{n} \left(\int_a^b L_j(x)dx \right) f(x_j).$$

Autrement dit

$$\int_a^b P_n(x)dx = \sum_{j=0}^{n} (\omega_j)\, f(x_j),$$

avec les $\omega_j = \displaystyle\int_a^b L_j(x)dx$, $i = 0, ..., n$, qui sont indépendants de f et ne dépendent que de a, b et des nœuds $x_0 < ... < x_n$.

Exemple

1. ***Pour un point*** *:*
   Si $x_0 = a$, on a $P_0(x) = f(a)$ constante.
   Et donc de façon assez simple

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

   C'est la méthode des rectangles à gauche.

2. ***Pour deux points*** *:*
   Si $x_0 = a$ et $x_1 = b$, on a

$$P_1(a) = f(a),\ P_1(b) = f(b) \text{ et } deg P_1 \leq 1.$$

Par conséquent

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

Et ainsi,

$$\int_a^b f(x)dx \simeq (b - a)\frac{f(a) + f(b)}{2}.$$

C'est la méthode des trapèzes.

2.2.2 Ordre d'une formule de quadrature, erreur

1. **Ordre** :

Définition 1 (ORDRE DE QUADRATURE)

On dit qu'une formule de quadrature $\sum_{j=0}^{n} (\omega_j)\, f(x_j)$ est d'ordre p si la formule est exacte sur $\mathscr{P}_p$ mais n'est pas exacte au moins pour un polynôme de degré $p + 1$.

Exemple

(a) ***Pour la méthode des rectangles à gauche*** : $\int_a^b f(x)dx \simeq (b - a)f(a)$.

Ordre :
Cette formule est exacte pour les fonctions constantes, c'est à dire sur $\mathscr{P}_0$.
Par contre, $\int_a^b (x - a)dx = \frac{(b-a)^2}{2} \neq (b - a)(a - a) = 0$. *La formule n'est donc pas d'ordre* 1.
Conclusion, la formule est d'ordre 0.

(b) ***Pour la méthode des trapèzes*** : $\int_a^b f(x)dx \simeq (b - a)\frac{f(a) + f(b)}{2}$.

Ordre :
Cette méthode est d'ordre 2.
Elle est exacte sur $\mathscr{P}_0$ *et* $\mathscr{P}_1$
Par contre $\int_a^b (x - a)^2 dx = \frac{(b-a)^3}{3} \neq \frac{(b-a)^3}{2}$

2. **Erreur** : Nous avons vu dans le chapitre 1 sur l'interpolation polynômiale que si f est de classe $\mathscr{C}^{n+1}([a, b])$, alors pour tout $x \in [a, b]$,

$$|f(x) - P_n(x)| \leq \frac{\|f^{(n+1)}\|_\infty}{(n + 1)!}|\pi_n(x)|,$$

où $\pi_n(x) = \prod_{k=0}^{n}(x - x_k)$ et $\|f^{(n+1)}\|_\infty = \max_{t\in[a,b]} |f^{(n+1)(t)}|$.

Alors, si la formule de quadrature est obtenue à partir de P_n, et si f est de classe $\mathscr{C}^{n+1}$, on a

$$\left|\int_a^b f(x)dx - \sum_{j=0}^{n}\omega_j f(x_j)\right| = \left|\int_a^b (f(x) - P_n(x))dx\right| \leq \int_a^b |f(x) - P_n(x)|dx,$$

par l'inégalité triangulaire.
Puis par la formule de l'erreur d'interpolation

$$\left|\int_a^b f(x)dx - \sum_{j=0}^{n}\omega_j f(x_j)\right| \leq \frac{\|f^{(n+1)}\|_\infty}{(n+1)!}\int_a^b |\pi_n(x)|dx.$$

Une estimation assez grossière s'obtient alors en écrivant

$$|\pi_n(x)| = |\prod_{k=0}^{n}(x - x_k)| \leq (b-a)^{n+1}.$$

Ainsi,

$$\int_a^b |\pi_n(x)|dx \leq (b-a)^{n+2}.$$

Par conséquent, si f est de classe $\mathscr{C}^{n+1}([a,b])$,

$$\left|\int_a^b f(x)dx - \sum_{j=0}^{n}\omega_j f(x_j)\right| \leq \frac{\|f^{(n+1)}\|_\infty}{(n+1)!}(b-a)^{n+2}.$$

Exemple

Pour la méthode des rectangles à gauche : $\int_a^b f(x)dx \simeq (b-a)f(a)$.

Erreur :

$$\left|\int_a^b f(x)dx - (b-a)f(a)\right| \leq \|f''\|_\infty \int_a^b (x-a)dx = \frac{(b-a)^2}{2}\|f''\|_\infty.$$

Pour la méthode des trapèzes : $\int_a^b f(x)dx \simeq (b-a)\frac{f(a)+f(b)}{2}$.

Erreur :

$$\left|\int_a^b f(x)dx - (b-a)\frac{f(a)+f(b)}{2}\right| \leq \frac{\|f''\|_\infty}{2}\int_a^b (x-a)(b-x)dx.$$

Or,

$$\int_a^b (x-a)(b-x)dx = \int_a^b \frac{(x-a)^2}{2}dx = \frac{(b-a)^3}{6}.$$

Par conséquent, l'erreur est estimée par la formule suivante

$$\left|\int_a^b f(x)dx - (b-a)\frac{f(a)+f(b)}{2}\right| \leq \frac{(b-a)^3}{12}\|f''\|_\infty.$$

2.3 Formules composées : formules de Newton-Cotes

Plutôt que d'utiliser un polynôme d'interpolation de degré élevé sur tout $[a, b]$, on utilise des formules d'ordre peu élevé sur des sous-intervalles de $[a, b]$.

Soit $N \in \mathbb{N}*$. On divise $[a, b]$ en N sous-intervalles de même longueur $h = \dfrac{b-a}{N}$. On pose ensuite $a_k = a + k\dfrac{b-a}{N}$, pour $0 \leq k \leq N$.

Sur chaque sous-intervalle $[a_k, a_{k+1}]$, on utilise la même formule simple, en général de degré peu élevé.

Méthode :

on part d'une formule de quadrature sur $[0, 1]$

$$\int_0^1 f(x)dx \simeq \sum_{j=0}^{n} \omega_j f(x_j),$$

où les $(x_j)_{0\leq j\leq n}$ sont $(n + 1)$ points distincts dans $[0, 1]$.

A partir de cette formule, on construit des formules pour calculer $\displaystyle\int_\alpha^\beta f(x)dx$, où α et β sont des réels quelconques tels que $\alpha < \beta$. Alors par le changement de variable $x = \alpha + t(\beta - \alpha)$ (on a $dx = (\beta - \alpha)dt$, et

$$\int_\alpha^\beta f(x)dx = (\beta - \alpha)\int_0^1 f(\alpha + t(\beta - \alpha))dt,$$

Par conséquent

$$\int_\alpha^\beta f(x)dx \simeq= (\beta - \alpha)\sum_{j=0}^{n} \omega_j g(x_j),$$

soit encore

$$\int_\alpha^\beta f(x)dx \simeq= (\beta - \alpha)\sum_{j=0}^{n} \omega_j f(\alpha + x_j(\beta - \alpha)).$$

Ensuite, pour écrire la formule composée, on découpe l'intégrale en somme de N intégrales correspondant aux sous-intervalles. On obtient de cette façon

$$\int_a^b f(x)dx = \sum_{k=0}^{N-1}\int_{a_k}^{a_{k+1}} f(x)dx \simeq \sum_{k=0}^{N-1}(a_{k+1} - a_k)\sum_{j=0}^{n} \omega_i f(a_k + (a_{k+1} - a_k)x_j),$$

soit encore

$$\int_a^b f(x)dx \simeq h\sum_{k=0}^{N-1}\sum_{j=0}^{n} \omega_i f(a_k + hx_j).$$

Exemple

Pour la méthode des rectangles à gauche : $\int_a^b f(x)dx \simeq h\sum_{k=0}^{N-1} f(a_k)$,

puisque $\sum_{j=0}^{n} \omega_i f(a_k + hx_j) = f(a_k)$.

Pour la méthode des trapèzes : $\int_a^b f(x)dx \simeq h = \sum_{k=0}^{N-1} \frac{f(a_k) + f(a_{k+1})}{2}$.

Erreur :
Supposons f de classe $\mathscr{C}n + 1([a, b])$. L'erreur pour cette méthode s'établit de la façon suivante

$$\begin{aligned}
& \left| \int_a^b f(x)dx - \frac{b-a}{N} \sum_{k=0}^{N-1} \sum_{j=0}^{n} \omega_j f(a_k + hx_j) \right| \\
= & \left| \sum_{k=0}^{N-1} \int_{a_k}^{a_{k+1}} f(x)dx - \frac{b-a}{N} \sum_{k=0}^{N-1} \sum_{j=0}^{n} \omega_j f(a_k + hx_j) \right| \\
\leq & \sum_{k=0}^{N-1} \left| \int_{a_k}^{a_{k+1}} f(x)dx - \frac{b-a}{N} \sum_{j=0}^{n} \omega_j f(a_k + hx_j) \right| \\
\leq & \sum_{k=0}^{N-1} \frac{\|f^{(n+1)}\|_\infty}{(n+1)!} \left(\frac{b-a)}{N} \right)^{n+2}.
\end{aligned}$$

Dans le dernier membre de l'inégalité, nous voyons qu'aucun des facteurs ne dépend de k. Par conséquent, nous obtenons,

$$\left| \int_a^b f(x)dx - \frac{b-a}{N} \sum_{k=0}^{N-1} \sum_{j=0}^{n} \omega_j f(a_k + hx_j) \right| \leq \frac{\|f^{(n+1)}\|_\infty}{(n+1)!} (b-a) h^{n+1},$$

puisque l'on multiplie $\left(\frac{b-a)}{N} \right)^{n+2}$ par N et que $h = \frac{b-a}{N}$.
L'erreur est donc en $O(h^{n+1})$. La formule de quadrature composée est proche de l'intégrale de f à $O(h^{n+1})$ près.

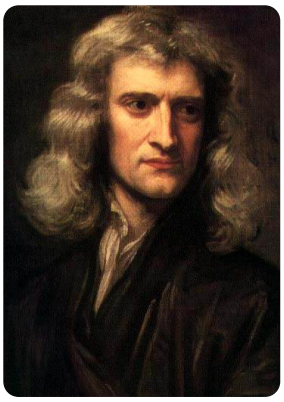
Chapitre 3

Résolution approchée d'équations non linéaires

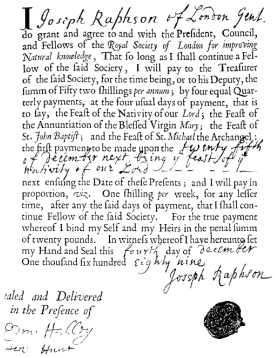
Non, la vie d'un individu n'a rien de linéaire, et pourtant son histoire est plus facile à raconter dans une apparente linéarité, dans un enchaînement qui se veut logique.

Douglas Kennedy

Sommaire



(a) Sir Isaac Newton (1642–1727), Newton qui partage avec Gottfried Wilhelm Leibniz la découverte du calcul infinitésimal, a également développé une méthode éponomyme pour trouver un point fixe.



I Joseph Raphson of London Gent.
do grant and agree to and with the President, Council, and Fellows of the *Royal Society of London for improving Natural knowledge*, That so long as I shall continue a Fellow of the said Society, I will pay to the Treasurer of the said Society, for the time being, or to his Deputy, the summ of Fifty two shillings *per annum*; by four equal Quarterly payments, at the four usual days of payment, that is to say, the Feast of the Nativity of our *Lord*; the Feast of the Annuntiation of the Blessed Virgin *Mary*; the Feast of St. *John Baptist*; and the Feast of St. *Michael* the Archangel; the first payment to be made upon the twenty fifth of December next being y^e feast of the Nativity of our Lord
next ensuing the Date of these Presents; and I will pay in proportion, viz. One shilling *per* week, for any lesser time, after any the said days of payment, that I shall continue Fellow of the said Society. For the true payment whereof I bind my Self and my Heirs in the penal summ of twenty pounds. In witness whereof I have hereunto set my Hand and Seal this fourth day of December One thousand six hundred eighty nine
Joseph Raphson
Sealed and Delivered in the Presence of

(b) Joseph Raphson (vers 1648-vers 1715), mathématicien anglais, connu pour la méthode de Newton-Raphson qu'il découvra en même temps que Newton de façon indépendante.

(c) Stefan Banach (1892-1945), mathématicien polonais est l'un des fondateurs de l'analyse fonctionnelle, à qui l'on doit entre autre un théorème éponyme du point fixe.

FIGURE 3.1 – Quelques mathématiciens célèbres liés à l'étude des nombres entiers, rationnels et réels.

Il est possible de trouver des cours et des exercices dans de nombreux ouvrages disponibles à la bibliothèque. Ceux que je suggère pour ce chapitre sont ceux de Burden et Faires [2], Filbet [4], Quarteroni *et al.* [5] et Schatzman [6].
L'objectif de ce chapitre de s'approcher le plus près possible de la solution, quand elle existe, d'un système non linéaire. Les systèmes linéaires (qui sont de la forme $f(x) = Ax + b$, où x et b sont des vecteurs de $\mathbb{R}^n$ et A une matrice carrée à n lignes et n colonnes), paradoxalement se verront en 3ème année de licence dans le cours d'analyse matricielle. Les méthodes développées pour ces systèmes linéaires sont en effet trop longues pour les exposer brièvement ici.
D'un autre côté, les systèmes non linéaires se retrouvent dans tous les problèmes issus de la physique, de la chimie et de la biologie. A titre de comparaison, considérer que tous les problèmes sont linéaires correspondrait en biologie à considérer que tous les animaux sont des éléphants. La non linéarité décrit de façon plus fine les problèmes. Mais d'un autre côté ils amènent une difficulté qui est de trouver la solution quand elle existe de ces systèmes.
La plupart du temps, nous sommes incapables de donner une forme explicite de la solution. La seule chose que nous pouvons faire est alors de l'approcher numériquement, d'aussi près que l'on veut mais surtout que l'on peut (contrainte du temps, de la puissance de l'ordinateur,...).
Les systèmes non linéaires sont de la forme

$$\begin{cases} \text{trouver} & x_* \in E, \\ \text{tel que} & f(x_*) = 0, \end{cases}$$

où E est un ensemble donné.
Ce chapitre se partagera en trois parties importantes : proposer un cadre mathématique adapté

à ce type de problème (existence, unicité des solutions et localisation de ces solutions), présenter des méthodes numériques pour approcher la solution (notion de convergence locale ou globale), définir la vitesse de convergence (ordre) de ces méthodes vers la solution exacte. Dans tout ce chapitre,

$$E \text{ désigne un sous ensemble de } \mathbb{R} \text{ et } f : E \to \mathbb{R}.$$

Nous donnerons à la fin de ce chapitre un résultat sur les systèmes d'équations linéaires où cette fois-ci

$$E \text{ désigne un sous ensemble de } \mathbb{R}^n \text{ et } f : E \to \mathbb{R}^n, n \geq 2.$$

Pour l'instant, intéressons nous seulement à la dimension 1.

3.1 Rappels sur la convergence

Commençons par rappeler quelques résultats généraux (dans n'importe quel espace métrique) sur la convergence de suite ainsi que l'ordre de convergence (c'est à dire la vitesse à laquelle la suite converge). Soient E un espace métrique (pour nous ce sera soit $\mathbb{R}$ soit $\mathbb{R}$) et F un sous $\mathbb{R}$-espace vectoriel, soit d la distance associée à E, soient K un sous ensemble de E, et $f : E \to F$. Le problème linéaire que nous étudierons se présente sous la forme

$$(\mathscr{S}) \quad \begin{cases} \text{trouver} & x_* \in E, \\ \text{tel que} & f(x_*) = 0, \end{cases}$$

Commençons par un bref rappel de la convergence d'une suite.

Définition 1 (CONVERGENCE)

Soit E un espace métrique (dans notre cas, ce sera soit $\mathbb{R}$ soit $\mathbb{R}^n$), muni de la distance $d(.,.)$ (qui est associée à $|.|$ dans $\mathbb{R}$ et $\|.\|$ dans $\mathbb{R}^n$).
Soit $(x_k)_{k\in\mathbb{N}}$ une suite de E approchant la solution x_* de $(\mathscr{S})$ donné par (3.1) et vérifiant $x_0 = x_0 \in \mathbb{K}$.
Nous disons que la suite $(x_k)_{k\in\mathbb{N}}$ converge vers $\overline{x}$ si

$$\lim_{k\to+\infty} d(x_k, x_*) = 0$$

(ou encore $\lim_{k\to+\infty} |x_k - x_*| = 0$ dans $\mathbb{R}$ et $\lim_{k\to+\infty} \|x_k - x_*\| = 0$ dans $\mathbb{R}^n$).

De plus, la méthode itérative construite avec la suite $(x_k)_{k\in\mathbb{N}}$ est convergente d'ordre p s'il existe deux constantes réelles c_1 et c_2 strictement positives telles que

$$c_1 \leq \lim_{k\to+\infty} \frac{d(x_{k+1}, x_*)}{d(x_k, x_*)^p} \leq c_2.$$

Remarque *L'ordre de convergence est important. En effet, plus il est élevé, plus le temps de calcul diminue. Et donc, nous privilégions le plus souvent les méthodes offrant des ordres de convergence les plus élevés. En effet, quand x_k est suffisamment proche de x_* pour la distance d, l'itéré suivant sera plus proche de d par la formule*

$$d(x_{k+1}, x_*) \leq c_2.d(x_k, x_*)^p.$$

Définissons ensuite les notions de convergence locale et globale. Certains de nos résultats ne sont valables que dans un voisinage assez proche de la solution (ce sera la convergence locale). Autrement dit, il faut absolument localiser la solution de notre problème pour choisir la condition initiale dans ce voisinage. Ceci peut s'avérer un inconvénient quand on n'a aucune idée de la localisation de la solution, même grossièrement. D'autres résultats sont valables quel que soit l'endroit d'où l'on démarrer l'itération. L'avantage de cette méthode, est que l'on n'a pas besoin de localiser la solution a priori.
Malheureusement, la plupart du temps nous n'avons pas le choix, et seuls des résultats locaux existent.

Définition 2 (CONVERGENCE LOCALE ET GLOBALE)

Soit E un espace métrique (dans notre cas, ce sera soit $\mathbb{R}$ soit $\mathbb{R}^n$), muni de la distance $d(.,.)$ (qui est associée à $|.|$ dans $\mathbb{R}$ et $\|.\|$ dans $\mathbb{R}^n$).
Soit $(x_k)_{k\in\mathbb{N}}$ une suite de E approchant la solution x_* de $(\mathscr{S})$ donné par (3.1) et vérifiant $x^{(0)} = x_0 \in K$.

1. Nous disons que la suite $(x_k)_{k\in\mathbb{N}}$ **converge globalement** vers $\overline{x}$ si pour tout $x_0 \in K$, la suite $(x_k)_{k\in\mathbb{N}}$ converge vers $\overline{x}$
2. Nous disons que la suite $(x_k)_{k\in\mathbb{N}}$ **converge localement** vers $\overline{x}$ s'il existe un voisinage V_{x_*} de x_*, tel que pour tout $x_0 \in V_{x_*}$, la suite $(x_k)_{k\in\mathbb{N}}$ converge vers $\overline{x}$

Commençons par la méthode la plus simple, celle qui semble être la plus intuitive, mais qui ne sera pas en général la plus efficace en terme de vitesse de convergence : la méthode de dichotomie.

3.2 Méthode de dichotomie

Cette méthode repose sur le théorème des valeurs intermédiaires.
Rappelons que si la fonction $f : [a, b] \to \mathbb{R}$ est continue et si $f(a)f(b) < 0$ alors il existe $c \in]a, b[$ telle que $f(c) = 0$.
Description de la méthode :
soit $f : [a, b] \to \mathbb{R}$ continue telle que $f(a)f(b) < 0$. On construit alors deux suites $(a_k)_{k\in\mathbb{N}}$ et $(b_k)_{k\in\mathbb{N}}$ de la façon suivante :

1. $a_0 = a$ et $b_0 = b$.
2. Si $f(\frac{a_0 + b_0}{2}) = 0$, la méthode s'arrête.

3. Sinon, on a deux sous cas :
   (a) si $f(\frac{a_0+b_0}{2}) < 0$ on pose $a_1 = a_0$ et $b_1 = \frac{a_0+b_0}{2}$,
   (b) si $f(\frac{a_0+b_0}{2}) > 0$ on pose $a_1 = \frac{a_0+b_0}{2}$ et $b_1 = b_0$.
4. On a a alors $b_1 - a_1 = \frac{b_0-a_0}{2}$.
   Et on continue l'itération jusqu'au rang k.
5. Si a_k et b_k sont construits et $f(\frac{a_k+b_k}{2}) = 0$, la méthode s'arrête.
6. Sinon, on a deux sous cas :
   (a) si $f(\frac{a_k+b_k}{2}) < 0$ on pose $a_{k+1} = a_k$ et $b_{k+1} = \frac{a_k+b_k}{2}$,
   (b) si $f(\frac{a_k+b_k}{2}) > 0$ on pose $a_{k+1} = \frac{a_k+b_k}{2}$ et $b_{k+1} = b_k$.
7. On a alors construit deux suites telles que $b_k - a_k = \frac{b_{k-1}-a_{k-1}}{2} = \frac{b_0-a_0}{2^k}$, pour tout $k \in \mathbb{N}$ et telles que la racine de f se trouve entre a_k et b_k.
8. En k étapes, on obtient une valeur de cette racine à $\frac{b_0-a_0}{2^k}$ près.

Remarque *Cette méthode est simple et elle converge puisque $\frac{b_0-a_0}{2^k}$ tend vers 0 quand k tend vers $+\infty$. Nous pouvons choisir d'arrêter l'itération quand la différence est inférieure à une précision ε donnée. Autrement dit, un rang $N \in \mathbb{N}$ tel que*

$$\frac{b_0-a_0}{2^N} \leq \varepsilon.$$

D'autre part, cette méthode ne nécessite que la continuité de f comme hypothèse.
Attention, dans ce qui précède on a supposé que f n'admettait qu'une seule racine sur $[a, b]$.
Sinon, on obtient une valeur approchée d'une des racines de f sur $[a, b]$.
Enfin, nous allons voir des méthodes moins intuitives mais plus efficaces dans ce qui suit.

La méthode suivante est basée sur le théorème du point fixe. C'est pour cette raison qu'on l'appelle la méthode du point fixe.

3.3 Méthode du point fixe

3.3.1 Théorèmes- énoncé général

Intéressons nous ici à la résolution du problème suivant :

$$\varphi(x) = x$$

Théorème 1 (THÉORÈME DU POINT FIXE)

Soit E un $\mathbb{R}$-espace vectoriel normé complet. Soit $\varphi : E \to E$ contractante, c'est à dire qu'il existe $K \in]0,1[$ tel que pour tous $x, y, in E$,

$$\|\varphi(x) - \varphi(y)\| \leq K\|x - y\|.$$

Alors

1. φ admet un unique point fixe $x_* \in E$.
2. Pour tout $x_k \in E$, la suite définie par

$$x_{k+1} = \varphi(x_k),\ x_0 \in E \text{ donné},$$

converge vers x_*.

Ici, nous allons utiliser ce théorème avec $E = \mathbb{R}$ ou $\mathbb{R}^n$ ou $[a, b]$.
La caractérisation du caractère contractant se fait en général à l'aide de la dérivée pour les fonctions définies sur $E = [a, b]$.
Supposons en effet φ de classe $\mathscr{C}^1$ sur E. Alors on a les équivalences entre

1. il existe $K \in]0,1[$, tel que $|\varphi(x) - \varphi(y)| \leq K|x - y|$, pour tous $x, y \in [a, b]$,
2. $|\varphi'(x)| \leq K$, pour tout $x \in [a, b]$,
3. $|\varphi'(x)| < 1$ pour tout $x \in [a, b]$.

Pour montrer cette caractérisation, on utilise d'autre part le théorème des accroissements finis : si $\varphi : [a, b] \to \mathbb{R}$ est continue sur $[a, b]$ et dérivable sur $]a, b[$ tel que $|\varphi'(x)| \leq K$ pour tout $x \in]a, b[$ alors

$$|\varphi(x) - \varphi(y) \leq \sup_{t \in]a,b[} |\varphi'(t)||x - y|$$

3.3.2 Construction de méthodes pour $f(x) = 0$

Le but ici est de trouver φ tel que résoudre $f(x) = 0$ soit équivalent à résoudre $\varphi(x) = x$.

1. Un premier choix simple serait de prendre $\varphi(x) = x - f(x)$.
   Supposons f de classe $\mathscr{C}^1$ sur $[a, b]$. Alors φ l'est aussi et nous avons

   $$\varphi'(x) = 1 - f'(x), \text{ pour tout } x \in [a, b].$$

   Nous souhaitons nous assurer de l'existence d'un réel positif $K < 1$ tel que

   $$|\varphi'(x)| \leq K,$$

   c'est à dire

   $$|1 - f'(x)| \leq K,$$

ou encore

$$1 - K \leq f'(x) \leq 1 + K. \tag{3.1}$$

En particulier, $f'(x) > 0$ pour tout $x \in [a, b]$. Donc f doit être strictement croissante sur $[a, b]$. C'est une condition nécessaire.
Par conséquent, la condition (3.1) est assez restrictive.

2. Un choix moins contraignant réside dans le choix suivant de φ :

$$\varphi(x) = x - \lambda f(x),$$

où λ est un réel non nul à choisir convenablement. Dans ce cas là,

$$|\varphi'(x)| \leq K \text{ pour tout } x \in [a, b]$$

est équivalent à

$$1 - K \leq \lambda f'(x) \leq 1 + K. \tag{3.2}$$

Nous devons encore avoir f' de signe constant (soit positif, soit négatif) mais la condition (3.2) reste beaucoup moins contraignante que la condition (3.1) pour un λ bien choisi.

Méthode de la corde :
La suite donnée par le théorème du point fixe, combinée avec la définition de φ ci-dessus,

$$\begin{cases} x_0 & \in [a, b] \quad \text{quelconque,} \\ x_{k+1} & = \varphi(x_k). \end{cases}$$

permet de déduire ce que nous appelons la méthode de la corde, décrite par la suite :

$$\begin{cases} x_0 & \in [a, b] \quad \text{quelconque,} \\ x_{k+1} & = x_k - \lambda f(x_k). \end{cases}$$

3.3.3 Vitesse de convergence

Nous reprenons le cadre du théorème général du point fixe :
soient E un espace de Banach, x_* l'unique point fixe de φ et K la constante de contraction.
Si nous notons $e_n = \|x_n - x_*\|$ pour tout $n \in \mathbb{N}$, alors

$$e_{n+1} = \|x_n - x_*\| = \|\varphi(x_n) - x_*\| = \|\varphi(x_n) - \varphi(x_*)\| \leq K\|x_n - x_*\|.$$

Autrement dit

$$e_{n+1} \leq K e_n, \text{ pour tout } n \in \mathbb{N}.$$

Par récurrence immédiate, nous obtenons

$$e_n \leq K^n e_0, \text{ pour tout } n \in \mathbb{N},$$

par conséquent, $\lim_{n\to+\infty} e_n = 0$ au plus à la même vitesse que K^n.
Par conséquent, plus K est petit, plus la convergence sera rapide.
Si nous regardons ce que ça donne pour la méthode de la corde citée un peu plus haut, d'après (3.3) nous avons, sous réserve que f soit de classe $\mathscr{C}^1$, $f(x) = 0$ équivalent à $\varphi(x) = x$ avec $\varphi(x) = x - \lambda f(x)$ où $\lambda \neq 0$ est à choisir tel qu'il existe $K < 1$ pour lequel

$$1 - K \leq \lambda f'(x) \leq 1 + K. \tag{3.3}$$

Ce qui pour un x fixé, nous donne pour le cas limite où $K = 0$ (le plus petit K positif possible). On se retrouve alors avec

$$1 \leq \lambda f'(x) \leq 1.$$

c'est à dire :

$$\lambda = \frac{1}{f'(x)}.$$

C'est l'idée de la méthode de Newton-Raphson que nous décrivons maintenant : en chaque point x_k, nous cherchons le meilleur λ (qui va dépendre du point x_k).

3.4 Méthode de Newton-Raphson

3.4.1 Principe

La dernière remarque de la section précédente conduit à l'écriture suivante

$$\begin{cases} x_0 \in [a,b] & \text{donné,} \\ x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)}, & n \in \mathbb{N}. \end{cases}$$

sous réserve bien sûr que f soit dérivable et de dérivée non nulle.
Une autre manière de voir la méthode de Newton-Raphson est la suivante :
nous cherchons à résoudre $f(x) = 0$ en se donnant un point de départ x_0. Nous ne savons pas en général résoudre une équation non linéaire, mais nous savons résoudre une équation affine.
L'idée ici est de remplacer le problème non linaire $f(x) = 0$ par un problème affine $g(x) = 0$, où g est la "meilleure" approximation affine" de f au voisinage de x_0.
De façon naturelle on identifie la représentation de g à la tangente à la courbe de f au point d'abscisse x_0 si f est de classe $\mathscr{C}^1$ au voisinage de x_0. En effet,

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \underset{x\to x_0}{o}(x - x_0).$$

Nous posons donc

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

et nous définissons x_1 tel que $g(x_1) = 0$. Notons que x_1 est bien défini si $f'(x_0) \neq 0$.
Nous avons alors

$$0 = f(x_0) + f'(x_0)(x_1 - x_0),$$

ou encore

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

En réitérant le procédé, nous obtenons la suite, définie par

$$\begin{cases} x_0 \in [a,b] & \text{donné,} \\ x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)}, & n \in \mathbb{N}. \end{cases}$$

Exemple
Calcul approché de $\sqrt{b}$ *pour* $b > 0$.
L'idée est de noter

$$\begin{cases} f : & \mathbb{R}_+ & \to & \mathbb{R} \\ & x & \mapsto & x^2 - b. \end{cases}$$

3.4.2 Théorème de convergence

Théorème 2 (THÉORÈME DE CONVERGENCE)

Soit $f : [a,b] \to \mathbb{R}$ une fonction de classe $\mathscr{C}^2$ sur $[a,b]$. Nous supposons qu'il existe $x_* \in]a,b[$ tel que $f(x_*) = 0$ et $f'(x_*) \neq 0$. Alors il existe $\varepsilon > 0$ tel que pour tout $x \in [x_* - \varepsilon, x_* + \varepsilon]$, la suite des itérés de Newton

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \text{ pour tout } n \in \mathbb{N},$$

1. soit bien définie pour tout $n \in \mathbb{N}$,
2. reste dans $[x_* - \varepsilon, x_* + \varepsilon]$,
3. converge vers x quand n tend vers $+\infty$,
4. admette l'existence d'un $C > 0$ tel que

$$|x_{n+1} - x_*| \leq C|x_n - x_*|^2, \text{ pour tout } n \in \mathbb{N}.$$

Remarque
La vitesse de convergence est telle que

$$e_{n+1} \leq e_n^2.$$

En général au bout de 3 ou 4 itérations, nous obtenons une précision de 10^{-8} *à* 10^{-16}.

Remarque
Cette méthode possède quand même quelques limites :

1. *il faut connaître la dérivée f' en chacun des points de la suite, ce qui peut être problématique quand f provient de données expérimentales,*
2. *il faut partir d'un point assez proche de la solution cherchée en général, donc nous avons besoin d'informations a priori précises sur f et x_*.*
   Une des solutions pour palier ce problème est d'utiliser la méthode de dichotomie pour localiser assez grossièrement x_* avant d'appliquer la méthode de Newton-Raphson.

3.5 Méthode de la sécante

L'idée ici, est de ne pas utiliser f' et donc de remplacer la dérivée $f'(x_n)$ par une différence finie

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

La méthode s'écrit alors

$$\begin{cases} x_0, x_1 \in [a,b] & \text{donnés,} \\ x_{n+1} = x_n - f(x_n)\dfrac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}, & n \in \mathbb{N}^*. \end{cases}$$

Géométriquement, pour tout $n \geq 1$ la relation précédent peut s'écrire

$$f(x_n) + \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}(x_{n+1} - x_n) = 0$$

et le terme

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

représente le coefficient directeur de la sécante au graphe de f passant par $(x_n, f(x_n))$ et $(x_{n-1}, f(x_{n-1}))$. C'est une méthode à deux pas : elle permet de calculer x_{n+1} en fonction des deux valeurs précédentes x_n et x_{n-1}.

Théorème 3 (MÉTHODE DE LA SÉCANTE)

Soit $f : [a,b] \to \mathbb{R}$ une fonction de classe $\mathscr{C}^2$ sur $[a,b]$. Nous supposons qu'il existe $x_* \in]a,b[$ tel que $f(x_*) = 0$ et $f'(x_*) \neq 0$. Alors il existe $\varepsilon > 0$ tel que pour tout $x \in [x_* - \varepsilon, x_* + \varepsilon]$, la suite de la méthode de la sécante définis par

$$x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}, n \in \mathbb{N}^*,$$

1. soit bien définie pour tout $n \in \mathbb{N}$,
2. reste dans $[x_* - \varepsilon, x_* + \varepsilon]$,
3. converge vers x quand n tend vers $+\infty$,

3.6 Ordre d'une méthode itérative

Soit $(x_n)_{n\in\mathbb{N}}$ une suite d'approximation de x_*, solution de $f(x_*) = 0$.
Nous notons

$$e_n = |x_n - x_*| \text{ pour tout } n \in \mathbb{N}.$$

Nous disons que la méthode itérative est d'ordre $\lambda \geq 1$ si λ est le $\sup$ des réels μ pour lesquels il existe $C > 0$ tel que

$$e_{n+1} \leq Ce_n^{\mu} \text{ pour tout } n \in \mathbb{N}.$$

1. Une méthode d'ordre 1 converge s'il existe un constante $C \leq 1$ telle que

   $$e_{n+1} \leq Ce_n \text{ pour tout } n \in \mathbb{N}.$$

2. Pour les méthodes d'ordre $\lambda > 1$, il faut que l'erreur initiale soit suffisamment petite.

   $$e_n \leq C^{1+\lambda+...+\lambda^{n-1}}(e_0)^{\lambda^n} \leq C^{\frac{1}{\lambda-1}}(Ce_0)^{\lambda^n}.$$

   Rappelons en effet que

   $$1 + \lambda + ... + \lambda^{n-1} = \frac{\lambda^n - 1}{\lambda - 1} \leq \frac{\lambda^n}{\lambda - 1}.$$

   Par conséquent, plus λ est grand, plus la convergence est rapide.

   (a) Méthode de Newton-Raphson : ordre 2,

   (b) Méthode sécante : ordre $\dfrac{1+\sqrt{5}}{2}$,

   (c) Méthode du point fixe : ordre 1,

   (d) Méthode de dichotomie : ordre 1.

Donnons enfin brièvement quelques résultats sur l'approximation de solutions pour les systèmes linéaire.

3.7 Systèmes d'équations non linéaires

Considérons le problème

$$f(x) = 0,$$

où cette fois-ci $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$.

3.7.1 Point fixe

Le théorème du point fixe s'applique dans $\mathbb{R}^n$. Nous considérons alors la fonction $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ définie par

$$\varphi(x) = x - f(x),$$

(ou $\varphi(x) = x - \lambda f(x)$), $\lambda \neq 0$ est à choisir comme précédemment).
Pour appliquer le théorème du point fixe, il faut que φ soit contractante. Il faut trouver une norme sur $\mathbb{R}^n$ pour laquelle il existe $K < 1$ tel que pour tous $x, y \in \mathbb{R}$,

$$\|\varphi(x) - \varphi(y)\| \leq K\|x - y\|.$$

Remarque
Si φ est de classe $\mathscr{C}^2$, nous avons encore une inégalité des accroissements finis sur $\mathbb{R}^n$.
Si nous fixons $\mathscr{J}_{\varphi(x)}$, la jacobienne de φ au point x définie par

$$\mathscr{J}_\varphi(x) = \left(\frac{\partial \varphi_i}{\partial x_j}\right)_{0 \leq i,j \leq n}(x),$$

alors

$$\|\varphi(x) - \varphi(y)\| \leq \sup_{t \in [0,1]} |||\mathscr{J}_\varphi(tx + (1-t)y)|||\|x - y\|.$$

3.7.2 Méthode de Newton-Raphson dans $\mathbb{R}^n$

C'est à peu près la même idée que dans $\mathbb{R}$. Nous remplaçons f au voisinage du point considéré par sa meilleure approximation affine

$$f(x) = f(x_0) + Df(x_0)(x - x_0) + \underset{x \to x_0}{o}\|x - x_0\|,$$

ou encore

$$f(x) = f(x_0) + \mathscr{J}_f(x_0)(x - x_0) + \underset{x \to x_0}{o}\|x - x_0\|.$$

On pose

$$g(x) = f(x_0) + \mathscr{J}_f(x_0)(x - x_0),$$

et x_1 est défini par $g(x_1) = 0$, et on obtient, si $\mathscr{J}_f(x_0)$ est inversible

$$x_1 = x_0 - (\mathscr{J}_f(x_0))^{-1} f(x_0).$$

La méthode s'écrit alors

$$\begin{cases} x_0, \in R^d & \text{donné,} \\ x_{n+1} = x_n - (\mathscr{J}_f(x_0))^{-1} f(x_n), & n \in \mathbb{N}^*. \end{cases}$$

On a un résultat de convergence similaire à celui vu pour $d = 1$.

3.7.3 A propos des systèmes linéaires et des méthodes itératives (hors programme)

Nous cherchons à résoudre $Ax = b$, où b est un vecteur de $\mathbb{R}^n$ donné et $A \in \mathscr{M}_n(\mathbb{R})$ est une matrice inversible.
Nous écrivons $A = M - N$, M inversible et nous proposons l'algorithme suivant :

$$\begin{cases} x_0, \in R^d & \text{donné,} \\ Mx_{k+1} = Nx_k + b, & k \in \mathbb{N}^*. \end{cases}$$

Ce qui se réécrit pour tout $k \in \mathbb{N}^*$ par

$$x_{k+1} = M^{-1}Nx_k + M^{-1}b.$$

Si nous posons

$$\begin{array}{rccl} f: & \mathbb{R}^n & \to & \mathbb{R}^n, \\ & x & \mapsto & M^{-1}Nx + M^{-1}b, \end{array}$$

alors pour tout $k \in \mathbb{N}^*$

$$x_{k+1} = f(x_k).$$

Si la suite $(x_k)_{k\in\mathbb{N}}$ converge, elle converge vers un point fixe de f.
Soit $\|.\|$ une norme de $\mathbb{R}^n$.
Alors pour tous $x, y \in \mathbb{R}^n$,

$$\|f(x) - f(y)\| = \|M^{-1}N(x - y)\| \leq |||M^{-1}N|||\|x - y\|.$$

Si pour une norme $|||M^{-1}N||| < 1$ on a f contractante et le théorème du point fixe s'applique.

Chapitre 4

Résolution approchée d'équations différentielles

Les schémas du mathématiciens, comme ceux du peintre ou du poète, doivent être beaux ; les idées, comme les couleurs ou les mots, doivent s'assembler de façon harmonieuse. La beauté est le premier test : il n'y a pas de place durable dans le monde pour les mathématiques laides

Godfrey Harold Hardy

(a) Peter Lax (1926 –), mathématicien hongrois de nationalité américaine, on lui doit la preuve d'une condition nécessaire et suffisante pour la convergence d'un schéma numérique consistant et stable.

(b) Carl Runge (1856 – 1927), mathématicien et physicien allemand, on lui doit avec Martin Kutta, une des méthodes les plus utilisées de résolution numérique des équations différentielles, la méthode de Runge-Kutta.

(c) Martin Kutta (1867 – 1944), mathématicien allemand, il a participé avec Carl Runge à l'élaboration de la méthode de Runge-Kutta, utilisée pour résoudre des équations différentielles.

FIGURE 4.1 – Quelques mathématiciens célèbres liés à l'étude des schémas numériques des équations différentielles.

Sommaire

Il est possible de trouver des cours et des exercices dans de nombreux ouvrages disponibles à la bibliothèque. Ceux que je suggère pour ce chapitre sont ceux de Burden et Faires [2],Demailly [3], Filbet [4], Quarteroni *et al*. [5] et Schatzman [6].
Pour un rappel sur le cours d'équations différentielles, je suggère le livre de Benzoni-Gavage [1] L'objectif de ce chapitre est de chercher à calculer de façon approchée la solution d'un problème de Cauchy. Il se trouve que très souvent, il est impossible de résoudre ce type de problème de façon explicite.
Il faut alors faire appel à des schémas numériques efficaces permettant d'approcher les solutions explicites. Nous verrons ce que sont les principaux schémas existants et surtout ce que signifie un schéma efficace.

4.1 Rappel sur les équations différentielles et le problème de Cauchy

4.1.1 Reconnaître les différents types d'équations

Introduisons ici quelques définitions essentielles pour la suite de ce cours.

Définition 1 (ÉQUATION DIFFÉRENTIELLE ORDINAIRE)

Une équation différentielle ordinaire, également notée EDO, d'ordre n est une relation entre la variable réelle t, une fonction inconnue $t \mapsto x(t)$ et ses dérivées x', x'',...,$x^{(n)}$ au point t définie par

$$F(t, x, x'', ..., x^{(n)}) = 0, \tag{4.1}$$

où F n'est pas indépendante de sa dernière variable $x^{(n)}$. On prendra t dans un intervalle I de $\mathbb{R}$ (I peut être $\mathbb{R}$ tout entier).
La solution x en général sera à valeurs dans $\mathbb{R}^N$, $N \in \mathbb{N}^*$ où N sera le plus souvent égal à 1, 2 ou 3. On dit que cette équation est scalaire si F est à valeurs dans $\mathbb{R}$.

Définition 2 (ÉQUATION DIFFÉRENTIELLE NORMALE)

On appelle équation différentielle normale d'ordre n toute équation de la forme

$$x^{(n)} = f(t, x, x'', ..., x^{(n-1)}). \tag{4.2}$$

Définition 3 (ÉQUATION DIFFÉRENTIELLE AUTONOME)

On appelle équation différentielle autonome d'ordre n toute équation de la forme

$$x^{(n)} = f(x, x'', ..., x^{(n-1)}). \tag{4.3}$$

Autrement dit, f ne dépend pas explicitement de t.

4.1.2 Equation linéaire

Donnons maintenant une classification par linéarité.

Définition 4 (ÉQUATION DIFFÉRENTIELLE LINÉAIRE)

Une EDO de type (4.1) d'ordre n est linéaire si elle est de la forme

$$a_n(t)x^{(n)}(t) + a_{n_1}(t)x^{(n-1)}(t) + ... + a_1(t)x'(t) + a_0(t)x(t) = g(t), \tag{4.4}$$

avec tous les $x^{(i)}$ de degré 1 et tous les coefficients dépendant au plus de t.

Exemple
Dire si les équations différentielles suivantes sont linéaires, ou non linéaires, et donner leur ordre (on justifiera la réponse) :

$$i.\ (x-t)dt + 4tdx = 0 \quad ii.\ x'' - 2x' + x = 0 \quad iii.\ \frac{d^3x}{dt^3} + t\frac{dx}{dt} - 5x = e^t$$

$$iv.\ (1-x)x' + 2x = e^t \quad v.\ \frac{d^2x}{dt^2} + \sin x = 0 \quad vi.\ \frac{d^4x}{dt^4} + x^2 = 0$$

4.1.3 Solutions

Définition 5 (SOLUTION)

On appelle solution (ou intégrale) d'une équation différentielle d'ordre n sur un certain intervalle I de $\mathbb{R}$, toute fonction x définie sur cet intervalle I, n fois dérivable en tout point de I et qui vérifie cette équation différentielle sur I.
On notera en général cette solution (x, I).
Si I contient sa borne inférieure notée a (respectivement sa borne supérieure b), ce sont des dérivées à droite (respectivement à gauche) qui interviennent au point $t = a$ (respectivement $t = b$).
Intégrer une équation différentielle consiste à déterminer l'ensemble de ses solutions.

Définition 6 (COURBE INTÉGRALE-ORBITE)

On appelle courbe intégrale l'ensemble des points $(t, x(t))$ où t parcourt I. Autrement dit, si x est à valeurs dans $\mathbb{R}^N$, la courbe intégrale est un ensemble de points de $\mathbb{R}^{N+1}$.
On appelle orbite, l'ensemble des points $x(t)$ où t parcourt I : c'est un ensemble de points de $\mathbb{R}^N$.
L'espace $\mathbb{R}^N$ où les solutions prennent leurs valeurs s'appelle espace de phases.

4.1.4 Equations à variables séparées

Exemple
Considérons l'EDO d'ordre 1 sous forme normale données par l'équation

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

L'idée est d'exprimer $f(t, x)$ sous la forme $g(t)h(x)$, où $g : I \to \mathbb{R}$ et $h : J \subset \mathbb{R} \to \mathbb{R}$. Ce qui permettra de résoudre une équation du type

$$x' = g(t)h(x).$$

Cas particulier :
Les équations les plus simples sont de la forme

$$x' = f(t),$$

avec $h \equiv 1$ et $g(t) = f(t)$ pour tout $t \in I$. On suppose en outre que $x(t_0) = x_O$ pour un $t_0 \in I$.
Si on suppose que f est continue sur un intervalle $I \subset \mathbb{R}$ d'intérieur non vide. Les solutions de cette équation sont données par

$$x(t) = x_0 + \int_{t_0}^{t} f(s)ds,$$

Définition 7 (EQ. A VARIABLES SÉPARÉES)

On appelle de façon générale équation à variables séparées, toute équation de la forme

$$b(x)x' = a(t), \tag{4.5}$$

où a et b sont deux fonctions définies respectivement sur I et K, et où I et K sont des intervalles de $\mathbb{R}$.

4.1.5 Résolution d'équations linéaires

Nous restons toujours sur les EDO d'ordre 1. Nous nous intéressons ici aux équations différentielles ordinaires linéaires.

Définition 8 (EDO LINÉAIRE)

Une équation différentielle du premier ordre est dite linéaire si elle est linéaire par rapport à la fonction inconnue x et par rapport à sa dérivée x'. Une telle équation peut toujours s'écrire sous la forme

$$a(t)x' + b(t)x = d(t). \tag{4.6}$$

Dans toute la suite, on supposera que a, b et d sont continues sur un intervalle $I \subset \mathbb{R}$.

EDO linéaire sans second membre

Commençons par résoudre une équation linéaire d'ordre 1 sans second membre. On l'appelle EDO linéaire du premier ordre homogène. C'est une équation de la forme

$$a(t)x' + b(t)x = 0. \tag{4.7}$$

C'est une équation à variables séparables sur $I \times J$ tel que $a(t) \neq 0$ pour tout $t \in I$.
Il est à noter que $x \equiv 0$ est une solution de l'équation linéaire homogène ci-dessus. On l'appelle solution triviale comme dans le cas des équations autonomes.

Proposition 1 (SOL. EQ. HOMOGENES)

L'ensemble des solutions de l'équation linéaire homogène

$$a(t)x' + b(t)x = 0.$$

sur le domaine I, avec pour un certain t_0 dans I tel que $x(t_0) = x_0$ est définie pour tout $t \in I$ par

$$x(t) = x_0 e^{F(t)},$$

avec $F(t) = \displaystyle\int_{t_0}^{t} -\frac{b(s)}{a(s)}$.

Proposition 2 (SOLUTION TRIVIALE)

Si une solution de l'équation linéaire homogène s'annule en au-moins un point t_0 alors elle est identiquement nulle (solution triviale).

Remarque
La solution $x \equiv 0$ sur I est appelée intégrale dégénérée de l'équation linéaire homogène.

EDO linéaire avec second membre

Considérons l'équation

$$a(t)x' + b(t)x = d(t),$$

sur l'intervalle I où a ne s'annule pas.
Soit x_h une solution particulière non dégénérée de l'équation homogène associée à l'équation ci-dessus sur I.

Proposition 3 (SOLUTION GÉNÉRALE)

La solution générale de l'équation

$$a(t)x' + b(t)x = d(t),$$

sur I avec pour un certain t_0 dans I tel que $x(t_0) = x_0$ est donnée par

$$x(t) = \exp\left(-\int_{t_0}^{t} \frac{b(s)}{a(s)} ds\right)\left(x_0 + \int_{t_0}^{t} \frac{d(s)}{a(s)} \exp\left(\int_{t_0}^{s} \frac{b(\sigma)}{a(\sigma)} d\sigma\right) ds\right).$$

Remarque
La méthode fréquemment utilisée pour trouver une solution de l'équation linéaire non homogène à partir de l'équation homogène est appelée méthode de variation de la constante.

Cas particulier

Proposition 4 (FORMULE DE DUHAMEL)

Soient une fonction continue sur un intervalle I de $\mathbb{R}$, α une constante réelle et $t_0 \in I$ tel que $x(t_0) = x_0$. La solution générale de l'équation scalaire

$$x' = \alpha x + f(t),$$

est donnée par

$$x(t) = x_0 e^{\alpha(t-t_0)} + \int_{t_0}^{t} e^{\alpha(t-s)} f(s) ds,$$

où c est une constante.

4.1.6 Cas particulier d'équations non linéaires : équations de Bernoulli

Définition 9 (ÉQUATION DE BERNOULLI)

Une équation de Bernoulli est une équation différentielle scalaire non linéaire de la forme

$$x' + P(t)x + Q(t)x^r = 0, \tag{4.8}$$

où $r \in \mathbb{R}$, P et Q sont deux fonctions définies et continues sur un intervalle I de $\mathbb{R}$.

Remarque

On peut éliminer les cas $r = 0$ et $r = 1$, car l'équation de Bernoulli correspond alors à une équation que l'on connaît déjà et que l'on a traité dans la section précédente.

Théorème 1 (SOLUTION BERNOULLI)

Une fonction dérivable strictement positive (au cas où $r = 1/2$ par exemple, où $r \leq 0$) x sur I est solution de l'équation de Bernoulli si et seulement si $u = x^{1-r}$ est une solution strictement positive de l'équation linéaire

$$u' + (1-r)P(t)u + (1-r)Q(t) = 0. \tag{4.9}$$

Remarque

1. *Connaissant la solution u de l'équation linéaire associée à l'équation de Bernoulli, on peut en déduire les solutions strictement positives de l'équation de Bernoulli.*
2. *Nous pouvons trouver quelques propriétés sur les solutions suivant les valeurs de r :*
   a. *Si $r > 0$ l'équation de Bernoulli admet la solution triviale $x \equiv 0$.*
   b. *Si $r > 1$ toute solution de l'équation de Bernoulli qui prend la valeur 0 en un point, est partout nulle.*
   c. *Si $0 < r < 1$, la fonction nulle n'est pas nécessairement la seule solution qui prenne la valeur 0 en un point.*
3. *L'équation particulière*
$$t^2x' + x + x^2 = 0, \tag{4.10}$$
est appelée équation de Ricatti.

4.1.7 Théorème de Cauchy-Lipschitz

Problème de Cauchy

Soit U un ouvert de $\mathbb{R} \times \mathbb{R}^m$ et $f : U \to \mathbb{R}^m$ une fonction. On note $\|.\|$ une norme quelconque sur $\mathbb{R}^m$ (on a vu en analyse III que toutes les normes sont équivalentes sur $\mathbb{R}^m$.

Définition 10 (PROBLÈME DE CAUCHY)

Etant donnée une équation différentielle du premier ordre sous forme normale

$$x' = f(t, x), \tag{4.11}$$

pour $(t, x(t)) \in U$, et un point $(t_0, x_0) \in U$, le problème de Cauchy correspondant est la recherche des solutions x telles que

$$x(t_0) = x_0. \tag{4.12}$$

Notation :
On note le problème de Cauchy de la façon suivante

$$\begin{cases} x' &= f(t, x), \\ x(t_0) &= x_0. \end{cases} \tag{4.13}$$

Définition 11 (SOLUTION DU PROBLÈME DE CAUCHY)

Une solution du problème de Cauchy (4.13) sur un intervalle ouvert I de $\mathbb{R}$ avec la condition initiale $(t_0, x_0) \in U$ et $t_0 \in I$ est une fonction dérivable $x : I \to \mathbb{R}^m$ telle que

i. pour tout $t \in I$, $(t, x(t)) \in U$,
ii. pour tout $t \in I$, $x'(t) = f(t, x(t))$,
iii. $x(t_0) = x_0$.

Théorème 2 (SOLUTIONS DE (4.13))

Supposons $f : U \to \mathbb{R}^m$ continue. Soit $(t_0, x_0) \in U$ et x une fonction définie sur un intervalle ouvert I contenant t_0 et à valeurs dans $\mathbb{R}^m$.
Une fonction x est solution de (4.13) sur I si et seulement si

i. pour tout $t \in I$, $(t, x(t)) \in U$,
ii. x est continue sur I,
iii. pour tout $t \in I$, $x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))ds$.

Preuve :
Faite en cours.

Existence et unicité locale

Énonçons tout d'abord un résultat local d'existence et d'unicité .

Théorème 3 (CAUCHY LIPSCHITZ)

Soient $f \in \mathscr{C}(U;\mathbb{R}^N)$ où U est un ouvert de $\mathbb{R}\times\mathbb{R}^m$, et $(t_0,x_0)\in U$. On suppose f lipschitzienne par rapport à sa variable x sur un voisinage de (t_0,x_0), c'est à dire qu'il existe un voisinage de (t_0,x_0) dans U et $L>0$ tel que pour tous (t,x) et (t,y) dans ce voisinage

$$\|f(t,x)-f(t,y)\|\leq L\|x-y\|. \tag{4.14}$$

Alors on a les propriétés suivantes.

1. Existence : Il existe $T>0$ et $x\in\mathscr{C}^1([t_0-T,t_0+T];J)$ solution du problème de Cauchy
$$\left\{\begin{array}{lcl} x' & = & f(t,x),\\ x(t_0) & = & x_0.\end{array}\right.$$
2. Unicité : Si y est une autre solution du problème de Cauchy ci-dessus, elle coïncide avec x sur un intervalle d'intérieur non vide inclus dans $[t_0-T,t_0+T]$.
3. Régularité Si de plus f est de classe $\mathscr{C}^r$, $r\geq 1$, alors x est de classe $\mathscr{C}^{r+1}$.

Preuve :
Faite en cours.

Remarque

1. *Dès que f est de classe $\mathscr{C}^1$ elle est localement lipschitzienne (ce résultat découle du théorème des accroissements finis). C'est un résultat connu découlant du théorème des accroissements finis.*
2. *A partir de maintenant, on considère un cas, légèrement plus particulier (pour simplifier les énoncés des propriétés), où f est définie sur $I\times J$, avec I un intervalle ouvert non vide de $\mathbb{R}$ et J un intervalle ouvert non vide de $\mathbb{R}^m$ et non plus sur un domaine ouvert quelconque U inclus dans $\mathbb{R}\times\mathbb{R}^m$.*

Théorème 4 (EXISTENCE ET UNICITÉ GLOBALE)

On suppose $f\in\mathscr{C}(I\times\mathbb{R}^m;\mathbb{R}^m)$ et globalement lipschitzienne par rapport à x.
Alors, quel que soit $(t_0,x_0)\in I\times\mathbb{R}^m$, il existe un unique $x\in\mathscr{C}^1(I;\mathbb{R}^m)$ solution de (4.13).

4.2 Simulations numériques des EDO : schémas explicites

Nous ne considérons dans cette partie que le problème de Cauchy d'ordre 1 (4.15) ci-dessous (nous verrons en cours d'équations différentielles que quel que soit l'ordre d'une EDO, il est toujours possible de se ramener à l'ordre 1) suivant :

$$(\mathscr{C}) \quad \begin{cases} x'(t) &= f(t,x), \\ x(t_0) &= x_0. \end{cases} \tag{4.15}$$

Mais attention, cette fois-ci $t \in [t_0, t_0+T]$ (avec $T \in \mathbb{R}_+^*$), étant donné que l'on ne peut pas faire de simulation pendant un temps infini, le temps doit rester fini.
D'autre part, $x : [t_0, t_0+T] \to \mathbb{R}^n$, $n \in \mathbb{N}$. Nous supposerons quasiment tout le temps pour les exercices d'applications que $n = 1$.

Remarque
Nous supposerons dans toute la suite, que la fonction $f : [t_0, t_0+T] \times \mathbb{R}^n \to \mathbb{R}^n$ est continue sur l'ensemble étudié, est globalement lipschitzienne par rapport à sa variable x sur l'intervalle étudié. En général, il nous suffira de montrer que la dérivée de f par rapport à sa variable x est bornée sur l'intervalle de t étudié. De telle sorte que nous n'aurons pas à nous soucier de l'existence globale et de l'unicité des solutions.

L'objectif est de calculer de façon approchée la solution du problème de Cauchy (4.15).
Pour cela nous discrétisons l'intervalle de temps $[t_0, t_0+T]$ de la façon suivante : on définit une subdivision

$$t_0 < t_1 < t_2 < ... < t_N = t_0 + T,$$

et on cherche à définir

$$x_0, x_1, x_2, ..., x_N,$$

tels que les x_i, $i = 0, ..., N$, approchent les $x(t_i)$ (solution exacte aux points t_i).
On note

$$h_i = t_i - t_{i-1} > 0,$$

le i-ème pas de temps et $h = \max\limits_{i=1,...,N} h_i > 0$.

Remarque

1. *Dans toute la suite, nous supposerons les t_i régulièrement espacés. Autrement dit*

$$h = h_1 = h_2 = ... = h_N > 0,$$

2. *Si nous divisons l'intervalle $[t_0, t_0+T]$ de longueur T, en N sous intervalles, nous aurons alors*

$$T = N.h \quad \textit{ou encore} \quad h = T/N \ (\textit{avec } N \in \mathbb{N}^*).$$

Commençons par le schéma le plus simple : le schéma d'Euler progressif ou explicite.

4.2.1 Schéma d'Euler progressif (ou explicite)

On considère le problème de Cauchy $(\mathscr{C})$ (4.15). Il existe deux méthodes pour construire ce schéma.

Méthode 1

Pour un point $t = t_n$, quelconque avec $n = 0, 1, ..., N-1$, nous avons l'EDO qui satisfait

$$\begin{cases} x'(t_n) &= f(t_n, x(t_n)), \\ x(t_0) &= x_0 \text{ donné.} \end{cases}$$

On approche alors $f(t_n, x(t_n))$ par $f(t_n, x_n)$, et il faut approcher ensuite $x'(t_n)$. Pour cela il semble naturel d'approcher la dérivée par son taux d'accroissement aux valeurs approchées x_{n+1} et x_n, autrement dit

$$x'(t_n) \approx \frac{x(t_{n+1}) - x(t_n)}{t_{n+1} - t_n} \approx \frac{x_{n+1} - x_n}{t_{n+1} - t_n} = \frac{x_{n+1} - x_n}{h}.$$

Le schéma d'approximation du système $(\mathscr{C})$ au point t_n est donc

$$\begin{cases} \dfrac{x_{n+1} - x_n}{h} &= f(t_n, x_n), \quad \text{pour } \ n = 0, ..., N-1, \\ x(t_0) &= x_0. \end{cases}$$

soit encore,

$$(\mathscr{E}) \quad \begin{cases} x_{n+1} &= x_n + hf(t_n, x_n), \quad \text{pour } \ n = 0, ..., N-1, \\ x(t_0) &= x_0. \end{cases}$$

C'est ce que l'on appelle le schéma d'Euler progressif ou explicite. On dit qu'il est à un pas parce que x_{n+1} est défini seulement à partir du pas précédent x_n et non pas des autres pas précédents (on l'aurait alors appelé multipas).

Méthode 2

Une autre méthode pour construire ce schéma est d'utiliser la méthode des rectangles à gauche. En effet, rappelons le problème au point $t \in [t_0, t_0 + T]$:

$$\begin{cases} x'(t) &= f(t, x(t)), \\ x(t_0) &= x_0 \text{ donné.} \end{cases}$$

D'après le théorème de Cauchy-Lipschitz (et par les hypothèses que nous avons prises), nous avons x' et f continues. Nous pouvons donc intégrer l'équation différentielle entre t_n et t_{n+1}. Nous avons ainsi, pour tout $n = 0, 1, ..., N-1$,

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

L'intégrale du premier membre est connue, il s'agit du théorème fondamental du calcul, c'est la valeur $x(t_{n+1}) - x(t_n)$. L'intégrale du second membre est quant à elle approchée par la méthode des rectangles à gauches :

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

Nous avons alors

$$x(t_{n+1}) - x(t_n) \approx f(t_n,x(t_n)).h,$$

et comme nous approchons les $x(t_n)$ par x_n pour $n = 0,...,N$, nous obtenons bien le même système qu'avec la première méthode, à savoir

$$(\mathscr{E}) \quad \left\{ \begin{array}{rcl} x_{n+1} & = & x_n + hf(t_n,x_n), \quad \text{pour} \quad n = 0,...,N-1, \\ x(t_0) & = & x_0. \end{array} \right.$$

L'avantage de cette méthode est qu'il est plus facile de comprendre la construction d'autres schémas que nous verrons dans la section 4.2.3 :

1. le schéma du point milieu où l'intégrale est approchée par le rectangle passant dont la hauteur est la solution exacte au milieu de l'intervalle,
2. le schéma de Heun donc l'approximation de l'intégrale se fait par la méthode des trapèzes.

Notons également que nous aurions pu prendre la méthode des rectangles à droite, nous n'aurions alors pas de schéma explicite mais un schéma d'Euler implicite (appelé aussi schéma rétrograde). Nous détaillerons ceci dans une des dernières sections de ce chapitre.
Nous voyons que les erreurs peuvent très vite s'accumuler. La question que l'on se pose est alors la suivantes :
est-ce que la solution approchée tend vers la solution de l'EDO quand h tend vers 0 (c'est à dire quand la discrétisation est très fine et les intervalles très petits) ?
Pour l'étudier nous allons introduire trois notions fondamentales dans ce chapitre : la consistance, la stabilité et la convergence pour des schémas explicites quelconques.

4.2.2 Schémas explicites à un pas

Définition 12 (SCHÉMAS EXPLICITES A UN PAS)

Un schéma explicite à un pas approchant la solution d'un problème de Cauchy $(\mathscr{C})$ donné par (4.15) est un schéma qui s'écrit sous la forme

$$(\mathscr{S}) \quad \left\{ \begin{array}{rcl} x_{n+1} & = & x_n + h\phi(t_n,x_n,h), \quad \text{pour} \quad n = 0,...,N-1, \\ x(t_0) & = & x_0, \end{array} \right. \tag{4.16}$$

où la fonction ϕ sera à définir suivant le schéma choisi.

Exemple *Par exemple, pour le schéma d'Euler explicite nous avons*

$$\phi(t_n,x_n,h) = f(t_n,x_n)$$

4.2.3 Consistance, stabilité et convergence

Nous pouvons alors introduire les notions fondamentales de ce chapitre. Commençons par la consistance.

Définition 13 (ERREUR DE CONSISTANCE)

On appelle erreur de consistance, que l'on note τ_n, le nombre réel défini par

$$\tau_{n+1}(h) = x(t_{n+1}) - x(t_n) - h.\phi(t_n, x(t_n), h), \text{ pour tout } n = 0, ..., N-1.$$

Remarque

Une autre façon d'écrire l'erreur de consistance est comme ceci :

$$x(t_{n+1}) = x(t_n) + h.\phi(t_n, x(t_n), h) + \tau_{n+1}(h), \textit{ pour tout } n = 0, ..., N-1.$$

Ca s'interprète assez simplement. L'erreur de consistance τ_{n+1} est l'erreur que l'on ferait si l'on partait de la solution $x(t_n)$ comme condition initiale et que l'on ne faisait qu'une seule itération. Nous voyons ainsi ce que chacun des pas séparément peut engendrer individuellement comme erreur en supposant que l'on n'a pas fait d'approximation dans les pas précédents.

Définition 14 (SCHÉMA CONSISTANT)

On dit que le schéma explicite à un pas ($\mathscr{S}$) donné par (4.16) est consistant avec le problème de Cuachy ($\mathscr{C}$) si et seulement si

$$\lim_{h \to 0} \sum_{n=1}^{N} |\tau_n(h)| = 0$$

Une fois l'erreur de consistance définie, nous nous intéressons à comment les erreurs peuvent s'accumuler. Il se trouve en effet que les ordinateurs sont obligés de faire des arrondis. Par exemple, si la solution trouvée est $1/3$, l'ordinateur va arrondir à $0,3333333$. Et même avec une précision assez fine, il aura quand même perturbé légèrement le calcul précédent. La question que l'on se pose est la suivante :
si le schéma est sensible à la moindre perturbation, les approximations risquent de ne pas être maîtrisée du tout. Il faut donc trouver un critère permettant de vérifier que ces petites perturbations sont sous contrôle. Comment ? Grâce à la notion de stabilité.

Définition 15 (SCHÉMA STABLE)

On dit que le schéma explicite à un pas $(\mathscr{S})$ donné par (4.16) est stable s'il existe un réel $M \geq 0$ tel que, étant donnés $\varepsilon_1, ..., \varepsilon_N$, si l'on considère pour le schéma explicite

$$x_{n+1} = x_n + h + \phi(t_n, x_n, h), \text{ avec } x_0 \text{ donné,}$$

une perturbation de ce schéma que l'on note

$$y_{n+1} = y_n + h + \phi(t_n, x_n, h) + \varepsilon_{n+1}, \text{ avec } y_0 = x_0,$$

alors

$$\max_{0 \leq n \leq N} \|y_n - x_n\| \leq M \sum_{n=1}^{N} |\varepsilon_n|.$$

Autrement dit, si les erreurs d'approximation à chaque pas de temps ne sont pas très grandes, l'erreur pour la solution approchée au pas suivant reste maîtrisée.
Nous pouvons enfin introduire la troisième et dernière notion fondamentale : la notion de convergence.

Définition 16 (CONVERGENCE)

On dit que le schéma explicite à un pas $(\mathscr{S})$ donné par (4.16) est convergent vers la solution du problème de Cauchy $(\mathscr{C})$ si

$$\lim_{h \to 0} \max_{1 \leq n \leq N} \|x(t_n) - x_n\| = 0.$$

Et le théorème permettant de relier les trois notions.

Théorème 5 (THÉORÈME DE LAX)

Si le schéma explicite à un pas $(\mathscr{S})$ donné par (4.16) est stable et consistant avec le problème de Cauchy $(\mathscr{C})$, alors il converge vers la solution du problème de Cauchy $(\mathscr{C})$.

Le problème reste de vérifier que le schéma est stable et consistant. Mais en utilisant la définition seulement, l'exercice peut très vite s'avérer difficile. Il faudrait trouver des critères plus faciles à manipuler pour obtenir ces deux notions sans trop de difficulté. C'est tout le but de ce qui suit. Commençons par un critère simple pour vérifier la consistance.

Proposition 5 (CRITÈRE DE CONSISTANCE)

Un schéma explicite à un pas ($\mathscr{S}$) donné par (4.16) est consistant avec le problème de Cauchy ($\mathscr{C}$) si et seulement si

$$\phi(t, x, 0) = f(t, x).$$

Puis un critère simple pour la stabilité.

Proposition 6 (CRITÈRE DE STABILITÉ)

S'il existe un réel $L > 0$ tel que pour tout $t \in [t_0, t_0 + T]$, pour tous x, $y \in \mathbb{R}$ et pour tout $h \leq T$,

$$\|\phi(t, x, h) - \phi(t, y, h)\| \leq L\|x - y\|,$$

alors le schéma explicite à un pas ($\mathscr{S}$) donné par (4.16) est stable et $M = e^{LT}$.

Quand un système est convergent, il nous reste à savoir maintenant à quelle vitesse il converge. Pour cela on introduit la notion d'ordre. L'ordre de la convergence est liée à l'ordre de la consistance, et plus l'ordre est élevé, plus le schéma converge rapidement vers la solution exact. Voyons ça d'un peu plus près.

Définition 17 (ORDRE DE CONSISTANCE)

On dit que le schéma explicite à un pas ($\mathscr{S}$) donné par (4.16) est consistant d'ordre p avec le problème de Cauchy ($\mathscr{C}$), s'il existe un réel $K > 0$ tel que

$$\sum_{n=1}^{N} |\tau_n(h)| \leq Kh^p.$$

Proposition 7 (ORDRE DE CONVERGENCE)

Si un schéma explicite à un pas ($\mathscr{S}$) donné par (4.16) est stable et consistant d'ordre p avec le problème de Cauchy ($\mathscr{C}$) alors

$$\max_{1 \leq n \leq N} \|x(t_n) - x_n\| \leq MKh^p.$$

On dit alors qu'il est convergent d'ordre p.

Là encore utiliser la définition peut s'avérer assez compliqué. Il faudrait un critère permettant de montrer l'ordre de consistance sans trop de difficulté.

Proposition 8 (ORDRE DE CONVERGENCE :CRITÈRES)

Pour toute fonction f de classe $\mathscr{C}^p$, le schéma explicite à un pas $(\mathscr{S})$ donné par (4.16) est consistant d'ordre au moins p avec le problème de Cauchy $(\mathscr{C})$ si

$$\begin{aligned} \phi(t,x,0) &= f(t,x), \\ \frac{\partial}{\partial h}\phi(t,x,0) &= \frac{1}{2}Df(t,x) = \frac{1}{2}[\frac{\partial}{\partial t}f(t,x) + f(t,x)\frac{\partial}{\partial x}f(t,x)], \\ &\vdots \\ \frac{\partial^{p-1}}{\partial h^{p-1}}\phi(t,x,0) &= \frac{1}{p}D^{p-1}f(t,x), \end{aligned}$$

avec $D^{p-1}f(t,x) = D[D^{p-2}f(t,x)]$.

Exemple *La différentielle seconde de f est donnée en (t,x) par :*

$$D^2 f = D[Df] = \frac{\partial^2}{\partial t^2}f + 2f\frac{\partial^2}{\partial x \partial t}f + \frac{\partial}{\partial t}f\frac{\partial}{\partial x}f + f(\frac{\partial}{\partial x}f)^2 + f^2\frac{\partial^2}{\partial x^2}f.$$

Exemple *On peut montrer que si f satisfait les hypothèse de l'existence et l'unicité globale, alors le schéma d'Euler explicite correspondant au problème de Cauchy $(\mathscr{C})$ est convergent d'ordre exactement* 1.

Exemple
Le schéma du point milieu :
le schéma du point milieu consiste à utiliser la méthode des rectangles, non plus à gauche (ni à droite d'ailleurs), mais au point milieu de chaque intervalle de la forme $[t_n, t_{n+1}]$. Il est donné par l'équation suivante

$$x_{n+1} = x_n + hf(t_n + \frac{h}{2}, x_n + \frac{h}{2}f(t_n, x_n)), \text{ pour } n = 0, ..., N-1, \text{ avec } x_0 \text{ donné.}$$

On montre que ce schéma est exactement d'ordre 2.
Le schéma de Heun (méthode des trapèzes) :
le schéma de Heun est donné par le système suivant

$$\begin{cases} \hat{x}_n &= x_n + hf(t_n, x_n), \\ x_{n+1} &= x_n + \dfrac{h}{2}\left(f(t_n, x_n) + f(t_{n+1}, \hat{x}_n)\right), \end{cases}$$

où $n = 0, ..., N-1$ et x_0 est donné.
On voit dans la deuxième équation que l'on estime la moyenne aux extrémités du segment $[t_n, t_{n+1}]$, ce qui correspond bien à l'approximation de l'aire sous l'intégrale par un trapèze.
On peut écrire ce système autrement de la façon suivante

$$x_{n+1} = x_n + \frac{h}{2}\left(f(t_n, x_n) + f(t_n + h, x_n + hf(t_n, x_n))\right),$$

où $n = 0, ..., N-1$ *et* x_0 *est donné.*
On montre que ce schéma est exactement d'ordre 2.

FIGURE 4.2 – Karl Heun (1859 – 1929), mathématicien allemand qui à qui l'on doit le schéma numérique qui porte son nom basé sur la méthode des trapèzes.

Remarque
Le schéma explicite à un pas est convergent d'ordre exactement p *quand il est d'ordre au moins* p *et qu'il n'est pas d'ordre au moins* $p+1$.

4.2.4 Les méthodes de Runge-Kutta

Certains schémas comme les schémas du point milieu et de Heun (méthode des trapèzes) permettent d'obtenir des convergences d'ordre plus élevé qu'Euler explicite.

1. Dans le cas du schéma du point milieu, la dérivée en l'instant intermédiaire entre t_n et t_{n+1} est estimée.

2. Dans le cas du schéma de Heun, la moyenne entre les estimations effectuées aux instant t_n et t_{n+1} est estimée.

La méthode du point milieu est due à Carl Runge, puis Martin Kutta a proposé diverses méthodes basées sur des moyennes à différents instants.
Cette idée a ensuite été généralisée pour construire ce que l'on appelle désormais les schémas de Runge-Kutta de n'importe quel ordre.

Définition 18 (SCHEMA DE RUNGE-KUTTA)

La structure générale d'un schéma de Runge-Kutta à s-stages explicite est donné par le système suivant

$$(RK) \begin{cases} X_i &= x_n + h\sum_{j=1}^{i-1} a_{ij} f(t_n + c_j h, X_j), \quad i = 1, ..., s \\ x_{n+1} &= x_n + h\sum_{i=1}^{s} b_i f(t_n + c_i h, X_i). \end{cases} \tag{4.17}$$

avec $n = 0, ..., N-1$ et x_0 donné. Par convention, nous notons $\sum_{i=1}^{0} ... = 0$

La donnée d'une méthode de Runge-Kutta explicite est donc la donnée de :

1. s instants compris entre t_n et t_{n+1} définis par $t_n + c_i h$, où $c_i \in [0, 1]$, pour $i = 1, ..., s$,
2. s poids b_i, $i = 1, ..., s$ permettant de faire la moyenne pondérée finale,
3. poids intermédiaires a_{ij} permettant de faire les estimations intermédiaires.

De façon usuelle, on représente alors une méthode de Runge-Kutta grâce au tableau de Butcher :

c_1	0	0	0	...	0	0
c_2	a_{21}	0	0	...	0	0
c_3	a_{31}	a_{32}	0	...	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
c_{s-1}	$a_{s-1,1}$	$a_{s-1,2}$	$a_{s-1,3}$	...	0	0
c_s	$a_{s,1}$	$a_{s,2}$	$a_{s,3}$	...	$a_{s,s-1}$	0
	b_1	b_2	b_3	...	b_{s_1}	b_s

FIGURE 4.3 – John Butcher (1933 –), mathématicien Néo-Zélandais, spécialisé dans les méthodes numériques de résolution d'équations différentielles. On lui doit le tableau qui porte son nom pour décrire les schémas de Runge-Kutta.

Exemple

1. *Le schéma représenté par*

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

est le schéma d'Euler explicite.

2. *Le schéma représenté par*

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

est le schéma du point milieu.

3. *Le schéma représenté par*

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

est le schéma de Heun. Ce dernier schéma est également appelé parfois le schéma de Runge-Kutta 2, même si le schéma du point milieu est lui aussi un schéma d'ordre 2 de type Runge-Kutta.

4. *Quand on parle "du" schéma de Runge Kutta, ou encore RK-4 on désigne le schéma à 4 stages suivant*

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

La question que l'on peut se poser est pourquoi ce choix des poids et des instants intermédiaires ?
La réponse est assez simple : pour avoir les propriétés de stabilité et de consistance d'ordre élevé.
Les schémas de Runge-Kutta explicites rentrent en effet dans le cadre général des schémas explicites à 1 pas puisqu'on peut les écrire sous la forme

$$\left\{ \begin{array}{lcl} x_{n+1} & = & x_n + h\phi(t_n, x_n, h), \quad \text{pour} \quad n = 0, ..., N-1, \\ x(t_0) & = & x_0. \end{array} \right.$$

Afin d'étudier leurs propriétés, on peut donc utiliser la théorie générale des schémas explicites à un pas. Nous avons alors le résultat suivant.

Proposition 9 (CRITÈRES DE CONSISTANCE DE RUNGE-KUTTA)

On considère un schéma de Runge-Kutta sous sa forme générale (RK) donné par (4.17).

1. Ce schéma est consistant d'ordre au moins 1 avec le problème de Cauchy $(\mathscr{C})$ si et seulement si
$$\sum_{i=1}^{s} b_i = 1.$$
2. Ce schéma est consistant d'ordre au moins 2 avec le problème de Cauchy $(\mathscr{C})$ si et seulement si
$$\sum_{i=1}^{s} b_i c_i = \frac{1}{2} \quad \text{et} \quad \sum_{i=1}^{s} b_i \sum_{j=1}^{i-1} a_{ij} = \frac{1}{2}.$$

Remarque

1. *La condition* $\sum_{i=1}^{s} b_i = 1$ *reflète bien le fait que les* b_i *sont bien des coefficients d'une moyenne pondérée.*
2. *On obtient des relations du même type pour tous les ordres. Et on peut ainsi montrer que le schéma RK-4 est consistant d'ordre 4.*

Nous avons également un critère de stabilité pour les schémas de Runge-Kutta.

Proposition 10 (CRITÈRES DE STABILITÉ DE RUNGE-KUTTA)

Tous les schémas du type Runge-Kutta explicites sont stables dès que f est continue et globalement lipschitzienne par rapport à sa variable x sur l'intervalle $[t_0, t_0 + T]$ étudié.

Remarque

En particulier le schéma RK-4 est convergent d'ordre 4. C'est le schéma le plus communément utilisé. Il est à la base du solveur ODE45 de Matlab. Le solveur ODE45 est à pas de temps adaptatif (à chaque instant, le pas de temps h_n *n'est pas prédéfini, on le détermine pour que l'erreur soit inférieure à une précision donnée).*
Pour ça, on fait une itération de RK-4 que l'on compare avec une itération d'un schéma de Runge-Kutta d'ordre 5 (dû à Dormand et Prince en 1980).
La différence $|RK-4 - RK5|$ *donne alors une estimation de l'erreur par RK-4. On fait donc comme si RK-5 donnait la solution exacte. Si l'erreur est supérieure à la précision que l'on a choisie préalablement, on diminue le pas de temps* h_n.

Les méthodes de Runge-Kutta sont donc basées sur l'introduction d'instants intermédiaires entre t_n et t_{n+1} pour augmenter l'ordre de la consistance.
Un autre moyen d'augmenter la précision du calcul consiste à faire intervenir plusieurs pas de temps lors du calcul de x_{n+1}. C'est ce qu'on appelle les schémas à pas multiples ou multipas explicites. Ces schémas sont de la forme

$$x_{n+1} = x_n + h\phi(t_n, x_n, h_n, t_{n-1}, x_{n-1}, h_{n-1}, t_{n-2}, x_{n-2}, h_{n-2}, ...).$$

Ils sortent évidemment du cadre des schémas à un pas, mais on peut pour elles aussi définir la convergence que l'on décompose en stabilité et consistance. Le schéma explicite multipas le plus couramment utilisé est celui d'Adams-Bashforth (dû à John Adams qui l'utilisa en 1977 pour résoudre une EDO décrivant un problème de capillarité introduit par Francis Bashforth en 1883). Les méthodes multipas sont préférables aux méthodes de Runge-Kutta lorsque la fonction est coûteuse (en temps de calcul) à évaluer.
Mais toutes ces méthodes, que ce soit Runge-Kutta explicite ou multipas explicites, ont un point commun : elles ne sont performantes que pour les problèmes non-raides (en anglais Non STIFF).

4.3 Problèmes raides et schémas implicites

Lorsque l'on applique un schéma explicite à un pas adaptatif (comme ODE45) à certains problèmes dont la solution présente des problèmes raides. On obtient un effet en accordéon.
Le programme doit placer beaucoup de points dans les problèmes raides pour obtenir une précision requise.
Si le problème est trop raide, le nombre de points nécessaires devient trop important, le temps de calcul devient trop long et le schéma est inefficace.
La question est alors la suivante : comment caractérise-t-on les problèmes raides ? Ce sont les zones où la constante de Lipschitz L est trop grande.
Ceci vient du fait que l'erreur d'un schéma explicite stable et d'ordre au moins p est

$$\max_{1\leq n\leq N} \|x(t_n) - x_n\| \leq MKh^p,$$

où K est la constante de consistance, et $M = e^{LT}$ la constante de stabilité.
Ainsi, lorsque L devient grande, M devient grande exponentiellement, et donc la valeur prise par h doit être très petite pour garder une précision initialement donnée.
A cause de ça, on introduit une autre notion de stabilité qui teste le schéma sur les différentes valeurs possibles de L.

4.3.1 Test Linéaire Standard

Cette notion est basée sur ce que l'on appelle l'équation Test Linéaire Standard (ou TLS).
Ce test est défini par le système suivant

$$(TLS) \quad \begin{cases} x'(t) &= -Lx(t), \\ x(0) &= x_0, \end{cases} \tag{4.18}$$

où $L > 0$.

1. La solution exacte est
$$x(t) = x_0 e^{-Lt}.$$
2. Elle vaut x_0 pour $t = 0$ et tend vers 0 quand t tend vers l'infini.
3. D'un autre côté, plus L est grand, plus $x(t)$ tend vers 0 de façon raide.
4. Enfin, nous remarquons que la constante de Lipschitz du problème (TLS) est L. En effet, si l'on pose $f(t,x) = -Lx$ pour $t \in \mathbb{R}_+$ et $x \in \mathbb{R}$, alors pour tous $t \in \mathbb{R}_+$ et pour tous $x, y \in \mathbb{R}$,
$$|f(t,x) - f(t,y)| = |-Lx + Ly| = L|x - y|.$$

Définition 19 (SCHEMA A-STABLE)

On dit qu'un schéma est A-stable si et seulement si ce schéma appliqué au problème (TLS) donne une solution x_n vérifiant $\lim_{n\to+\infty} x_n = 0$ quel que soit $L > 0$, et quel que soit le pas de temps constant h.
On dit aussi que le schéma est **inconditionnellement stable**.

Remarque

1. *Un schéma est A-stable est donc un schéma qui peut traiter les problèmes de n'importe quelle raideur sans condition sur le pas de temps.*
2. *La stabilité d'un schéma décrit la façon dont les erreurs s'accumulent sur un intervalle de temps borné $[t_0, t_0 + T]$ tandis que la A-stabilité décrit la façon dont les erreurs faussent le comportement de la solution pour $t \to +\infty$.*

Exemple *On peut montrer que le schéma d'Euler explicite n'est pas A-stable.*

De même on peut montrer que tous les schémas de Runge-Kutta explicites ont des conditions sur le pas de temps dépendant de la raideur du problème. Ces schémas ne sont donc pas A-stables.
Pour cette raison, on a introduit d'autres schémas qui cette fois-ci seront A-stables. En contre partie, ils sont plus difficiles à analyser et à coder. Le plus simple de ces schémas est le schéma d'Euler implicite.

4.3.2 Schéma d'Euler implicite (ou rétrograde)

Le schéma d'Euler implicite est construit de la même manière que le schéma d'Euler explicite : par la méthode des rectangles. Mais au lieu de considérer le rectangle à gauche, on considère le rectangle à droite cette fois-ci.
Le schéma s'écrit alors de la façon suivante :

$$(\mathscr{E}_r) \quad \begin{cases} x_{n+1} &= x_n + hf(t_n + h, x_{n+1}), \quad \text{pour} \quad n = 0, ..., N-1, \\ x(t_0) &= x_0. \end{cases}$$

Ce schéma est appelé implicite parce qu'il ya du x_{n+1} dans le second membre. A chaque itération, le schéma définit donc une équation qu'il faut résoudre pour trouver x_{n+1} de façon implicite.

Proposition 11 (SCHEMA EULER IMPLICITE A-STABLE)

Les schéma d'Euler implicite est A-stable.

4.3.3 Shémas de Runge-Kutta implicite

Définition 20 (SCHEMAS RK IMPLICITES)

Un schéma de Runge-Kutta implicite à S-stages est décrit par le tableau de Butcher suivant

$$\begin{array}{c|cccccc} c_1 & a_{11} & a_{12} & a_{13} & \dots & a_{1,s-1} & a_{1s} \\ c_2 & a_{21} & a_{22} & a_{23} & \dots & a_{2,s-1} & a_{2,s} \\ c_3 & a_{31} & a_{32} & a_{33} & \dots & a_{3,s-1} & a_{3,s} \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ c_{s-1} & a_{s-1,1} & a_{s-1,2} & a_{s-1,3} & \dots & a_{s-1,s-1} & a_{s-1,s} \\ c_s & a_{s,1} & a_{s,2} & a_{s,3} & \dots & a_{s,s-1} & a_{s,s} \\ \hline & b_1 & b_2 & b_3 & \dots & b_{s_1} & b_s \end{array}$$

ou par la forme développé suivante

$$(RK_r) \begin{cases} X_i &= x_n + h\sum_{j=1}^{s} a_{ij} f(t_n + c_j h, X_j), \quad i = 1, ..., s \\ x_{n+1} &= x_n + h\sum_{i=1}^{s} b_i f(t_n + c_i h, X_i). \end{cases} \tag{4.19}$$

avec $n = 0, ..., N-1$ et x_0 donné.

On pourrait montrer que tous les schémas de implicite à S-stages sont A-stables.
Le cas explicite est d'ailleurs un cas particulier du cas implicite où le tableau de Butcher contient une matrice "strictement triangulaire inférieure".

Exemple
Le schéma représenté par

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

est le schéma d'Euler implicite.

La convergence pour les schémas de Runge-Kutta implicites se définit comme pour les schémas explicites.

$$\lim_{h\to 0}\max_{1\leq n\leq N}\|x(t_n)-x_n\|=0.$$

La notion de consistance et d'ordre de consistance sont légèrement différentes par rapport au cas explicite, mais sont assez similaires quand même.
Les conditions sur l'ordre de consistance sont les mêmes à savoir :

1. Ce schéma est consistant d'ordre au moins 1 avec le problème de Cauchy ($\mathscr{C}$) si et seulement si
$$\sum_{i=1}^{s} b_i = 1.$$
2. Ce schéma est consistant d'ordre au moins 2 avec le problème de Cauchy ($\mathscr{C}$) si et seulement si
$$\sum_{i=1}^{s} b_i c_i = \frac{1}{2} \quad \text{et} \quad \sum_{i=1}^{s} b_i \sum_{j=1}^{s} a_{ij} = \frac{1}{2}.$$
mais la matrice constituée des a_{ij} n'est plus strictement triangulaire inférieure.

La convergence est toujours vraie pour les méthodes de Runge-Kutta implicites dès qu'elles sont consistantes et que f vérifie les conditions du théorème de Cauchy-Lipschitz global.

4.3.4 Résolution des itérations des schémas implicites

On a vu que les schémas implicites définissent à chaque itération une équation qu'il faut résoudre pour trouver x_{n+1}.

Exemple *Pour Euler implicite c'est*

$$x_{n+1} = x_n + hf(t_{n+1}, x_{n+1})$$

C'est à dire que x_{n+1} est un point fixe de la fonction G définie pour tout $x \in \mathbb{R}$ par

$$G(x) = x_n + hf(t_{n+1}, x).$$

il existe une théorie classique de l'existence et l'unicité des points fixes des fonctions (c'est le chapitre 3 de ce cours). Elle donne une condition suffisante d'existence et d'unicité de x_{n+1}. Mais pour un problème raide, cette condition n'a que très peu d'intérêt car il faut par exemple pour le problème de Cauchy, si f est L-lipschitzienne, que $h < 1/L$ pour avoir l'existence et l'unicité du point fixe x_{n+1}.
Autrement dit, il faut une condition sur la valeur du pas de temps.
Pour des problèmes raides, cette condition est donc peu utile.
D'autres conditions sont actuellement développées qui donnent l'existence et l'unicité de x_{n+1} avec une condition moins restrictive sur h. Mais

1. il y a quand même une condition sur h,
2. ça ne marche pas pour tous les problèmes raides.

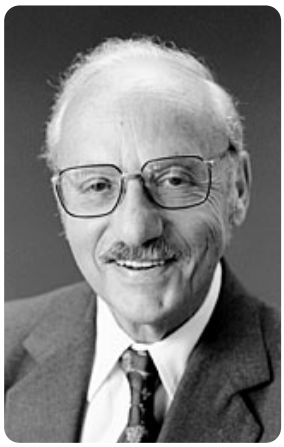
toutefois, en pratique, la solution x_{n+1} est en général bien définie pour les schémas de Runge-Kutta implicites, même pour des problèmes très raides et de grands pas de temps.
Enfin, quand x_{n+1} est bien définie, la méthode la mieux adaptée pour la calculer est la méthode de Newton que nous avons vus au chapitre précédent.

Chapitre 5

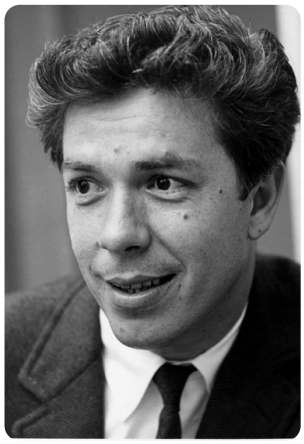
Optimisation : méthode du gradient

La liberté, c'est la faculté de choisir ses contraintes.

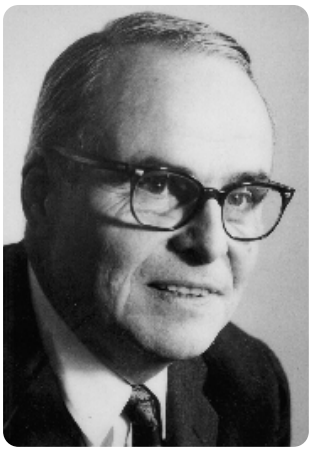
Jean-Louis Barrault



(a) George Bernard Dantzig (1914 – 2005), mathématicien américain, a contribué à l'étude de l'optimisation en créant l'algorithme du simplexe en optimisation linéaire.



(b) Harold William Kuhn (1925 – 2014), mathématicien et économiste américain, qui a publié avec Albert Tucker plusieurs résultats sur l'optimisation avec contrainte.



(c) Albert William Tucker (1905 – 1995) mathématicien américain d'origine canadienne. On lui doit d'importants résultats, notamment en topologie et théorie des jeux et optimisation non linéaire.

FIGURE 5.1 – Quelques mathématiciens célèbres liés à l'étude des schémas numériques pour l'optimisation.

Sommaire

Il est possible de trouver des cours et des exercices dans de nombreux ouvrages disponibles à la bibliothèque. Celui dont je me suis principalement inspiré est de Filbet [4]. L'objectif de ce chapitre est de chercher des méthodes efficaces pour minimiser ou maximiser de fonctions sachant que l'inconnue satisfait ou non des contraintes de tous ordres (physiques, économiques, écologiques par exemple). Commençons par un exemple simple pour fixer les idées. Cet exemple est tiré de [4].

5.1 L'exemple du sac à dos

L'exemple qui suit est appelé "problème du sac à dos" par Filbet dans [4].
On considère un randonneur qui souhaite emporter dans son sac n objets différents. Chaque objet i possède un poids connu $p_i > 0$, un volume $v_i > 0$ et une utilité $u_i > 0$. Plus l'utilité de l'objet est grande, plus il est important de le considérer.
Mais bien évidemment, le porteur est limité par une charge maximale $p > 0$ que peut supporter son sac à dos, ou qu'il peut lui même supporter, et par le volume $v > 0$ que contient ce même sac.
Le problème est donc de bien sélectionner les objets à emporter.
Nous posons le vecteur $x = (x_1, x_2, ..., x_m)^T \in \mathbb{R}^m$ tel que pour $i = 1, ..., m$, $x_i = 1$ si le randonneur emporte l'objet et $x_i = 0$ s'il décide de ne pas le prendre.
Les deux contraintes imposent :

1. la somme des poids de tous les objets pris doit être la plus proche possible de p en étant inférieure ou égale, c'est à dire

$$\sum_{i=1}^{m} x_i p_i \leq p,$$

2. la somme des volumes doit être inférieure ou égale à v, en étant la plus proche possible de v, c'est à dire

$$\sum_{i=1}^{m} x_i v_i \leq v.$$

N'oublions pas l'utilité des objets. Il faut donner la priorité aux objets essentiels comme les boissons ou la nourriture. Par conséquent, il faut maximiser cette fonction d'utilité définie sur $\mathbb{R}^m$ par

$$f(x) = \sum_{i=1}^{m} x_i u_i.$$

Le problème se résume alors à maximiser la fonction f sous les contraintes. La fonction f est linéaire, nous appelons donc ce problème : problème de programmation linéaire sous contraintes.
De façon générale dans ce chapitre, nous considérons E un $\mathbb{R}$-espace vectoriel et une fonction $f : K \subset E \to \mathbb{R}$. Nous recherchons à minimiser (maximiser revient à minimiser $-f$), c'est à dire trouver les minima $\overline{x}$ tels que

$$\begin{cases} \overline{x} \in K, \quad \text{tel que,} \\ \\ f(\overline{x}) = \min\limits_{x \in K} f(x). \end{cases}$$

Commençons par les problèmes linéaires (comme dans l'exemple du sac à dos).

5.2 Programmation linéaire

Soit $x \in \mathbb{R}^m$. Par abus, nous écrirons $x \geq 0$ qui signifiera que toutes les composantes x_i du vecteur x, $i = 1, ..., m$ sont positives ($x_i \geq 0$).

5.2.1 Définition

Définition 1 (PROGRAMMATION LINÉAIRE)

Nous appelons problème de programmation linéaire sous la forme standard, tout problème qui s'écrit

$$(PL) \quad \begin{cases} \overline{x} \in \mathbb{R}^m, \quad A\overline{x} = b, \overline{x} \geq 0 \\ \\ f(\overline{x}) = c^T \overline{x} = \min\limits_{\substack{Ax=b \\ x \geq 0}} c^T x. \end{cases} \tag{5.1}$$

où $A \in \mathscr{M}_{n,m}(\mathbb{R})$ (matrice à n lignes et m colonnes), $b \in \mathbb{R}^n$ et $c \in \mathbb{R}^m$ sont donnés.

5.2.2 Résolution du problème de programmation linéaire

On considère un problème de programmation linéaire sous forme standard (PL) défini par (5.1). Regardons de plus près l'ensemble qui définit les contraintes.
Nous supposons $A := (a_1, ..., a_m) \in \mathscr{M}_{n,m}(\mathbb{R})$ de rang $n \leq m$. C'est à dire qu'il existe une famille de n vecteurs colonnes parmi $\{a_1, ..., a_m\}$ formant une famille libre (et donc une base) de $\mathbb{R}^n$.

Définition 2 (ENSEMBLE ADMISSIBLE ET SOLUTION OPTIMALE)

Nous appelons ensemble admissible pour le problème de programmation linéaire sous forme standard (PL) défini par (5.1), l'ensemble

$$M = \{x \in \mathbb{R}^m, Ax = b, x \geq 0\}.$$

Lorsque $x \in M$, nous disons que x est une solution admissible et lorsque $x \in M$ et qu'elle satisfait (PL) , nous disons que x est une solution optimale.

Remarque
L'ensemble admissible M est l'intersection du sous-espace affine de dimension n donné par

$$\{x \in \mathbb{R}^m, Ax = b\},$$

et du cône convexe fermé défini par

$$\{x \in \mathbb{R}^m, x \geq 0\}.$$

Définition 3 (COMBINAISON CONVEXE)

Soient $x_1, ..., x_k$ des vecteurs de $\mathbb{R}^m$. Un vecteur x de $\mathbb{R}^m$ est dit combinaison convexe des vecteurs $(x_i)_{1 \leq k \leq k}$ s'il peut s'écrire sous la forme

$$x = \sum_{i=1}^{k} \lambda_i x_i,$$

où les λ_i sont des réels tels que $0 \leq \lambda_i \leq 1$ et $\sum_{i=1}^{k} \lambda_i = 1$.

Définition 4 (POINT EXTRÉMAL)

Soient M une partie convexe et fermée de $\mathbb{R}^m$ et x un point (vecteur) de M. Nous disons que le point x est extrémal s'il ne peut pas s'écrire comme combinaison convexe de deux éléments distincts de M, c'est à dire que l'égalité

$$x = ty + (1 - t)z,$$

avec y et $z \in M$, $0 < t < 1$ n'a pas d'autre solution que $y = z = x$.
Lorsque M est un polytope convexe, ses points extrémaux sont également appelés sommets.

Remarque
Rappelons qu'un polytope dans $\mathbb{R}^n$ est l'enveloppe convexe d'un nombre fini de points de $\mathbb{R}^n$.

La suite de cette section est consacrée à la description de méthodes pour rechercher ces ponts extrémaux.
Introduisons pour cela les notions de variables de bases et hors bases.

Définition 5 (BASE ASSOCIÉE)

Soit $A \in \mathscr{M}_{n,m}(\mathbb{R})$ de rang $n \leq m$. Nous appelons base associée au problème de programmation linéaire (PL) donné par (5.1) toute matrice carrée $B \in \mathscr{M}_{n,n}(\mathbb{R})$ inversible issue de A.

Méthode pour la construction d'une base :
la construction d'une base revient à rechercher n colonnes de A linéairement indépendantes. On note $B \in \mathscr{M}_{n,n}(\mathbb{R})$ la matrice carrée formée par ces n colonnes. On note $P \in \mathscr{M}_{n,m}(\mathbb{R})$ une matrice de permutation (c'est à dire que $P^T P = Id_m$) et $N \in \mathscr{M}_{n,m-n}(\mathbb{R})$ telles que

$$AP = (B|N).$$

Il est alors possible de partitionner n'importe quel point $x \in M$ de la façon suivante :

$$P^T x = \begin{pmatrix} x_B \\ x_N \end{pmatrix},$$

où $x_B \geq 0$ est appelée variable de base et $x_N \geq 0$ variable hors-base.
Par ce qui précède nous avons

$$Ax = APP^T x = (B|N)P^T x.$$

Ainsi

$$Bx_B + Nx_N = b,$$

Finalement, comme B est inversible cela donne

$$x_B = B^{-1}(b - Nx_N). \tag{5.2}$$

Le problème de programmation linéaire (PL) peut donc être reformulé en décomposant le vecteur $c \in \mathbb{R}^m$ comme

$$P^T c = \begin{pmatrix} c_B \\ c_N \end{pmatrix}, \tag{5.3}$$

et donc pour tout $x \in M$

$$f(x) = c_B^T x_B + c_N^T x_N.$$

D'après la définition de x_B donné par (5.2) nous obtenons

$$f(x) = c_B^T B^{-1} b + (c_N^T - c_B^T B^{-1} N) x_N.$$

D'autre part, si $B^{-1}b$ est positif ou nul, le point $x \in \mathbb{R}^m$ formé par $x = (x_B, x_N)^T$ avec $x_N = 0_{\mathbb{R}^{m-n}}$ et $x_B = B - 1b \geq 0$ vérifie bien $Ax = b$ et $x \geq 0$, ce qui signifie que $x \in M$.
Ceci permet de définir ce que l'on appelle une solution de base.

Définition 6 (SOLUTION DE BASE)

Un vecteur $x \in \mathbb{R}^m$ est appelé solution de base s'il existe une matrice de permutation P telle que

$$AP = (B|N) \text{ et } P^T x = (x_B, x_N)^T,$$

avec $x_B = B^{-1}b$ et $x_N = 0_{\mathbb{R}^{m-n}}$.
Si de plus, le vecteur $x_B = B^{-1}b \geq 0$, le point x est une solution de base admissible.

Théorème 1 (SOLUTION DE BASE ET ADMISSIBLE)

Considérons le problème programmation linéaire (PL) où A est de rang $n \leq m$.

1. S'il existe une solution admissible de (PL) alors il existe une solution admissible de base.
2. S'il existe une solution admissible optimale de (PL) alors il existe une solution optimale admissible de base.

Remarque

Ce théorème signifie qu'il suffit de rechercher les solutions optimales de (PL) parmi les solutions de base admissibles, c'est à dire sous la forme $P^T x = (B^{-1}b, 0)$ avec $B^{-1}b \geq 0$.

Théorème 2 (LOCALISATION D'UN SOMMET)

Soient M l'ensemble des états admissibles du problème de programmation linéaire (PL) et $x \in M$. Les deux propriétés suivantes sont équivalentes :

1. le point x est un sommet de M,
2. le point x est une solution de base admissible.

Il existe alors deux méthodes pour calculer une solution du programmation linéaire :

1. **la méthode d'énumération**.
   Elle s'appuie sur les deux théorèmes précédents. Le minimum de la fonction f sur l'ensemble admissible M est atteint en un sommet. Autrement dit, nous cherchons les solutions de base admissibles. La méthode d'énumération (la plus naturelle intuitivement) consiste à parcourir l'ensemble des sommets du polytope M et d'évaluer la fonction en ces points. Puis de choisir celui qui possède la valeur minimale.
   Mais cette méthode est coûteuse en calculs.
   Une autre méthode proposée est la suivante.
2. **Méthode du simplexe** :
   Plutôt que chercher toutes les solutions de bases admissibles, le but ici est d'identifier les solutions de base optimales.
   Procédons de la façon suivante.
   Étant donnée une base B formée par n colonnes de la matrice A, la fonction coût

$$f(x) = c^T x$$

   s'écrit $f(x) = c^T x = c^T P P^T x = (P^T c)^T P^T x$.
   En décomposant le vecteur $c \in \mathbb{R}^m$ de la façon suivante :

$$P^T c = \begin{pmatrix} c_B \\ c_N \end{pmatrix},$$

comme précédemment (voir (5.3)). Il vient alors pour $x \in M$,

$$x_B = B^{-1}(b - Nx_N) \text{ et } f(x) = c_B^T x_B + c_N^T x_N.$$

Ce qui nous donne

$$f(x) = c_B^T B^{-1} b + (c_N^T - c_B^T B^{-1} N) x_N.$$

Nous avons alors la proposition suivante

Proposition 1 (SOLUTION DE BASE ADMISSIBLE ET VECTEUR DES COUTS)

Supposons que la matrice B issue de A soit inversible et telle que $B^{-1}b > 0$. Alors le point $\overline{x}$ donné par

$$P^T \overline{x} = \begin{pmatrix} B^{-1}b \\ 0_{\mathbb{R}^{m-n}} \end{pmatrix}$$

est une solution de base admissible optimale si et seulement si le vecteur des coûts $\overline{c_N}$ vérifie

$$\overline{c_N}^T := c_N^T - c_B^T B^{-1} N \geq 0.$$

La méthode du simplexe prend appuie sur cette proposition pour parcourir les sommets du polytope plus efficacement que la méthode d'énumération.
Si l'on désigne par $E(M)$, l'ensemble des sommets de M, l'algorithme permettant d'éviter de parcourir tous les sommets est le suivant :

1. on détermine un sommet x de M,
2. on applique à x le critère du coût réduit permettant de savoir s'il existe un point z de M tel que $f(z) > f(x)$,
   (a) si ce n'est pas le cas, la fonction f atteint son minimum en x et le problème est résolu,
   (b) si c'est le cas, nous déterminons un autre sommet x_0 de M voisin de x tel que $f(x_0) < f(x)$,
3. on continue en remplaçant x par x_0.

5.2.3 Introduction à la dualité

Considérons le problème de programmation linéaire sous la forme standard

$$(PL) \begin{cases} \overline{x} \in \mathbb{R}^m, \quad A\overline{x} = b, \overline{x} \geq 0 \\ \\ f(\overline{x}) = c^T \overline{x} = \min\limits_{\substack{Ax=b \\ x \geq 0}} c^T x. \end{cases}$$

où $A \in \mathscr{M}_{n,m}(\mathbb{R})$ (matrice à n lignes et m colonnes), $b \in \mathbb{R}^n$ et $c \in \mathbb{R}^m$ sont donnés.
Nous disons que x est la variable primale et $\overline{x} \in \mathbb{R}^m$ est la solution du problème primal.
Nous introduisons les contraintes dans une fonction coût par le lagrangien $\mathscr{L} : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ définie pour tout $(x, q)\mathbb{R}^m \times \mathbb{R}^n$ par

$$\begin{aligned} \mathscr{L}(x,q) &= c^T x + q^T(b - Ax), \\ &= (c - A^T q)^T x + q^T b. \end{aligned}$$

Il y a simplement un ajout de nouvelles variables q pour tenir compte des contraintes, et nous cherchons à minimiser le lagrangien $\mathscr{L}$.
Ceci peut s'expliquer de la façon suivante : lorsque la contrainte $Ax - b = 0$ n'est pas vérifiée, le terme $q^T(b - Ax)$ est susceptible d'augmenter.
Par conséquent, pour chaque valeur $q \in \mathbb{R}^n$, l'objectif est de minimiser le lagrangien $\mathscr{L}$. Ceci induit une nouvelle fonction $\eth(q) : \mathbb{R}^n \to \mathbb{R}$ que nous appelons ici fonction duale et qui est définie pour tout $q \in \mathbb{R}^n$ par

$$\mathscr{G}(q) = \min_{\substack{x \in \mathbb{R}^m \\ x \geq 0}} \mathscr{L}(x, q).$$

Nous disons alors q est la variable duale. On obtient alors

$$\mathscr{G}(q) = \begin{cases} q^T b, & \text{si } A^T q - c \leq 0, \\ \\ -\infty, & \text{sinon.} \end{cases} \tag{5.4}$$

Comme la solution du problème primal $\overline{x} \in \mathbb{R}^m$ vérifie exactement les contraintes, il vient

$$\mathscr{G}(q) \leq \mathscr{L}(\overline{x}, q) = c^T \overline{x}.$$

On considère maintenant le problème du point de vue des contraintes. Nous cherchons $q \in \mathbb{R}^n$ tel que la fonction coût soit la plus proche possible du coût optimal $c^T \overline{x}$. C'est ce qui est appelé le problème dual.

Définition 7 (PROBLÈME DUAL)

Bibliographie

[1] S. BENZONI-GAVAGE, *Calcul différentiel et équations différentielles*, Dunod, 2014. 39

[2] R.-L. BURDEN ET J.-D. FAIRES, *Numerical Analysis*, Brooks Cole, 2001. 3, 16, 24, 39

[3] J.-P. DEMAILLY, *Analyse numérique et équations différentielles*, PUG Grenoble, 1996. 39

[4] F. FILBET, *Analyse numérique : algorithme et étude mathématique*, Dunod, 2009. 3, 16, 24, 39, 65

[5] R. QUARTERONI, A. ET SACCO ET F. SALERI, *Méthodes numériques pour le calcul scientifique. Programmes en MATLAB*, Springer, 2005. 3, 16, 24, 39

[6] M. SCHATZMAN, *Analyse numérique*, InterEditions, Paris, 1991. 3, 16, 24, 39