Table des matières

Introduction générale

En mathématiques, il existe des problèmes qui ne peuvent pas être résolu analytiquement, par exemple, les équations algébriques de degré ≥ 5. Pour la résolution de ce type d'équations il a été prouver qu'il n'existe pas de formule par radicaux.

Aussi, dans certains cas les solutions analytiques existent mais sont complètement inefficaces à mettre en oeuvre en pratique, comme la résolution d'un système linéaire $Ax = b$ dont la solution est $x = A^{-1}b$ où le calcul de A^{-1} devient rapidement intolérable lorsque la taille de la matrice augmente. Donc souvent pour résoudre un problème mathématique on fait appel à l'analyse numérique.

L'*analyse numérique* comprend deux mots : l'analyse qui fait référence aux mathématiques et le mot numérique qui fait référence au traitement informatique.

En d'autres termes c'est l'élaboration des méthodes de calcul mathématiques adaptées aux traitement par ordinateur qui ont pour but la résolution des problèmes concrets qui se posent dans différentes disciplines : physique, économie, ingénierie, etc. Ces problèmes doivent être bien posés c'est-à-dire que la solution existe, unique et dépend continûment des données du système.

Chapitre 1

Rappels sur quelques méthodes numériques

1.1 Résolution des équations non linéaires

Calculer les racines d'une équation $f(x) = 0$ est un problème que l'on rencontre très souvent en calcul scientifique, comme par exemple déterminer le point de fonctionnement d'une diode d'après sa caractéristique, la concentration d'une espèce chimique dans un mélange réactionnel ou la fréquence de coupure d'un filtre électrique.

Il est rare qu'on puisse écrire une solution analytique de l'équation $f(x) = 0$, cela ce produira que pour les polynômes de degré inférieur ou égal à 4 ou pour quelques fonctions simples. En conséquence, toutes les méthodes générales de recherche de racine sont des méthodes itératives ; partant d'une solution approchée, on obtient une suite d'approximations de plus en plus précises. On doit étudié alors la convergence et la vitesse de convergence de la méthode ensuite définir un critère d'arrêt des itérations.

Comme exemple d'équation non linéaire on a l'équation d'état d'un gaz. On veut déterminer le volume V occupé par un gaz de température T et de pression p. L'équation d'état (c'est-à-dire l'équation qui lie p, V et T) est

$$\left[p + a\left(\frac{N}{V}\right)^2\right](V - Nb) = kNT,$$

où a et b sont deux coefficients dépendants de la nature du Gaz, N est le nombre de molécules contenues dans le volume V et k est la constante de Boltzmann. Il faut donc résoudre une équation non linéaire d'inconnue V. Pour la résolution de ce type d'équation, les méthodes analytiques sont limitées à certaines formes algébriques $(a_n x^n + a_{n-1}x^{n-1} + ... + a_0)$, $n < 5$ ou formes particulières. Par conséquent pour les autres formes il faut utiliser les méthodes numériques pour trouver ou approcher les racines.

Ces méthodes là consistent d'abord à :

1. Localiser le (ou les) zéro(s) de f en procédant à l'étude du graphe de f, puis utiliser le théorème des valeurs intermédiaires afin de trouver un intervalle qui contient une unique racine.

   Théorème 1 *(Théorème des valeurs intermédiaires)*
   Si f une fonction réelle à variable réelle, définie et continue dans un intervalle $[a, b]$ de $\mathbb{R}$ et $f(a) \times f(b) < 0$ alors $\exists \alpha \in \,]\, a, b \,[$ tel que $f(\alpha) = 0$.
   Si de plus f est strictement monotone dans $[a, b]$ alors α est unique dans $[a, b]$.

2. Trouver la solution en utilisant l'une des méthodes suivantes.

1.1.1 Méthode de Dichotomie

La dichotomie est un mot grec qui signifie :

tomie : vient de couper
dicho : $\cdots$ en deux.

On considère un intervalle $[a, b]$ et une fonction f de $[a, b]$ dans $\mathbb{R}$. On suppose que $f(a) \times f(b) < 0$ et que l'équation $f(x) = 0$, admet une unique solution $\alpha \in [a, b]$.

La méthode de dichotomie consiste à construire une suite (x_n) qui converge vers α de la manière suivante :

Soit c_0 le milieu de $[a, b]$. La racine se trouve alors dans l'un des deux intervalles $[a, c_0]$ ou $[c_0, b]$ où bien elle est égale à c_0.

— Si $f(a) \times f(c_0) < 0$, alors $\alpha \in]\, a, c_0\, [$. On pose $a_1 = a$, $b_1 = c_0$.
— Si $f(a) \times f(c_0) = 0$, alors $\alpha = c_0$.
— Si $f(a) \times f(c_0) > 0$, alors $\alpha \in]\, c_0, b\, [$. On pose $a_1 = c_0$, $b_1 = b$.

On prend alors c_1 le milieu de $[a_1, b_1]$. On construit ainsi une suite

$$c_0 = \frac{a+b}{2}, c_1 = \frac{a_1+b_1}{2}, \cdots, c_n = \frac{a_n+b_n}{2}$$

telle que $|a_n - b_n| < \epsilon$ avec $\epsilon > 0$.

Convergence de la méthode

Théorème 2 *Soit f une fonction continue sur $[a, b]$, vérifiant $f(a) \times f(b) < 0$ et soit $\alpha \in [a, b]$ l'unique solution de $f(x) = 0$. Si l'algorithme de dichotomie arrive jusqu'à l'étape n alors on a l'estimation :*

$$|\alpha - c_n| \leq \frac{b-a}{2^{n+1}}$$

Par conséquent la suite $(c_n)_{n\in\mathbb{N}}$ converge vers α. C'est ainsi vrai si $c_n = \alpha$ [?].

test d'arrêt

Pour que la valeur de c_n de la suite de la $n^{\text{ième}}$ itération soit une valeur approchée de α à $\epsilon > 0$ près. Il suffit que n vérifie :

$$\frac{b-a}{2^{n+1}} \leq \epsilon$$

On a alors

$$|\alpha - c_n| \leq \frac{b-a}{2^{n+1}} \leq \epsilon$$

Ce qui permet de calculer à l'avance le nombre nécessaire $n \in \mathbb{N}$ d'itérations assurant la précision ϵ.

$$\frac{b-a}{2^{n+1}} \leq \epsilon \iff \frac{b-a}{\epsilon} \leq 2^{n+1} \iff n \geq \frac{\ln\left(\frac{b-a}{\epsilon}\right)}{\ln(2)} - 1$$

1.1.2 Méthode de point fixe

Le principe de cette méthode consiste à transformer l'équation $f(x) = 0$ en une équation équivalente $g(x) = x$ où g est une fonction bien choisie. Ceci est toujours possible en posant $g(x) = x - f(x)$. Le point α est alors un point fixe de g.
Approcher les zéros de f revient à approcher les points fixe de g. Le choix de la fonction g est motivé par les exigences du théorème du point fixe.

Convergence de la méthode

Théorème 3 *(**Théorème du point fixe**) (condition suffisante de convergence globale)*
Soit g une fonction de classe C^1 sur un intervalle $[a, b] \subset \mathbb{R}$ telle que :

1. $g(x) \in [a, b], \forall x \in [a, b]$ *(on dit que l'intervalle $[a, b]$ est stable par $g(x)$. Si on écrit $I = [a, b]$, alors $g(I) \subset I$.*
2. $\exists k \in \mathbb{R}, 0 < k < 1$ *tel que* $|g'(x)| \leq k < 1, k = \max_{x \in [a,b]} |g'(x)|$ *On dit que g est strictement contractante*

Alors

— *g admet un unique point fixe dans $I = [a, b]$.*
— *Pour tout $x_0 \in [a, b]$, la suite $(x_n)_{n \in \mathbb{N}}$ définie par la récurrence :*

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

Vérifie :

$$\forall n \in \mathbb{N}, x_n \in [a, b] \text{ et } \lim_{n \to +\infty} x_n = \alpha$$

Corollaire 1 *Soit α une solution de l'équation $g(\alpha) = \alpha$ et g' continue au voisinage de α, alors on a les trois cas suivants :*

— ***Point fixe attractif :*** *Si $|g'(x)| < 1$ alors il existe un intervalle $[a, b]$ contenant α pour lequel $\forall x_0 \in [a, b]$ la suite $(x_n)_{n \in \mathbb{N}}$ définie par $x_{n+1} = g(x_n)$ converge vers α.*
— ***Point répulsif :*** *Si $|g'(x)| > 1$ alors $\forall x_0 \neq 0$ la suite $(x_n)_{n \in \mathbb{N}}$ définie par $x_{n+1} = g(x_n)$ ne converge pas vers α.*
— ***Point fixe douteux :*** *Si $|g'(x)| = 1$, on ne peut pas conclure, il peut y avoir convergence ou divergence. Pour cela on doit trouver un bon g qui vérifie $|g'(x)| < 1$.*

test d'arrêt

On a la suite (x_n) converge vers α tel que $g(\alpha) = \alpha$. En fixant la tolérance ϵ on estime qu'on atteint la précision ϵ dès qu'il existe un $n_0 \in \mathbb{N}$ tel que :

$$|x_{n_0+1} - x_{n_0}| < \epsilon$$

Estimation de l'erreur

Le nombre minimal d'itérations pour que la solution soit approchée avec une précision ϵ est :

$$|x_n - \alpha| < \epsilon$$

Sachant que

$$|x_n - \alpha| \leq |x_1 - x_0| \frac{k^n}{1-k}$$

Donc

$$n > \frac{ln\left[\frac{(1-k)\epsilon}{|x_1-x_0|}\right]}{ln(k)}, k = \max_I |g'(x)|$$

Ordre de convergence

Définition 1 *On dit que la convergence de (x_n) vers α est d'ordre p ($p \in \mathbb{R}_+^*$) s'il existe une constante $c > 0$ telle que*

$$\lim_{n \to +\infty} \frac{e_{n+1}}{e_n^p} = c \text{ avec } e_n = |x_n - \alpha|$$

Définition 2

- *Lorsque $p = 1$, la convergence de x_n vers α est dite linéaire.*
- *Lorsque $p = 2$, la convergence de x_n vers α est dite quadratique.*

Théorème 4 *Si la suite $(x_n)_n$, définie par $x_{n+1} = g(x_n)$ converge vers α et si g est suffisamment dérivable au voisinage de α, alors l'ordre de la méthode est donné par :*

$$\begin{cases} g'(\alpha) = g''(\alpha) = ...g^{(p-1)}(\alpha) = 0 \\ g^{(p)}(\alpha) \neq 0 \end{cases}$$

De plus on :

$$\lim_{n \to +\infty} \frac{e_{n+1}}{e_n^p} = \frac{g^{(p)}(\alpha)}{p!}$$

Donc on remarque que plus p est grand plus la vitesse de la convergence est rapide.

1.1.3 Méthode de Newton

On suppose que la fonction est dérivable sur l'intervalle $[a, b]$. Le principe de la méthode de Newton qu'on appelle aussi la méthode de la tangente, consiste à choisir un point x_0 de l'intervalle de définition de f, et on considère la tangente à la courbe représentative de f en $(x_0, f(x_0))$. Soit x_1 l'abscisse de l'intervalle de la tangente avec l'axe des abscisses. Puisque la tangente est proche de la courbe, on peut espérer que x_1 donne une meilleure estimation d'une solution de l'équation $f(x) = 0$ que x_0. On recommence alors le procédé à partir de x_1 et on construit par récurrence une suite (x_n) définie par :

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

!

Convergence de la méthode

Théorème 5 *(**Théorème de Newton**) (condition suffisante de convergence globale) Soit $f : [a, b] \to \mathbb{R}$ de classe C^2 vérifiant :*

1. $f(a) \times f(b) < 0$
2. $f'(x) \neq 0, \forall x \in [a, b]$
3. $f''(x) \neq 0, \forall x \in [a, b]$

4. Partant d'un point x_0 qui satisfait l'inégalité

$$f(x_0).f''(x_0) > 0$$

(vérifié par un certain choix de $x_0 \in [a, b]$)

Si les conditions énoncées, ci-dessus, sont satisfaites, alors le processus de Newton :

$$\begin{cases} x_0 \text{ choisi} \\ x_{n+1} = g(x_n) = x_n - \dfrac{f(x_n)}{f'(x_n)} \end{cases}$$

Converge pour ce choix de x_0 vers l'unique solution α de $f(x)$.

Test d'arrêt

Les itérations s'achèvent dés que

$$|x_{n+1} - x_n| < \epsilon \qquad (*)$$

où ϵ est une tolérance fixée.

Ordre de convergence

Définition 3 *On dit que la convergence de $(x_n)_{n\in\mathbb{N}}$ vers α est d'ordre p ($p \in \mathbb{R}_+^*$) s'il existe une constante $c > 0$ telle que*

$$\lim_{n\to+\infty} \frac{e_{n+1}}{e_n^p} = c \text{ avec } e_n = |x_n - \alpha|$$

Théorème 6 *On suppose que $f'(\alpha) \neq 0$, alors si la suite $(x_n)_{n\in\mathbb{N}}$ des itérés de Newton converge, sa vitesse de convergence est (au moins) quadratique.*

1.2 Résolution des systèmes linéaires par les méthodes itératives

Étant donné le système $Ax = b$ à résoudre, les méthodes itératives de résolution des systèmes linéaires consistent à calculer les valeurs successives d'une suite de vecteurs $x^{(k)}$ convergeant vers la solution x quand $k \longrightarrow \infty$.

$$\lim_{k\to+\infty} x^{(k)} = x \qquad (1)$$

On décompose A en $A = M - N$ où M est une matrice inversible alors

$$Ax = b \Longleftrightarrow Mx = Nx + b$$

conduit à l'itération

$$Mx^{(k+1)} = Nx^{(k)} + b, \qquad k \geq 0$$
$$x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b$$

Alors

$$M^{-1}N = B \text{ et } M^{-1}b = c$$

La suite $(x^{(k)})_{k\in\mathbb{N}}$ est obtenue à partir d'un vecteur initial arbitraire $x^{(0)}$, par une relation de récurrence de la forme

$$x^{(k+1)} = Bx^{(k)} + c, \quad \forall k \in \mathbb{N} \qquad (2)$$

où la matrice B est carrée, appelée matrice d'itération de la méthode, et le vecteur c dépendant de la matrice A et du second membre de b du système à résoudre.

Le problème est de trouver les conditions sous les quelles la suite (x^k) est convergente.

De manière générale, on décompose la matrice A sous la forme :

$$A = D - E - F$$

où

— D est la matrice diagonale de A.
— E est la matrice triangulaire inférieure de A.
— F est la matrice triangulaire supérieure de A.

1.2.1 Méthode de Jacobi

Pour la méthode de jacobie, on prend

$$M = D \text{ et } N = E + F$$

d'où

$$M^{-1}N = D^{-1}(E + F) \text{ et } M^{-1}b = D^{-1}b$$

La matrice d'itération de Jacobi est donnée par

$$J = D^{-1}(E + F)$$

d'où la procédure de jacobi s'écrit alors :

$$x^{(k+1)} = D^{-1}(E + F) + D^{-1}b$$

On peut exprimer la procédure de jacobie en fonction des éléments de la matrice A et du vecteur b :

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1, j\neq i}^{n} a_{ij} x_j^{(k)} \right), i = 1, \cdots, n, k \in \mathbb{N}$$

1.2.2 Méthode de Gauss-Seidel

Dans cette méthode, on prend

$$M = D - E \text{ et } N = F$$

d'où

$$M^{-1}N = (D - E)^{-1} \text{ et } M^{-1}b = (D - E)^{-1}b$$

La matrice d'itération de Gauss-Seidel est donnée par

$$G = (D - E)^{-1}F$$

La méthode de Gauss-Seidel s'exprime alors par la suite

$$\begin{cases} x^{(0)} & \text{donné} \\ x^{(k+1)} = (D-E)^{-1}Fx^{(k)} + (D-E)^{-1}b & \text{pour tout entier } k \geq 0 \end{cases}$$

On observe que la méthode de Gauss-Seidel correspond à l'écriture ligne par ligne suivante

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right), i = 1, \cdots, n, k \in \mathbb{N}$$

1.2.3 Méthode de Relaxation

C'est une méthode intermédiaire entre les deux méthodes précédentes dépendant d'un paramètre w. Ce paramètre est déterminé pour obtenir une convergence plus rapide. Dans cette méthode, on prend

$$M = \frac{1}{\omega}D - E \text{ et } N = \frac{1-\omega}{\omega}D + F$$

où ω est un paramètre de relaxation. Alors

$$M^{-1}N = \left(\frac{1}{\omega}D - E\right)^{-1}\left(\frac{1-\omega}{\omega}D + F\right) \text{ et } M^{-1}b = \left(\frac{1}{\omega}D - E\right)^{-1}b$$

La matrice de relaxation est donnée par

$$\mathcal{L}_\omega = \left(\frac{1}{\omega}D - E\right)^{-1}\left(\frac{1-\omega}{\omega}D + F\right)$$

La méthode de relaxation s'exprime alors par la suite

$$\begin{cases} x^{(0)} & \text{donné} \\ x^{(k+1)} = \left(\frac{1}{\omega}D - E\right)^{-1}\left(\frac{1-\omega}{\omega}D + F\right)x^{(k)} + \left(\frac{1}{\omega}D - E\right)^{-1}b & \text{pour tout entier } k \geq 0 \end{cases}$$

Remarque 1
— *La méthode de Gauss-Seidel correspond au cas $\omega = 1$.*
— *Le but de la méthode de relaxation est de trouver le ω qui assure la meilleure convergence.*

La méthode de relaxation s'écrit ligne par ligne comme une extrapolation de la composante obtenue par Gauss-Seidel. On a

$$x_i^{(k+1)}(\text{relax}) = \omega x_i^{(k+1)}(\text{Gauss-Seidel}) + (1-\omega)x_i^{(k)}(\text{relax})$$

1.2.4 Convergence des méthodes itératives

Définition 4 ***(Norme vectorielle) :*** *Pour tout $x = (x_1, \ldots, x_n)^t \in \mathbb{R}^\kappa$, on note par :*

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

Définition 5 ***(Norme matricielle) :*** *Pour toute matrice* $A = (a_{ij}), 1 \leq i \leq n, 1 \leq i \leq n$

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}|$$

$$\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|$$

Définition 6 ***(Rayon spectral) :***

On appelle le rayon spectral de la matrice A le nombe réel donné par :

$$\rho(A) = \max_{1 \leq i \leq n} \{|\lambda_i| \, ; \lambda_i \text{ valeurs propres de } A\}$$

Théorème 7 ***(Condition nécessaires et suffisantes de convergence)***

Le procédé itératif

$$x^{(k+1)} = Bx^{(k)} + c$$

est convergent si et seulement si

$$\rho(B) < 1$$

Théorème 8 ***(le cas d'une matrice symétrique, définie positive)***

Soit A une matrice $n \times n$. *Alors*

1. $\rho(\mathcal{L}_\omega) < 1$ *pour tout* $0 < \omega < 2$
2. *Si de plus la matrice A est symétrique, définie positive, alors on a* $\rho(\mathcal{L}_\omega) < 1$ *si et seulement si* $0 < \omega < 2$

Définition 7 *Une matrice est dite tridiagonale si toutes ses éléments sont nuls sauf les éléments de la diagonale, ceux de la sur-diagonale et ceux de la sous-diagonale.*

Théorème 9 ***(le cas d'une matrice tridiagonale)***

Si A est une matrice tridiagonale, alors les rayons spectraux des matrices de Jacobi et de Gauss Seidel sont liés par la relation :

$$\rho(G) = (\rho(J))^2$$

Donc, si les deux méthodes sont convergente, alors la méthode de Gauss Seidel converge plus rapidement que celle de jacobi.

Théorème 10 *Si une matrice tridiagonale A est symétrique, définie positive, alors la méthode de relaxation converge pour tout* $0 < \omega < 2$ *et il existe un unique* ω_* *optimal assurant la convergence la plus rapide. Il est donné par la formule suivante :*

$$\omega^* = \frac{2}{1 + \sqrt{1 - \rho(J)^2}}$$

Dans ce cas, le rayon spectral de $\mathcal{L}_{\omega_*}$ *est donné par*

$$\rho(\mathcal{L}_{\omega_*}) = |\omega_* - 1|$$

1.3 Intégration numérique

Dans cette section ,on va représenter des méthodes pour approcher les dérivées et les intégrales de fonctions. Concernant l'intégration, on sait bien qu'il n'est pas toujours possible, pour une fonction arbitraire, de trouver la forme explicite d'une primitive. Mais même quand on la connaît, il est parfois difficile de l'utiliser. On a comme exemples

$$\int_0^1 cos(x^2), \int_0^1 \exp -(x^2)$$

Exemple 1 ***(Electromagnétisme) :*** *Considérons un conducteur électrique sphérique de rayon r et de conductivité σ. On veut calculer la distribution de la densité de courant $\boldsymbol{j}$ en fonction de r et t (le temps), connaissant la distribution initiale de la densité de charge $\rho(r)$. Le problème peut être résolu en utilisant les relations entre la densité de courant, champ électrique et la densité de charge, et en remarquant qu'avec la symétrie de la configuration, $\boldsymbol{j}(r,t) = \dfrac{j(r,t)r}{|r|}$, où $j = /\boldsymbol{j}/$. On obtient*

$$j(r,t) = \gamma(r)e^{-\frac{\sigma t}{\epsilon_0}}, \gamma(r) = \frac{\sigma}{\epsilon_0 r2}\int_0^r \rho(\xi)\xi_2 d\xi,$$

où $\epsilon_0 = 8.859 * 10^{-12}$ $farad/m$ est la constante diélectrique du vide.

Pour toutes ces raisons, on fait appel aux méthodes numériques.

1.3.1 Intégration numérique

Les méthodes numériques se divisent en deux grandes familles : Les méthodes de Newton-Cotes et Les méthodes de Gauss.

Méthodes de Newton-cotes

Les méthodes de Newton-Cotes sont basées sur la théorie de l'interpolation polynomiale. L'idée est de diviser l'intervalle $[a,b]$ en n parties égales à l'aide de points $x_i; i = 0; \ldots; n$ avec $x_0 = a$ et $x_n = b$. Le pas d'intégration est :

$$h = x_{i+1} - x_i = \frac{b-a}{n}$$

On note par

$$M_k = \max_{x\in[a,b]} \left|f^{(k)}(x)\right|$$

Sur chaque segment $[x_i, x_{i+1}]$, on remplace la fonction f par son polynôme d'interpolation P_i. On a l'égalité approchée :

$$\int_{x_{i+1}}^{x_i} f(x)dx \simeq \int_{x_{i+1}}^{x_i} P_i(x)dx$$

Suivant le degré du polynôme P_i, trois méthodes sont mises en exergue : La méthode des rectangles, la méthode des trapèzes et la méthode de Simpson.

1. **Méthodes des rectangles :**

   — **Méthode des rectangles à gauche :**
   Sur chaque segment $[x_i, x_{i+1}]$, on remplace la fonction f par son polynôme d'interpolation P_i de degré 0, donc $P_i(x)$ est constant

   $$P_i(x) = f(x_i)$$

   Proposition 1 *La valeur approchée de l'intégrale de f sur l'intervalle $[a, b]$ par la méthode des rectangles à gauche est donnée par :*

   $$I_n = \frac{b-a}{n} \sum_{i=0}^{n-1} f(x_i)$$

   — **Méthodes des rectangles à droite :**
   Sur chaque segment $[x_i, x_{i+1}]$, on remplace la fonction f par son polynôme d'interpolation P_i de degré 0, donc $P_i(x)$ est constant

   $$P_i(x) = f(x_{i+1})$$

   Proposition 2 *La valeur approchée de l'intégrale de f sur l'intervalle $[a, b]$ par la méthode des rectangles à droite est donnée par :*

   $$I_n = \frac{b-a}{n} \sum_{i=0}^{n-1} f(x_{i+1})$$

   — **Méthodes des rectangles milieu :**
   Sur chaque segment $[x_i, x_{i+1}]$, on remplace la fonction f par son polynôme d'interpolation P_i de degré 0, donc $P_i(x)$ est constant

   $$P_i(x) = f(x_{i+1})$$

   Proposition 3 *La valeur approchée de l'intégrale de f sur l'intervalle $[a, b]$ par la méthode des rectangles milieu est donnée par :*

   $$I_n = \frac{b-a}{n} \sum_{i=0}^{n-1} \left(\frac{f(x_i) + f(x_{i+1}}{2} \right)$$

   Proposition 4 *Si f est de classe C^1, la borne supérieure de l'erreur entre la valeur exacte I et la valeur approchée I_n est donnée par :*

   $$|I - I_n| \leq \frac{M_1}{2n} (b-a)^2$$

2. **Méthodes des trapèzes :**
   Sur chaque segment $[x_i, x_{i+1}]$, on remplace la fonction f par son polynôme d'interpolation P_i de degrée 1.

   Proposition 5 *La valeur approchée de l'intégrale de f sur l'intervalle $[a, b]$ par la méthode des trapèzes est donnée par :*

   $$I_n = \frac{b-a}{n} \left(\frac{f(a) + f(b)}{2} \sum_{i=1}^{n-1} f(x_i) \right)$$

Proposition 6 *Si f est de classe C^2, la borne supérieure de l'erreur entre la valeur exacte I et la valeur approchée I_n est donnée par :*

$$|I - I_n| \leq \frac{M_2}{12n^2} (b-a)^3$$

3. **Méthodes de Simpson :**
Sur chaque segment $[x_i, x_{i+1}]$, on remplace la fonction f par son polynôme d'interpolation P_i de degré 2 qui prend les mêmes valeurs que la fonction f aux points x_i,x_{i+1} et au milieu $\xi_i = \dfrac{x_i + x_{i+1}}{2}$.
Sur l'ensemble de l'intervalle $[a, b]$ la méthode de Simpson nécessite $2n+1$ points d'appui..

Proposition 7 *La valeur approchée de l'intégrale de f sur l'intervalle $[a, b]$ par la méthode de Simpson est donnée par :*

$$I_n = \frac{b-a}{6n} \left(f(a) + f(b) + 2\sum_{i=1}^{n-1} f(x_i) + 4\sum_{i=0}^{n-1} f(\xi_i) \right)$$

Proposition 8 *Si f est de classe C^3, la borne supérieure de l'erreur entre la valeur exacte I et la valeur approchée I_n est donnée par :*

$$|I - I_n| \leq \frac{M_3}{192n^3} (b-a)^4$$

Remarque 2 *Dans la cas où la fonction f est de classe C^4, on peut raffiner le résultat précédent et on obtient la majoration suivante, qui est la plus utilisée :*

Proposition 9

$$|I - I_n| \leq \frac{M_4}{2880n^4} (b-a)^5$$

Ordre des méthodes numériques

Définition 8 *Une méthode numérique est dite d'ordre k si elle est exacte pour tout polynôme de degré inférieur ou égal à k. Par linéarité de l'intégration, il suffit qu'elle soit exacte pour tout monôme de degré inférieur ou égal à k.*

Proposition 10 *Pour les trois méthodes étudiées, on a :*
a) La méthode des rectangles est d'ordre 0.
b) La méthode des trapèzes est d'ordre 1.
c) La méthode de Simpson est d'ordre 3.

Méthode de Gauss

Pour la famille de méthodes de Gauss, l'intégrale approchée est donnée par la formule :

$$\tilde{I} = (b-a)\sum_{i=0}^{n} \omega_i f(x_i)$$

On cherche alors les positions x_i et les coefficients correspondants ω_i de manière à ce que la méthode obtenue soit d'ordre $2n+1$, c.à.d, exacte pour les monômes :

$$1, x, x_2, \ldots, x_{2n+1}$$

Les ω_i sont appelées : **Poids**.
Il existe plusieurs méthodes de Gauss, on s'intéressera particulièrement à la méthode de Gauss-Legendre.
Remarque 3 *Généralement les méthodes de Gauss sont présentées sur l'intervalle* $[-1, 1]$*, pour un intervalle quelconque* $[a, b]$*, on utilise le changement de variable :*

$$x \mapsto \left(\frac{b-a}{2}\right) x + \frac{a+b}{2}$$

1. **Méthode de Gauss-Legendre à un point** $(n = 0)$ **:**
   On cherche l'unique point x_0 ainsi que son coefficient correspondant ω_0 de manière à ce que la méthode soit d'ordre :

   $$2n + 1 = 2 * 0 + 1 = 1$$

   donc elle doit être exacte pour les monômes :

   $$P_0(x) = 1, P_1(x) = x$$

   -Pour $P_0(x) = 1$:
   L'intégrale exacte est donnée par :

   $$I = \int_{-1}^{1} 1dx = 2$$

   L'intégrale approchée est donnée par :

   $$\tilde{I} = 2\omega_0$$

   d'où

   $$I = \tilde{I} \Rightarrow 2\omega_0 = 2 \Rightarrow \omega_0 = 1$$

   -Pour $P_0(x) = x$:
   L'intégrale exacte est donnée par :

   $$I = \int_{-1}^{1} xdx = 0$$

   L'intégrale approchée est donnée par :

   $$\tilde{I} = 2\omega_0 x_0$$

   d'où

   $$I = \tilde{I} \Rightarrow 2\omega_0 x_0 = 0 \Rightarrow x_0 = 0$$

   Par suite la méthode approchée est donnée par :

   $$\tilde{I} = 2f(0)$$

2. **Méthode de Gauss-Legendre à deux points** $(n = 1)$ **:**
On cherche les positions x_0 et x_1 ainsi que les coefficients correspondants ω_0 et ω_1 de manière à ce que la méthode soit d'ordre :

$$2n + 1 = 2 * 1 + 1 = 3$$

donc elle doit être exacte pour les monômes :

$$P_0(x) = 1; P_1(x) = x; P_2(x) = x^2; P_3(x) = x^3$$

-Pour $P_0(x) = 1$:
L'intégrale exacte est $I = 2$, l'intégrale approchée est donnée par :

$$\tilde{I} = 2\sum_{i=0}^{1} \omega_i P_0(x_i) = 2\sum_{i=0}^{1} \omega_i 1 = 2\left(\omega_0 + \omega_1\right)$$

On obtient alors l'équation :

$$\omega_0 + \omega_1 = 1$$

-Pour $P_1(x) = x$:
L'intégrale exacte est $I = 0$, l'intégrale approchée est donnée par :

$$\tilde{I} = 2\sum_{i=0}^{1} \omega_i P_1(x_i) = 2\sum_{i=0}^{1} \omega_i x_i = 2\left(\omega_0 x_0 + \omega_1 x_1\right)$$

On obtient alors l'équation :

$$\omega_0 x_0 + \omega_1 x_1 = 0$$

-Pour $P_2(x) = x^2$:
L'intégrale exacte est $I = \dfrac{2}{3}$, l'intégrale approchée est donnée par :

$$\tilde{I} = 2\sum_{i=0}^{1} \omega_i P_2(x_i) = 2\sum_{i=0}^{1} \omega_i x_i^2 = 2\left(\omega_0 x_0^2 + \omega_1 x_1^2\right)$$

On obtient alors l'équation :

$$\omega_0 x_0^2 + \omega_1 x_1^2 = \frac{1}{3}$$

-Pour $P_3(x) = x^3$:
L'intégrale exacte est $I = 0$, l'intégrale approchée est donnée par :

$$\tilde{I} = 2\sum_{i=0}^{1} \omega_i P_3(x_i) = 2\sum_{i=0}^{1} \omega_i x_i^3 = 2\left(\omega_0 x_0^3 + \omega_1 x_1^3\right)$$

On obtient alors l'équation :

$$\omega_0 x_0^3 + \omega_1 x_1^3 = 3$$

On obtient alors quatre équations avec quatre inconnues : En résolvant ce système on obtient :

$$\omega_0 + \omega_1 = \frac{1}{2}, x_0 = -\frac{1}{\sqrt{3}}, x_1 = \frac{1}{\sqrt{3}}$$

La méthode est alors donnée par la formule suivante :

$$\tilde{I} = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

3. **Méthode de Gauss-Legendre à $(n+1)$ points :**

Définition 9 *On appelle polynômes de Legendre, les polynômes donnés par la formule de récurence suivante :*

$$\begin{aligned} &P_0(x) = 1 \\ &P_1(x) = x \\ &\vdots \\ &P_n(x) = \frac{1}{n}\left[(2n-1)xP_{n-1}(x) - (n-1)P_{n-2}(x)\right] \end{aligned}$$

Proposition 11 *L'intégrale approchée est donnée par :*

$$\tilde{I} = 2\sum_{i=0}^{1} \omega_i f(x_i)$$

où
Les $x_i$sont les racines du polynôme de Legendre P_{n+1} de degré $n+1$.
Les ω_i sont donnés par :

$$\omega_i = \frac{1-x_i^2}{(n+1)^2 P_n^2(x_i)}$$

Remarque 4 *A partir de $n = 4$ le calcul des x_i et des ω_i se complique. On a le tableau des valeurs obtenues pour différentes valeurs de n.*

$n+1$	$2\omega_i$	x_i
1	2	0
2	1, 1	$-\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}$
3	$\frac{5}{9}, \frac{8}{9}, \frac{5}{9}$	$-\frac{\sqrt{3}}{5}, \frac{\sqrt{3}}{5}$

1.4 Différentiation numérique

Considérons une fonction f définie par une formule ou un algorithme, dérivable dans un intervalle. Cette fonction est trop compliquée pour que le calcul analytique de sa dérivée soit pratique, mais on souhaite pourtant connaître la valeur numérique de f'. Pour cela il existe deux approches numériques, l'une utilise les développements en série de Taylor et l'autre utilise les formules d'interpolation.

Formule de Taylor : Soit $f(x)$ une fonction possédant $(n+1)$ dérivées continues sur l'intervalle $[a, b]$ et soient x et $\bar{x}$ deux points de cet intervalle. Alors

$$f(x) = p_n(x) + R_{n+1}(x)$$

avec :

$$p_n(x) = f(\bar{x}) + \frac{f'(\bar{x})}{1!}(x-\bar{x}) + \ldots + \frac{f^n(\bar{x})}{n!}(x-\bar{x})^n$$

$$R_{n+1}(x) = \frac{f^{n+1}(\xi)}{(n+1)!}(x-\bar{x})^{n+1}, \bar{x} < \xi < x$$

Utilisation de la formule de Taylor

Considérons une fonction $f : [a, b] \to \mathbb{R}$ continûment dérivable dans $[a, b]$.
Dérivée première :

1. **Différence finie à droite :**
   Pour connaitre une approximation de la dérivée première de f en un point $\bar{x}$ de $]a, b[$. Pour un h assez petit et positif, le procédé le plus simple est

$$f'(\bar{x}) \simeq \frac{f(\bar{x}+h) - f(\bar{x})}{h} \tag{1.4.1}$$

   En effet, si $f \in C^2(]a, b[)$ et en utilisant le développement en série de Taylor au voisinage de $\bar{x}$, on obtient :

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + \frac{f''(\xi)}{2}(x - \bar{x})^2, \xi \in]x, \bar{x}[\tag{1.4.2}$$

   On pose $x = \bar{x} + h$, alors $h = x - \bar{x}$, on le remplace dans (2.5.13), on obtient

$$f(\bar{x}+h) = f(\bar{x}) + hf'(\bar{x}) + \frac{h^2}{2}f''(\xi_1), \xi_1 \in]\bar{x}, \bar{x}+h[\tag{1.4.3}$$

   d'où

$$f'(\bar{x}) = \frac{f(\bar{x}+h) - f(\bar{x})}{h} - \frac{h}{2}f''(\xi_1), \xi_1 \in]\bar{x}, \bar{x}+h[\tag{1.4.4}$$

   ainsi

$$f'_d(\bar{x}) \simeq \frac{f(\bar{x}+h) - f(\bar{x})}{h} \tag{1.4.5}$$

   avec l'erreur $o(h)$ est

$$\frac{h}{2}f''(\xi_1)$$

2. **Différence finie à gauche :**
   De la même manière on définit la dérivée d'ordre 1 de f.

$$f'_g(\bar{x}) \simeq \frac{f(\bar{x}) - f(\bar{x}-h)}{h} \tag{1.4.6}$$

   En effet, si $f \in C^2(]a, b[)$ et en utilisant le développement en série de Taylor au voisinage de $\bar{x}$, on obtient :

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + \frac{f''(\xi_2)}{2}(x - \bar{x})^2, \qquad \xi_2 \in]x, \bar{x}[\tag{1.4.7}$$

   On pose $x = \bar{x} - h$, alors $-h = x - \bar{x}$, on le remplace dans (1.4.7), on obtient

$$f(\bar{x}-h) = f(\bar{x}) - hf'(\bar{x}) + \frac{h^2}{2}f''(\xi_2), \qquad \xi_2 \in]\bar{x}-h, \bar{x}[\tag{1.4.8}$$

d'où

$$f'(\bar{x}) = \frac{f(\bar{x}) - f(\bar{x}-h)}{h} - \frac{h}{2}f''(\xi_2), \qquad \xi_2 \in]\bar{x}-h, \bar{x}[\tag{1.4.9}$$

ainsi

$$f'_g(\bar{x}) \simeq \frac{f(\bar{x}) - f(\bar{x}-h)}{h} \tag{1.4.10}$$

avec l'erreur $o(h)$ est

$$\frac{h}{2}f''(\xi_2)$$

3. **Différence finie centrée :**

$$f'_c(\bar{x}) \simeq \frac{f(\bar{x}+h) - f(\bar{x}-h)}{2h} \tag{1.4.11}$$

En effet, si $f \in C^3(]a,b[)$ et en utilisant le développement en série de Taylor au voisinage de $\bar{x}$ de $f(\bar{x}+h)$ et de $f(\bar{x}-h)$, on obtient :

$$f(\bar{x}+h) = f(\bar{x}) + hf'(\bar{x}) + \frac{h^2}{2}f''(\bar{x}) + \frac{h^3}{2.3!}f^{(3)}(\xi_3), \qquad \xi_3 \in]\bar{x}, \bar{x}+h[\tag{1.4.12}$$

$$f(\bar{x}-h) = f(\bar{x}) - hf'(\bar{x}) + \frac{h^2}{2}f''(\bar{x}) - \frac{h^3}{2.3!}f^{(3)}(\xi_4), \qquad \xi_4 \in]\bar{x}-h, \bar{x}[\tag{1.4.13}$$

en faisant (1.4.12)-(1.4.13), on obtient :

$$f'(\bar{x}) = \frac{f(\bar{x}+h) - f(\bar{x}-h)}{2h} - \frac{h^2}{12}\left[f^{(3)}(\xi_3) - f^{(3)}(\xi_4)\right] \tag{1.4.14}$$

d'où la formule centrée d'ordre 2 avec une erreur en $o(h^2)$

$$f'_c(\bar{x}) \simeq \frac{f(\bar{x}+h) - f(\bar{x}-h)}{2h} \tag{1.4.15}$$

Dérivée seconde : Si $f \in C^4(]a,b[)$, on a

$$f(\bar{x}+h) = f(\bar{x}) + hf'(\bar{x}) + \frac{h^2}{2}f''(\bar{x}) + \frac{h^3}{6}f^{(3)}(\bar{x}) + \frac{h^4}{4!}f^{(4)}(\xi_1), \qquad \xi_1 \in]\bar{x}, \bar{x}+h[$$

et

$$f(\bar{x}-h) = f(\bar{x}) - hf'(\bar{x}) + \frac{h^2}{2}f''(\bar{x}) - \frac{h^3}{2.3!}f^{(3)}(\bar{x}) + \frac{h^4}{4!}f^{(4)}(\xi_2), \qquad \xi_2 \in]\bar{x}-h, \bar{x}[$$

Donc

$$f(\bar{x}+h) + f(\bar{x}-h) = 2f(\bar{x}) + h^2f''(x) + \frac{h^4}{4!}\left[f^{(4)}(\xi_1) + f^{(4)}(\xi_2)\right]$$

pour $\xi_1 \in]\bar{x}, \bar{x}+h[$,$\xi_2 \in]\bar{x}-h, \bar{x}[$, ainsi

$$f''(\bar{x}) = \frac{f(\bar{x}+h) - 2f(\bar{x}) + f(\bar{x}-h)}{h^2} - \frac{h^2}{4!}\left[f^{(4)}(\xi_1) + f^{(4)}(\xi_2)\right]$$

alors

$$f''(\bar{x}) = \frac{f(\bar{x}+h) - 2f(\bar{x}) + f(\bar{x}-h)}{h^2} - \frac{h^2}{4!}\left[f^{(4)}(\xi)\right]$$

avec $\xi \in [\bar{x}+h, \bar{x}-h]$. d'où la formule centrée d'ordre 2 de la deuxième dérivée de f est

$$f'_c(\bar{x}) \simeq \frac{f(\bar{x}+h) - 2f(\bar{x}) + f(\bar{x}-h)}{h^2}$$

Utilisation des formules d'interpolation

Dérivée première :

La formule de Newton s'écrit :

$$f(x) = f(x_0)+f[x_0,x_1](x-x_0)+f[x_0,x_1,x_2](x-x_0)(x-x_1)+\ldots+f[x_0,x_1,\ldots,x_n](x-x_0)\ldots(x-x_{n-1})$$
$$+\frac{f^{(n+1)}(\xi)}{(n+1)!}\prod_{i=0}^{n}(x-x_i), \quad \xi\in]x_0,x_n[$$

On pose $w_n(x)=\prod_{i=0}^{n}(x-x_i)$. En dérivant :

$$f'(x) = f[x_0,x_1]w_0'(x)+\ldots+f[x_0,x_1,\ldots,x_n]w_{n-1}'(x)+\frac{f^{(n+1)}(\xi)}{(n+1)!}w_n'(x)+\frac{f^{(n+2)}(\xi)}{(n+1)!}w_n(x), \quad \xi\in]x_0,x_n[$$

Pour $n=1$, on a

$$f'(x)=\frac{f(x_1)-f(x_0)}{h}+\frac{f^2(\xi)}{2}[x-x_0+x-x_1]+\frac{f^3(\xi)}{2}(x-x_0)(x-x_1), \quad \xi\in]x_0,x_1[$$

avec $h=x_1-x_0$.

Si $x=x_0$, on obtient

$$f'(x)=\frac{f(x_0+h)-f(x_0)}{h}+\frac{h}{2}f^2(\xi), \quad \xi\in]x_0,x_1[$$

Alors la dérivée à droite de f au point x_0 est

$$f'(x_0)\simeq\frac{f(x_0+h)-f(x_0)}{h}$$

Si on utilise la formule de Newton régressive, on obtient

$$f'(x_0)=\frac{f(x_0)-f(x_0-h)}{h}+\frac{f^2(\xi)}{2!h}, \quad \xi\in]x_0,x_1[$$

Alors la dérivée à gauche de f au point x_0 est

$$f'(x_0)\simeq\frac{f(x_0)-f(x_0-h)}{h}$$

Maintenant, on choisit $x=\dfrac{x_0+x_1}{2}$ et on prend $x-x_0=h$, on obtient

$$f'(x)=\frac{f(x+h)-f(x-h)}{2h}-\frac{h^2}{6}f^3(\xi), \quad \xi\in]x_0,x_1[$$

Alors la dérivée centrée d'ordre 2 de f est

$$f'(x)\simeq\frac{f(x+h)-f(x-h)}{2h}$$

Pour $n=2$, si on prend $x=x_0$, $x_1=x+h$ et $x_2=x+2h$, on a

$$f'(x)=\frac{4f(x+h)-3f(x)-f(x+2h)}{2h}+\frac{h^2}{3}f^3(\xi), \quad \xi\in]x,x+2h[$$

d'où la dérivée à droite d'ordre 2 est

$$f'(x)\simeq\frac{4f(x+h)-3f(x)-f(x+2h)}{2h}$$

si on prend $x = x_2$, $x_1 = x - h$ et $x = x_0 - 2h$, on a La formule de la dérivée à gauche d'ordre 2 est

$$f'(x) = \frac{3f(x) - 4f(x-h) + f(x-2h)}{2h} + \frac{h^2}{3}f^3(\xi), \quad \xi \in]x-2h, x[$$

d'où

$$f'(x) \simeq \frac{3f(x) - 4f(x-h) + f(x-2h)}{2h}$$

Et maintenant, si $x = x_0$, $x_1 = x - h$ et $x_2 = x + h$, on a

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{h^2}{6}f^3(\xi), \quad \xi \in]x-h, x+h[$$

d'où la dérivée centrée d'ordre 2 est

$$f'(x) \simeq \frac{f(x+h) - f(x-h)}{2h}$$

Dérivée seconde :
Par le même principe, pour la dérivée seconde on choisit en général d'interpoler sur trois points, ce qui donne (dans le cas des points équidistants) :

$$f(x) = f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + \frac{f^{(3)}(\xi)}{3!}\prod_{i=0}^{2}(x - x_i), \quad \xi \in]x_0, x_2[$$

Alors

$$f'(x) = f[x_0, x_1] + f[x_0, x_1, x_2](2x - x_1 - x_0) + \frac{f^{(4)}(\xi)}{3!}\prod_{i=0}^{2}(x - x_i)$$

$$+ \frac{f^{(3)}(\xi)}{3!}\left[(2x - x_1 - x_0)(x - x_2) + (x - x_0)(x - x_1)\right], \quad \xi \in]x_0, x_2[$$

Ainsi

$$f''(x) = 2f[x_0, x_1, x_2] + \frac{f^{(4)}(\xi)}{3}\left[(2x - x_1 - x_0)(x - x_2) + (x - x_0)(x - x_1)\right]$$

$$-\frac{f^{(5)}(\xi)}{3!} + \frac{f^{(3)}(\xi)}{3!}\left[2(x - x_2)(2x - x_1 - x_0) + (2x - x_0 - x_1)\right]$$

Si, $x = x_0$,$x_1 = x + h$ et $x_2 = x + 2h$, on a

$$f''(x) = \frac{f(x+2h) - 2f(x+h) + f(x)}{h^2} - \frac{2h^2}{3}f^4(\xi) - \frac{h}{3}f^3(\xi), \quad \xi \in]x-h, x+h[$$

d'où

$$f''(x) \simeq \frac{f(x+2h) - 2f(x+h) + f(x)}{h^2}$$

1.4.1 Erreur

Dérivée premier ordre

Théorème 11 *Soient $f : \mathbb{R} \to \mathbb{R}$, de classe C^2, $x_0 \in \mathbb{R}$ et $h > 0$. Alors*

$$E_d = |f'(x_0) - f'_d(x_0)| \leq \frac{h}{2}\max_{x \in [x_0, x_0+1]}|f''(x)|$$

Théorème 12 *Soient $f : \mathbb{R} \to \mathbb{R}$, de classe C^2, $x_0 \in \mathbb{R}$ et $h > 0$. Alors*

$$E_g = \left|f'(x_0) - f'_g(x_0)\right| \leq \frac{h}{2} \max_{x\in[x_0-1,x_0]} |f''(x)|$$

Théorème 13 *Soient $f : \mathbb{R} \to \mathbb{R}$, de classe C^3, $x_0 \in \mathbb{R}$ et $h > 0$. Alors*

$$E_c = |f'(x_0) - f'_c(x_0)| \leq \frac{h^2}{24} \max_{x\in[x_0-\frac{1}{2},x_0+\frac{1}{2}]} \left|f^{(3)}(x)\right|$$

Dérivée seconde

Soient $f : \mathbb{R} \to \mathbb{R}$, de classe C^4, $x_0 \in \mathbb{R}$ et $h > 0$. Alors

$$E_c = |f''(x_0) - f''_c(x_0)| \leq \frac{h^2}{24} \max_{x\in[x_0-1,x_0+1]} \left|f^{(4)}(x)\right|$$

1.5 Équations différentielles ordinaires

1.5.1 Introduction

Définition 10 *On appelle équation différentielle ordinaire du premier ordre toute équation reliant une fonction $x(t)$ et sa dérivée d'ordre 1. D'une manière générale, elle se met sous la forme suivante :*

$$\frac{dx}{dt} = f(t, x), \quad t \in I \subseteq \mathbb{R} \tag{1.5.1}$$

Résoudre une équation différentielle revient à chercher toutes les fonctions $x(t)$ de classe C^1 sur l'intervalle I, telle que

$$\frac{dx(t)}{dt} = f(t, x(t))$$

Remarque 5
- *La dérivée $\left(\frac{dx(t)}{dt}\right)$ peut être notée x'.*
- *Une équation différentielle est d'ordre p si elle implique ds dérivées d'ordre au plus p.*

1.5.2 Problème de Cauchy

Une équation différentielle admet généralement une infinité de solutions. Pour en sélectionner une on doit imposer une condition supplémentaire qui correspond à la valeur prise par la solution par un point de l'intervalle.

On considérera par conséquent les problèmes dits de Cauchy, de la forme suivante : trouver $x : I \subset \mathbb{R} \to \mathbb{R}$ tel que :

$$\begin{cases} x'(t) = f(t, x(t)), \forall t \in I \\ x(t_0) = x_0 \end{cases} \tag{1.5.2}$$

où $f : I \times \mathbb{R} \to \mathbb{R}$ est une fonction donnée, x' est la dérivée de x par rapport à t, t_0 est un point de I et x_0 est une valeur appelée donnée initiale.

Proposition 12 *On suppose que la fonction $f(t,x)$ est*

1. *continue par rapport à ses deux variables ;*
2. *lipschitzienne par rapport à sa deuxième variable, c'est-à-dire qu'il existe une constante positive L (appelée constante de Lipschitz) telle que*

$$|f(t,x_1) - f(t,x_2)| \leq L\,|x_1 - x_2|\,, \forall t \in I, \forall x_1, x_2 \in \mathbb{R}$$

Alors la solution $x = x(t)$ du problème de Cauchy (1.5.2) existe, est unique et appartient à $C^1(I)$.

1.5.3 Méthodes exactes de résolution

a-Équations à variables séparables :

L'équation (1.5.1) est dite à variables séparables si elle se met sous la forme suivante :

$$\frac{dx}{dt} = h(t).g(x)$$

qu'on peut l'écrire sous la forme

$$\frac{dx}{g(x)} = h(t)dt$$

En intégrant les deux membres, on obtient une relation entre x et t de laquelle on peut déduire la solution générale de l'équation donnée.

Exemple 2 *Soit à résoudre le problème de Cauchy suivant :*

$$\begin{cases} (1+e^t)xx' = e^t \\ x(0) = 1 \end{cases}$$

L'équation donnée peut se mettre sous la forme suivante :

$$xdx = \frac{e^t}{1+e^t}dt$$

b-Équations homogènes :

L'équation (1.5.1) est dite homogène si

$$f(\lambda t, \lambda x) = f(t,x)$$

En particulier si

$$\lambda = \frac{1}{t}$$

on a

$$f(t,x) = f(1, \frac{x}{t})$$

En posant

$$x = u.t$$

et en dérivant, on obtient

$$\frac{dx}{dt} = t.\frac{du}{dt} + u$$

l'équation différentielle donnée est alors :

$$\frac{dx}{dt} = f(1, u)$$

des deux égalités précédentes, on obtient :

$$u + t.\frac{du}{dt} = f(1, u)$$

d'où

$$\frac{du}{f(1,u) - u} = \frac{dt}{t}$$

qui est une équation à variables séparables.

Exemple 3 *Soit à résoudre l'équation différentielle :*

$$tx' = \sqrt{t^2 - x^2} + x$$

qui peut se mettre sous la forme :

$$x' = \sqrt{1 - \left(\frac{x}{t}\right)^2} + \frac{x}{t}$$

qui est une équation homogène

On pose

$$x = u.t$$

l'équation devient

$$\frac{du}{\sqrt{1 - u^2}} = \frac{dt}{t}$$

c-Équations linéaires :

L'équation (1.5.1) est dite linéaire si elle se met sous la forme suivante :

$$\frac{dx}{dt} = a(t).x + b(t) \tag{1.5.3}$$

Cette équation est dite avec second membre.

A cette équation on associe l'équation suivante, dite sans second membre :

$$\frac{dx}{dt} = a(t).x \tag{1.5.4}$$

Proposition 13 *La solution générale de l'équation (1.5.3) est la somme de la solution générale de l'équation (1.5.4) et la soltion particulière de l'équation (1.5.3).*

L'équation (1.5.4) est une équation à variables séparables, sa solution générale est donnée par :

$$x(t) = ke^{A(t)}, k \in \mathbb{R} \tag{1.5.5}$$

Pour déterminer la solution particulière de l'équation (1.5.3), on utilise la méthode de la variation de la constante.
On cherche la solution particulière sous la forme suivante :

$$x(t) = k(t)e^{A(t)}$$

Pour déterminer $k(t)$, on dérive x par rapport à t et on obtient

$$x'(t) = k'(t).e^{A(t)} + k(t).a(t).e^{A(t)}$$

En remplaçant dans l'équation(1.5.3), on a

$$k'(t).e^{A(t)} + k(t).a(t).e^{A(t)} = k(t).a(t).e^{A(t)} + b(t)$$

d'où

$$k'(t) = b(t).e^{-A(t)}$$

En intégrant, on détermine $k(t)$.

1.5.4 Systèmes d'équations différentielles

On considère le système d'équations différentielles du premier ordre dont les inconnus sont $x_1(t) \dots x_n(t)$.

$$\begin{cases} x_1' = f_1(t, x_1, \dots x_n) \\ \vdots \\ x_n' = f_n(t, x_1, \dots x_n) \end{cases}$$

où $t \in]t_0, T]$, avec les conditions initiales

$$x_1(t_0) = x_{0,1}, \dots x_n(t_0) = x_{0,n}$$

Ce système peut s'écrire sous la forme

$$X'(t) = AX(t) + B(t)$$

où A est une matrice et $B(t)$ une fonction continue qui admet une solution unique.

En pratique, on résout d'abord l'équation homogène

$$X'(t) = AX(t)$$

puis on détermine une solution particulière de l'équation globale.

On peut aussi utiliser l'une des méthodes numériques

1.5.5 Méthodes numériques

On s'intéresse à la résolution par des méthodes numériques du problème de Cauchy suivant

$$\begin{cases} \dfrac{dy}{dx} = f(x, y), x \in [a, b] \\ \\ x(a) = y_0 \end{cases} \tag{1.5.6}$$

(a, y_0) étant la condition initiale, y_0 est une donnée du problème.
On subdivise l'intervalle $[a, b]$ en n parties égales à l'aide des points :

$$x_0, x_1, \dots, x_{n-1}, x_n = b$$

en utilisant un pas constant

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

Trois méthodes seront mis en exergue : La méthode d'Euler, les méthodes de Runge Kutta et la méthode d'Adams.

a-Méthode d'Euler :

La méthode de Leonhard Euler $(1707-1783)$ est une méthode à pas séparé du premier ordre. Elle consiste à remplacer l'opérateur de dérivation $\frac{d}{dx}$ par le schéma discret $\left(\frac{y_{i+1}-y_i}{h}\right)$ dans notre problème et on obtient le schéma suivant :

$$\begin{cases} x_{i+1} = x_i + h \\ y_{i+1} = y_i + hf(x_i, y_i), 0 \leq i \leq n-1 \end{cases} \tag{1.5.7}$$

En pratique, la méthode d'Euler n'est pas utilisée, car elle n'offre pas une précision suffisante. Cette méthode est convergente et du premier ordre, car l'erreur de consistance vaut

$$|y(x_i) - y_i| = \frac{1}{2}h^2 f'(c, y_i), c \in [x_{i-1}, x_i]$$

b-Méthode de Runge Kutta :

Carl Runge $(1856-1927)$ et Martin Kutta $(1867-1944)$ ont proposé en 1895 de résoudre le problème de Cauchy (1.5.6), le shéma numérique devient :

$$\begin{cases} x_{i+1} = x_i + h_i \\ y_{i+1} = y_i + h\phi(x_i, y_i, h_i) \end{cases} \tag{1.5.8}$$

où la fonction d'incrémentation ϕ est une approximation de la fonction $f(x, y)$ sur l'intervalle $[x_i, x_{i+1}]$.

Runge Kutta d'ordre 2 : Revenons au problème de Cauchy Intégrons les deux cotés sur l'intervalle $[x_i, x_{i+1}]$, on obtient

$$y_{i+1} - y_i = \int_{x_i}^{x_{i+1}} f(x, y(t))dt \tag{1.5.9}$$

Au lieu d'approcher y' à l'aide des formules de dérivation approchée on pourrai approcher l'intégrale dans (1.5.9) à l'aide d'une formule de quadrature numérique. Par exemple la formule du trapèze donne le schéma numérique suivant :

$$y_{i+1} - y_i \simeq \left(\frac{h}{2}f(x_i, y_i) + f(x_{i+1}, y_{i+1})\right)$$

Le schéma devient :

$$y_{i+1} = y_i + \left(\frac{k_1}{2} + \frac{k_2}{2}\right)$$

avec

$$\begin{cases} k_1 = f(x_i, y_i) \\ k_2 = hf(x_i + h, y_i + k_1) \end{cases} \tag{1.5.10}$$

Runge Kutta d'ordre 4 : On introduit le point milieu $x_{i+1/2} = x_i + \frac{h}{2}$ et approchant l'intégrale (1.5.9) par la formule de Simpson, on obtient

$$\int_{x_i}^{x_{i+1}} f(x, y(t))dt \simeq \frac{h}{6}\left(f(x_i, y_i) + 4f(x_{i+1/2}, y_{i+1/2}) + f(x_{i+1}, y_{i+1})\right)$$

Le schéma implicite devient :

$$y_{i+1} = y_i + \frac{h}{6}\left[k_1 + 2k_2 + 2k_3 + k_4\right]$$

avec

$$\begin{cases} k_1 = f(x_i, y_i) \\ k_2 = f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2}k_1\right) \\ k_3 = f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2}k_2\right) \\ k_4 = f(x_i + h, y_i + hk_3) \end{cases} \tag{1.5.11}$$

c-Méthode d'Adams :

-Méthodes d'Adams-Bashfort :

Revenons à la relation (1.5.9) et écrivons le polynôme d'interpolation P_1 de $f(x, y(x))$ aux noeuds (x_{i-1}, y_{i-1}),(x_i, y_i).

$$p_1(x) = f_{i-1} + [f_{i-1}, f_i](x - x_i)$$

où

$$[f_{i-1}, f_i] = \frac{f_{i-1} - f_i}{h}$$

sont les différences divisées d'ordre 1. En approchant f dans l'intégrale (1.5.9) par P_1 on obtient

$$\begin{aligned} y_{i+1} - y_i &\simeq \int_{x_i}^{x_{i+1}} P_1(t)dt \\ &= h\left(\frac{3}{2}f_i - \frac{1}{2}f_{i-1}\right) \end{aligned}$$

d'où l'on tire le schéma

$$y_{i+1} = y_i + h\left(\frac{3}{2}f_i - \frac{1}{2}f_{i-1}\right), 1 \leq i \leq n-1$$

Ce schéma est appelé schéma d'Adams-Bashforth. C'est un schéma explicite à deux pas car y_{i+1} dépend de y_i et y_{i-1}.

De la même manière si on interpole f aux noeuds x_{i-2}, x_{i-1}, x_i par son polynôme de degré 2, on obtient le schéma suivant :

$$y_{i+1} = y_i + h\left(\frac{5}{12}f_{i-2} - \frac{4}{3}f_{i-1} + \frac{23}{12}f_i\right), 2 \leq i \leq n-1$$

-Méthodes d'Adams-Moulton

Pour obtenir des schémas implicites par la méthode d'Adams on approche f par son polynôme d'interpolation P_2 aux noeuds x_{i-1}, x_i, x_{i+1}. P_2 qui s'écrit alors

$$p_2(x) = f_{i-1} + [f_{i-1}, f_i](x - x_{i-1}) + [f_{i-1}, f_i, f_{i+1}](x - x_{i-1})(x - x_i)$$

d'où en reportant l'expression de P_2 dans (1.5.9) on arrive à

$$y_{i+1} - y_i = h\left(-\frac{1}{12}f_{i-1} - \frac{2}{3}f_i + \frac{5}{12}f_{i+1}\right)$$

d'où le schéma

$$y_{i+1} = y_i + h\left(-\frac{1}{12}f_{i-1} - \frac{2}{3}f_i + \frac{5}{12}f_{i+1}\right), 2 \leq i \leq n-1$$

appelé schéma d'Adams-Moulton. C'est un schéma implicite à deux pas.

Les schémas d'Adams-Moulton à p pas ($p \geq 1$) sont des schémas implicites, ils sont obtenus en approchant f dans (1.5.9) par son polynôme d'interpolation obtenu avec les noeuds $t_{i-p+1}, \ldots, t_{i+1}$.

Chapitre 2

équations aux dérivées partielles

2.1 Introduction

En mathématiques, plus précisément en calcul différentiel, une équation aux dérivées partielles abrégée en EDP, est une équation différentielle dont les solutions sont les fonctions inconnues dépendant de plusieurs variables vérifiant certaines conditions concernant leurs dérivées partielles.

Les EDP sont omniprésentes dans les sciences puisqu'elles apparaissent aussi bien dans la mécanique des fluides que dans les théories de la gravitation ou dans l'électromagnétisme (équations de Maxwell).

la résolution des équations aux dérivées partielles est un sujet important. C'est aussi un domaine très travaillé et encore en plein développement. A part dans quelques cas particuliers, il est impossible de calculer explicitement des solutions des différents modèles qu'on présentera par la suite. Il est donc nécessaire d'avoir recours au calcul numérique sur ordinateur pour estimer quantitativement et qualitativement ces solutions.

Il existe de nombreuses méthodes d'approximation numérique des solutions d'équations aux dérivées partielles. On va présenter dans ce chapitre une des plus anciennes et plus simples appelée méthode des différences finies et une autre méthode appelée éléments finis.

2.2 C'est quoi une EDP ?

Une équation différentielle partielle très simple est :

$$\frac{\partial u}{\partial x} = 0$$

où u est une fonction inconnue de x et y. Cette équation implique que les valeurs $u(x, y)$ sont indépendantes de x. Les solutions de cette équation sont :

$$u(x, y) = f(y)$$

où f est une fonction de y.

L'équation différentielle ordinaire

$$\frac{du}{dx} = 0$$

a pour solution :

$$u(x) = c$$

avec c une valeur constante (indépendante de x). Ces deux exemples illustrent qu'en général, la solution d'une équation différentielle ordinaire met en jeu une constante arbitraire, tandis que les équations aux dérivées partielles mettent en jeu des fonctions arbitraires. Une solution des équations aux dérivées partielles n'est généralement pas unique.

Pour les EDP, on peut écrire u la fonction inconnue et $D_x u$ (notation française) ou u_x (notation anglo-saxonne, plus répandue) sa dérivée partielle par rapport à x, soit avec les notations habituelles du calcul différentiel :

$$u_x = \frac{\partial u}{\partial x}$$

et pour les dérivées partielles secondes :

$$u_{xy} = \frac{\partial^2 u}{\partial y \partial x} = \frac{\partial}{\partial y}\left(\frac{\partial u}{\partial x}\right)$$

2.3 Classification des EDP

Définition 11 *On appelle* ***ordre*** *d'une équation aux dérivées partielles l'ordre de la plus grande dérivée présente dans l'équation.*

Les EDP peuvent être classées suivant une autre méthode si nous considérons des EDP linéaires du second ordre de la forme

$$a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} + d\frac{\partial u}{\partial x} + e\frac{\partial u}{\partial y} + fu = g \tag{2.3.1}$$

oé a, b, c, f sont des fonctions qui dépendent de $x$et y.

a une analogie avec la classification des coniques selon la forme quadratique

$$q(x, y) = ax^2 + bxy + cy^2 + dx + ey + f = 0$$

1. Si $b^2 - 4ac > 0$, l'équation (2.3.1) est hyperbolique.

2. Si $b^2 - 4ac = 0$, l'équation (2.3.1) est parabolique.

3. Si $b^2 - 4ac < 0$, l'équation (2.3.1) est elliptique.

Si on applique les conditions précédentes aux divers modèles du deuxième ordre, en remplaçant le couple (x, y) par les variables (x, t), on obtient les trois types d'équations les plus connues :

1. **équation des ondes :** c'est une équation hyperbolique

$$\frac{\partial^2 u}{\partial x^2}(x, t) - \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}(x, t) = f(x, t)$$

où $c > 0$ est la vitesse de propagation de la déformation de l'onde.

2. **équation de la chaleur :** c'est une équation parabolique

$$\frac{\partial u}{\partial t}(x, t) - \mu\frac{\partial^2 u}{\partial x^2}(x, t) = f(x, t)$$

où $\mu > 0$ est un coefficient donné qui correspond à la diffusion thermique.

3. **équation de Laplace ou équation de poisson (une seule variable) :** c'est une équation elliptique

$$\frac{\partial^2 u}{\partial x^2}(x, y) + \frac{\partial^2 u}{\partial t^2}(x, y) = f(x, y)$$

2.4 Conditions aux limites des EDP

Les problèmes aux limites sont des problèmes différentielles posés sur un intervalle $]a, b[$ de la droite réelle, ou sur un ouvert à plusieurs dimensions $\Omega \subset \mathbb{R}$ $(d = 2, 3)$, pour lesquelles les valeurs de l'inconnu (ou de ses dérivées) sont fixées aux extrémités a et b ou sur le bord $\partial\Omega$ dans le cas multidimensionnel.

Dans le cas multidimensionnel, l'équation différentielle met en jeu les dérivées partielles de la solution par rapport aux coordonnées d'espaces. Les équations qui dépendent aussi du temps (t), comme l'équation de la chaleur ou l'équation des ondes, sont appelées problèmes aux limites et aux valeurs initiales. Pour ce type d'équation ont doit aussi fournir la valeur de la solution au point $t = 0$.

En mathématiques, une condition aux limites de Dirichlet (nommée d'après Johann Dirichlet) est imposée à une équation différentielle ou à une équation aux dérivées partielles lorsque l'on spécifie les valeurs que la solution doit vérifier sur les frontières ou limites du domaine. Voici quelques exemples :

— Pour une équation différentielles :

$$y'' + y' = 0$$

La condition aux limites de Dirichlet sur le l'intervalle $[a, b]$ s'exprime alors

$$y(a) = \alpha \text{ et } y(b) = \beta$$

où α et β sont deux nombres donnés.

— Pour une équation aux dérivées partielles :

$$\Delta y + y = 0$$

où Δ est l'opérateur de Laplace ou le Laplacien définit par

$$\Delta y = \sum_{i=1}^{d} \frac{\partial^2 y}{\partial x_i^2}$$

la condition aux limites de Dirichlet sur un domaine $\Omega \subset \mathbb{R}$ s'exprime par :

$$y(x) = f(x), \quad \forall x \in \partial\Omega$$

où f est une fonction connue définie sur $\partial\Omega$

Il y a plusieurs conditions aux limites comme celles de Neumann, de Robin,etc.

2.5 Méthodes numériques de résolution

2.5.1 Méthode des différences finies

Le principe de toutes les méthodes de résolution numériques des équations aux dérivées partielles et d'obtenir des valeurs numériques discrètes (c'est-à-dire en nombre fini) qui approchent la solution exacte.

Les problèmes différentiels présentés précédemment admettent une infinité de solutions. Pour avoir l'unicité, il faut imposer des conditions aux limites sur le bord $\partial\Omega$ de Ω et pour les problèmes dépendant du temps, des conditions initiales en $t = 0$.

On considère l'équation de Poisson

$$-u''(x) = f(x), x \in]a, b[\tag{2.5.1}$$

ou en plusieurs dimensions

$$-\Delta u(X) = f(X), X = (x_1, \ldots, x_d)^T \in \Omega \tag{2.5.2}$$

où f est une fonction donnée et Δ est le Laplacien.

a- Le cas monodimensionnel :

Dans le cas monodimensionnel, une possibilité pour déterminer de manière unique la solution, consiste à imposer la valeur de u en $x = a$ et $x = b$.

$$\begin{cases} -u''(x) = f(x), x \in]a, b[\\ u(a) = \alpha, u(b) = \beta \end{cases} \tag{2.5.3}$$

où α et β sont deux réels donnés.

L'équation différentielle (2.5.3) doit être satisfaite en particulier pour x_j (noeuds) intérieure à$]a, b[$, c'est-à-dire :

$$-u''(x_j) = f(x_j), \;\; j = 1, \ldots, N$$

On peut approcher cet ensemble de N équation en remplaçant la dérivée seconde par une formule de différences finies. Par exemple, si $u :]a, b[\mapsto \mathbb{R}$ est une fonction assez régulière au voisinage d'un point $\bar{x} \in]a, b[$, alors la quantité

$$u''(\bar{x}) \simeq \frac{u(\bar{x} + h) - 2u(\bar{x}) + u(\bar{x} - h)}{h^2}$$

est une approximation de u'' par rapport à h. Ceci suggère d'approcher ainsi le problème (2.5.3) : trouver $\{u_j\}_{j=1}^N$ tels que :

$$\begin{cases} -\dfrac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = f(x_j), \;\; j = 1, \ldots, N \\ u_0 = \alpha, u_{N+1} = \beta \end{cases} \tag{2.5.4}$$

On obtient alors le système suivant :

$$\begin{cases} -\dfrac{u_2 - 2u_1 + u_0}{h^2} = f(x_1) \\ \\ -\dfrac{u_3 - 2u_2 + u_1}{h^2} = f(x_2) \\ \\ \vdots \\ \vdots - \dfrac{u_N - 2u_{N-1} + u_{N-2}}{h^2} = f(x_{N-1}) \\ \\ -\dfrac{u_{N+1} - 2u_N + u_{N-1}}{h^2} = f(x_N) \end{cases} \tag{2.5.5}$$

$$\Rightarrow \begin{cases} -u_2 + 2u_1 - \alpha = h^2 f(x_1) \\ u_3 + 2u_2 - u_1 = h^2 f(x_2) \\ \vdots \\ -u_N + 2u_{N-1} - u_{N-2} = h^2 f(x_{N-1}) \\ -\beta + 2u_N - u_{N-1} = h^2 f(x_N) \end{cases} \tag{2.5.6}$$

$$\Rightarrow \begin{cases} -u_2 + 2u_1 = h^2 f(x_1) + \alpha \\ u_3 + 2u_2 - u_1 = h^2 f(x_2) \\ \vdots \\ -u_N + 2u_{N-1} - u_{N-2} = h^2 f(x_{N-1}) \\ 2u_N - u_{N-1} = h^2 f(x_N) + \beta \end{cases} \tag{2.5.7}$$

$$\Leftrightarrow \begin{pmatrix} 2 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & -1 & & 0 \\ \vdots & & \ddots & \ddots & \ddots & \\ & & \ddots & -1 & 2 & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix} = h^2 \begin{pmatrix} f(x_1) + \dfrac{\alpha}{h^2} \\ f(x_2) \\ f(x_3) \\ \vdots \\ f(x_{N-1}) \\ f(x_N) + \dfrac{\beta}{h^2} \end{pmatrix}$$

$\Leftrightarrow$

$$Au_h = h^2 f \tag{2.5.8}$$

avec $u_h = (u_1, \cdots, u_N)^T$ est le vecteur des inconnus.

$f = \left(f(x_1) + \dfrac{\alpha}{h^2}, f(x_2), \cdots, f(x_{N-1}), f(x_N) + \dfrac{\beta}{h^2} \right)^T$, et $A = \text{tridiag}(-1, 2, -1)$ est la matrice tridiagonale.

b- Le cas bidimensionnel :

On considère le problème de Poisson (2.5.2), dans une région bidimensionnelle Ω.

La méthode des différences finies consiste à approcher les dérivées partielles présentes dans l'EDP à l'aide de taux d'accroissement calculés sur une grille constituée d'un nombre fini de noeuds. La solution u est alors approchée seulement en ces noeuds.

La première étape consiste alors à définir la grille de calcul. Supposons pour simplifier que Ω soit un rectangle $]a, b[\times]c, d[$. Introduisons une partition de $]a, b[$ en sous-intervalles $]x_i, x_{i+1}[$ pour $i = 0, \cdots, N_x$ avec $x_0 = a$ et $x_{N_x+1} = b$.

Notons par $\Delta_x = \{x_0, \cdots, x_{N_x+1}\}$ l'ensemble des extrémités de ces intervalles et $h_x = \max\limits_{i=0,\cdots,N_x} (x_{i+1} - x_i)$ leur longueur maximale.

On discrétise de la même manière y, $\Delta_y = \{y_0, \cdots, y_{N_x+1}\}$ avec $y_0 = c$, $y_{N_y+1} = d$ et $h_y = \max_{j=0,\cdots,N_y} (y_{j+1} - y_j)$. Le produit cartésien $\Delta_h = \Delta_x \times \Delta_y$ définit la grille de calcul sur Ω (voir figure ?) , et $h = \max\{h_x, h_y\}$ mesure le pas de discrétisation.

On cherche les valeurs $u_{i,j}$ qui approchent $u(x_i, y_j)$. On suppose pour simplifier que les noeuds sont uniformément espacés, c'est-à-dire $x_i = x_0 + ih_x$ pour $i = 0, \cdots, N_x + 1$ et $y_j = y_0 + ih_y$ $j = 0, \cdots, N_y + 1$.

Les dérivées partielles du second ordre peuvent être approchées par des taux d'accroissements, comme on l'a fait pour les dérivées ordinaires.

Dans le cas d'une fonction de deux variables, on définit le taux d'accroissement suivant

$$\begin{aligned} \frac{\partial^2 u(x_i, y_i)}{\partial x^2} &= \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h_x^2} \\ \frac{\partial^2 u(x_i, y_i)}{\partial y^2} &= \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{h_y^2} \end{aligned} \tag{2.5.9}$$

Soit le problème

$$\begin{cases} -\Delta u = f, \text{ sur } \Omega \\ u = g, \text{ sur } \partial\Omega \end{cases} \tag{2.5.10}$$

avec

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$$

$\Leftrightarrow$

$$\begin{cases} -\left(\dfrac{\partial^2 u(x_i, y_i)}{\partial x^2} + \dfrac{\partial^2 u(x_i, y_i)}{\partial y^2}\right) = f(x_i, y_j), \text{ sur } \Omega \\ u_{i,j} = g_{i,j}, \forall i, j \text{ tels que } (x_i, y_j) \in \partial\Delta_h \end{cases} \tag{2.5.11}$$

En remplaçant (2.5.9) dans notre problème on obtient

$$-\left[\frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h_x^2} + \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{h_y^2} = f_{i,j}\right]$$

$\Rightarrow$

$$-\left[\frac{1}{h_x^2}u_{i-1,j} + \frac{1}{h_y^2}u_{i,j-1} 2u_{i,j}\left(\frac{1}{h_x^2} + \frac{1}{h_y^2}\right) + \frac{1}{h_y^2}u_{i,j+1} + \frac{1}{h_x^2}u_{i+1,j}\right] = f_{i,j}$$

Le système obtenu peut être écrit sous une forme agréable, c'est-à-dire en numérotant les noeuds (les inconnus) de gauche à droite et de bas en haut. On obtient un système de la forme (2.5.8), avec une matrice $A \in \mathbb{R}^N \times \mathbb{R}^N$ tridiagonale par bloc

$$A = \text{tridiag}(D, T, D)$$

Elle comporte N_y lignes et N_y colonnes et chaque terme est une matrice $N_x \times N_x$.
La matrice $D \in \mathbb{R}^{N_x \times N_x}$ est diagonale et ses coefficients sont $-\dfrac{1}{h_y^2}$.
La matrice $T \in \mathbb{R}^{N_x \times N_x}$ est tridiagonale symétrique

$$T = \text{tridiag}\left(-\frac{1}{h_x^2}, \frac{2}{h_x^2} + \frac{2}{h_y^2}, -\frac{1}{h_x^2}\right)$$

La matrice A est symétrique définie positive.

Exemple 4 *Le déplacement transverse u par rapport au plan de référence $z = 0$ d'une membrane élastique soumise à un chargement $f(x, y) = 8\pi^2 \sin(2\pi x)\cos(2\pi y)$ qui vérifie le problème de Poisson (2.5.2) dans le domaine $\Omega =]0, 1[\times]0, 1[$.*

On choisit les données de Dirichlet sur le bord $\partial\Omega$ de la manière suivante :

$$\begin{cases} g(0, y) = g(1, y) = 0, \forall y \in]0, 1[\\ g(x, 0) = g(x, 1) = sin(2\pi x), \forall x \in]0, 1[\end{cases}$$

On prend $h_x = 0.25$, alors $N_x = 3$ et $h_y = \dfrac{1}{3}$, alors $N_y = 2$. La solution exacte de ce problème est donnée par $u(x, y) = \sin(2\pi x)\cos(2\pi y)$.

*On reprend le rectangle Ω (voir **??**), on le discrétise (voir figure **??**) et on applique le schéma des différences finies.*

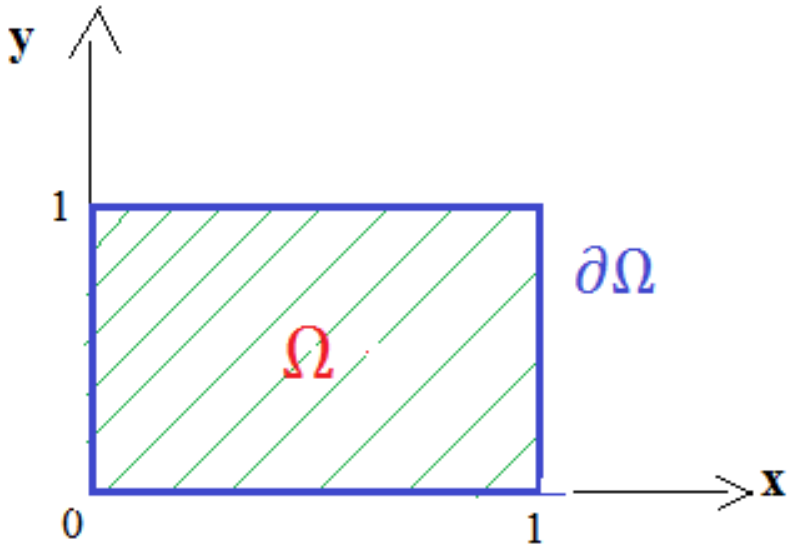
Pour $i = 1, j = 1$

$$-\frac{1}{h_x^2}u_{0,1} - \frac{1}{h_y^2}u_{1,0} + 2u_{1,1}\left(\frac{1}{h_x^2} + \frac{1}{h_y^2}\right) - \frac{1}{h_y^2}u_{1,2} - \frac{1}{h_x^2}u_{2,1} = f_{1,1}$$

Pour $i = 2, j = 1$

$$-\frac{1}{h_x^2}u_{1,1} - \frac{1}{h_y^2}u_{2,0} + 2u_{2,1}\left(\frac{1}{h_x^2} + \frac{1}{h_y^2}\right) - \frac{1}{h_y^2}u_{2,2} - \frac{1}{h_x^2}u_{3,1} = f_{2,1}$$

math et num 3ELT/exempleEDp.png



Pour $i = 3, j = 1$

$$-\frac{1}{h_x^2}u_{2,1} - \frac{1}{h_y^2}u_{3,0} + 2u_{3,1}\left(\frac{1}{h_x^2} + \frac{1}{h_y^2}\right) - \frac{1}{h_y^2}u_{3,2} - \frac{1}{h_x^2}u_{4,1} = f_{3,1}$$

Pour $i = 1, j = 2$

$$-\frac{1}{h_x^2}u_{0,2} - \frac{1}{h_y^2}u_{1,1} + 2u_{1,2}\left(\frac{1}{h_x^2} + \frac{1}{h_y^2}\right) - \frac{1}{h_y^2}u_{1,3} - \frac{1}{h_x^2}u_{2,2} = f_{1,2}$$

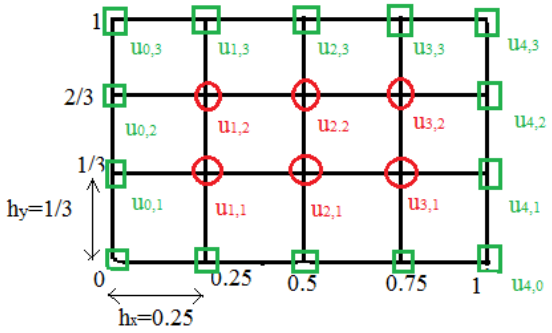
Pour $i = 2, j = 2$

$$-\frac{1}{h_x^2}u_{1,2} - \frac{1}{h_y^2}u_{2,1} + 2u_{2,2}\left(\frac{1}{h_x^2} + \frac{1}{h_y^2}\right) - \frac{1}{h_y^2}u_{2,3} - \frac{1}{h_x^2}u_{3,2} = f_{2,2}$$

Pour $i = 3, j = 2$

$$-\frac{1}{h_x^2}u_{2,2} - \frac{1}{h_y^2}u_{3,1} + 2u_{3,2}\left(\frac{1}{h_x^2} + \frac{1}{h_y^2}\right) - \frac{1}{h_y^2}u_{3,3} - \frac{1}{h_x^2}u_{4,2} = f_{3,2}$$

math et num 3ELT/discrétisation.png



On obtient le système d'équations suivant :

$$\begin{cases} 2u_{1,1}\left(\frac{1}{h_x^2}+\frac{1}{h_y^2}\right)-\frac{1}{h_y^2}u_{1,2}-\frac{1}{h_x^2}u_{2,1}=f_{1,1}+\frac{1}{h_x^2}u_{0,1}+\frac{1}{h_y^2}u_{1,0} \\ -\frac{1}{h_x^2}u_{1,1}+2u_{2,1}\left(\frac{1}{h_x^2}+\frac{1}{h_y^2}\right)-\frac{1}{h_y^2}u_{2,2}-\frac{1}{h_x^2}u_{3,1}=f_{2,1}+\frac{1}{h_y^2}u_{2,0} \\ -\frac{1}{h_x^2}u_{2,1}+2u_{3,1}\left(\frac{1}{h_x^2}+\frac{1}{h_y^2}\right)-\frac{1}{h_y^2}u_{3,2}=f_{3,1}+\frac{1}{h_y^2}u_{3,0}+\frac{1}{h_x^2}u_{4,1} \\ -\frac{1}{h_x^2}u_{0,2}-\frac{1}{h_y^2}u_{1,1}+2u_{1,2}\left(\frac{1}{h_x^2}+\frac{1}{h_y^2}\right)-\frac{1}{h_x^2}u_{2,2}=f_{1,2}+\frac{1}{h_x^2}u_{0,2}+\frac{1}{h_y^2}u_{1,3} \\ -\frac{1}{h_x^2}u_{1,2}-\frac{1}{h_y^2}u_{2,1}+2u_{2,2}\left(\frac{1}{h_x^2}+\frac{1}{h_y^2}\right)-\frac{1}{h_x^2}u_{3,2}=f_{2,2}+\frac{1}{h_y^2}u_{2,3} \\ -\frac{1}{h_x^2}u_{2,2}-\frac{1}{h_y^2}u_{3,1}+2u_{3,2}\left(\frac{1}{h_x^2}+\frac{1}{h_y^2}\right)=f_{3,2}+\frac{1}{h_y^2}u_{3,3}+\frac{1}{h_x^2}u_{4,2} \end{cases}$$

d'où on obtient les système matriciel

$$Au_h = F$$

avec

$$A = \begin{pmatrix} \boxed{\begin{matrix} 2\left(\frac{1}{h_x^2}+\frac{1}{h_y^2}\right) & -\frac{1}{h_x^2} & 0 \\ -\frac{1}{h_x^2} & 2\left(\frac{1}{h_x^2}+\frac{1}{h_y^2}\right) & -\frac{1}{h_x^2} \\ 0 & -\frac{1}{h_x^2} & 2\left(\frac{1}{h_x^2}+\frac{1}{h_y^2}\right) \end{matrix}} & \boxed{\begin{matrix} -\frac{1}{h_y^2} & 0 & 0 \\ 0 & -\frac{1}{h_y^2} & 0 \\ 0 & 0 & -\frac{1}{h_y^2} \end{matrix}} \\ \boxed{\begin{matrix} -\frac{1}{h_y^2} & 0 & 0 \\ 0 & -\frac{1}{h_y^2} & 0 \\ 0 & 0 & -\frac{1}{h_y^2} \end{matrix}} & \boxed{\begin{matrix} 2\left(\frac{1}{h_x^2}+\frac{1}{h_y^2}\right) & -\frac{1}{h_x^2} & 0 \\ -\frac{1}{h_x^2} & 2\left(\frac{1}{h_x^2}+\frac{1}{h_y^2}\right) & -\frac{1}{h_x^2} \\ 0 & -\frac{1}{h_x^2} & 2\left(\frac{1}{h_x^2}+ \end{matrix}} \end{pmatrix}$$

$$= \begin{pmatrix} \boxed{\begin{matrix} 50 & -16 & 0 \\ -16 & 50 & -16 \\ 0 & -16 & 50 \end{matrix}} & \boxed{\begin{matrix} -9 & 0 & 0 \\ 0 & -9 & 0 \\ 0 & 0 & -9 \end{matrix}} \\ \boxed{\begin{matrix} -9 & 0 & 0 \\ 0 & -9 & 0 \\ 0 & 0 & -9 \end{matrix}} & \boxed{\begin{matrix} -16 & 50 & -16 \\ 50 & -16 & 0 \\ 0 & -16 & 50 \end{matrix}} \end{pmatrix}$$

,

$$F = \begin{pmatrix} f_{1,1} + \frac{1}{h_x^2}u_{0,1} + \frac{1}{h_y^2}u_{1,0} \\ f_{2,1} + \frac{1}{h_y^2}u_{2,0} \\ f_{3,1} + \frac{1}{h_y^2}u_{3,0} + \frac{1}{h_x^2}u_{4,1} \\ f_{1,2} + \frac{1}{h_x^2}u_{0,2} + \frac{1}{h_y^2}u_{1,3} \\ f_{2,2} + \frac{1}{h_y^2}u_{2,3} \\ f_{3,2} + \frac{1}{h_y^2}u_{3,3} + \frac{1}{h_x^2}u_{4,2} \end{pmatrix} = \begin{pmatrix} -30.4784 \\ 0 \\ 30.4784 \\ -30.4784 \\ 0 \\ 30.4784 \end{pmatrix}$$

et

$$u_h = \begin{pmatrix} u_{1,1} \\ u_{2,1} \\ u_{3,1} \\ u_{1,2} \\ u_{2,2} \\ u_{3,2} \end{pmatrix} = \begin{pmatrix} -0.7434 \\ 0 \\ 0.7434 \\ -0.7434 \\ 0 \\ 0.7434 \end{pmatrix}$$

2.5.2 Méthode des éléments finis

La méthode des éléments finis est une alternative à la méthode des différences finies pour approcher les problèmes aux limites. Elle est basée sur la reformulation du problème (2.5.3).

a- Le cas monodimensionnel :

Considérons le problème (2.5.3) et multiplions les deux membres de l'égalité par une fonction $v \in C^1([a,b])$ et on intègre l'égalité sur l'intervalle $[a,b]$, on obtient :

$$-u'' = f$$

$$\Leftrightarrow -u''(x)v(x) = f(x)v(x)$$

$$\Leftrightarrow -\int_a^b u''(x)v(x)dx = \int_a^b f(x)v(x)dx$$

En effectuant une intégration par partie, on obtient :

$$(u'(x)v(x))' = u''(x)v(x) + u'(x)v'(x)$$

$$\Rightarrow \int_a^b (u'(x)v(x))'\,dx = \int_a^b u''(x)v(x)dx + \int_a^b u'(x)v'(x)dx$$

$$\Rightarrow \int_a^b u''(x)v(x)dx = \int_a^b \left(u'(x)v(x)\right)' dx - \int_a^b u'(x)v'(x)dx = \left[u'(x)v(x)\right]_a^b - \int_a^b u'(x)v'(x)dx$$

Alors

$$-\int_a^b u''(x)v(x)dx = \int_a^b f(x)v(x)dx$$

$$\Leftrightarrow -\left[u'(x)v(x)\right]_a^b + \int_a^b u'(x)v'(x)dx = \int_a^b f(x)v(x)dx$$

Si on suppose de plus que v s'annule aux extrémités $x = a$ et $x = 0$ (c'est-à-dire : $v(a) = v(b) = 0$), alors le problème (2.5.3) devient :

trouver $u \in C^1([a, b])$ tel que $u(a) = \alpha, u(b) = \beta$ et

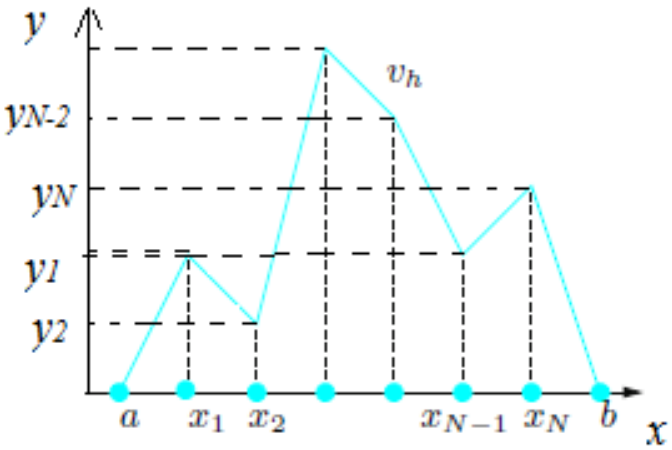
$$\int_a^b u'(x)v'(x)dx = \int_a^b f(x)v(x)dx \tag{2.5.12}$$

Cette équation s'appelle formulation faible du problème (2.5.3).

On discrétise maintenant l'intervalle $[a, b]$, en le subdivisant en $(N + 1)$ intervalles où $h = x_{i+1} - x_i, i = 0, \ldots, N$, alors l'équation (2.5.12) devient :

$$\sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} u'(x)v'(x)dx = \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} f(x)v(x)dx \tag{2.5.13}$$

En interpolant v sur chaque intervalle $I_i = [x_i, x_{i+1}]$, on obtient la fonction affine par morceaux v_h (voir la figure suivante).



On utilise l'interpolation de Lagrange sur chaque intervalle $I_i = [x_i, x_{i+1}]$, on a $v_h = L(x)$, un polynôme de degré inférieure ou égale à 1 et sur chaque noeud (x_i, y_i), on a $v_h(x_i) \simeq v(x_i)$.

$$L(x) = \sum_{j=0}^{N} l_j(x) y_j$$

avec $y_j = v(x_j)$.

On prend $N = 4$, donc on a $(x_0 = a, x_1, x_2, x_3, x_4 = b)$ et $(y_0 = v(a) = 0, y_1, y_2, y_3, y_4 = v(b) = 0)$, on obtient alors

$$v_h = \begin{cases} \dfrac{x - x_1}{x_0 - x_1} y_0 + \dfrac{x - x_0}{x_1 - x_0} y_1, & x \in [x_0, x_1] \\ \dfrac{x - x_2}{x_1 - x_2} y_1 + \dfrac{x - x_1}{x_2 - x_1} y_2, & x \in [x_1, x_2] \\ \dfrac{x - x_3}{x_2 - x_3} y_2 + \dfrac{x - x_2}{x_3 - x_2} y_3, & x \in [x_2, x_3] \\ \dfrac{x - x_4}{x_3 - x_4} y_3 + \dfrac{x - x_3}{x_4 - x_3} y_4, & x \in [x_3, x_4] \end{cases}$$

Soit

$$\varphi_1(x) = \begin{cases} \dfrac{x - x_0}{x_1 - x_0}, & x \in [x_0, x_1] \\ \dfrac{x - x_2}{x_1 - x_2}, & x \in [x_1, x_2] \\ 0 & \text{ailleurs} \end{cases}$$

$$\varphi_2(x) = \begin{cases} \dfrac{x - x_1}{x_2 - x_1}, & x \in [x_1, x_2] \\ \dfrac{x - x_3}{x_2 - x_3}, & x \in [x_2, x_3] \\ 0 & \text{ailleurs} \end{cases}$$

$$\varphi_3(x) = \begin{cases} \dfrac{x - x_2}{x_3 - x_2}, & x \in [x_2, x_3] \\ \dfrac{x - x_4}{x_3 - x_4}, & x \in [x_3, x_4] \\ 0 & \text{ailleurs} \end{cases}$$

Alors

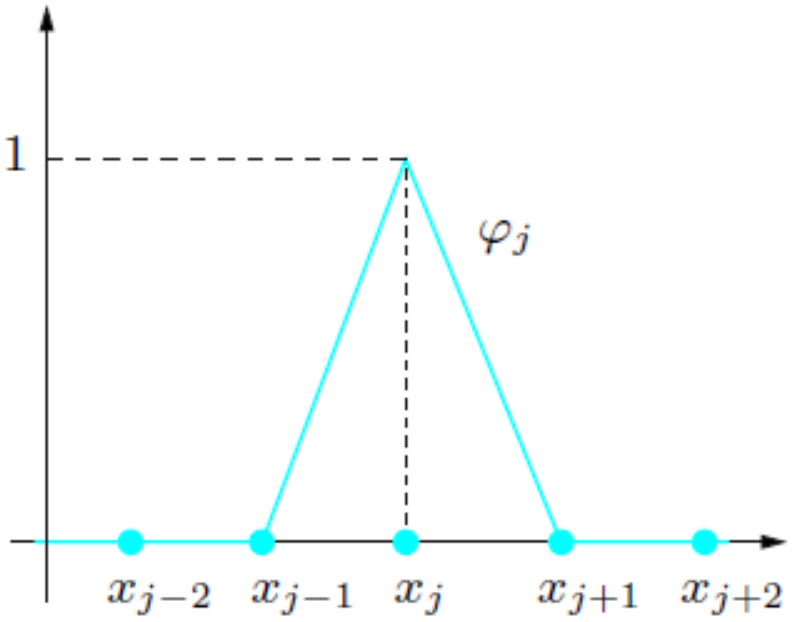
$$v_h(x) = \varphi_1(x) y_1 + \varphi_2(x) y_2 + \varphi_3(x) y_3$$

On remarque que

$$\varphi_i(x_j) = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{sinon} \end{cases}$$

qu'on appelle les fonction de base (voir la figure suivante).

math et num 3ELT/lafonctionphi.png



De la même manière on interpole u sur l'intervalle I_i, on obtient la fonction affine par morceaux u_h.

$$u_h(x) = \varphi_0(x)u_0 + \varphi_1(x)u_1 + \varphi_2(x)u_2 + \varphi_3(x)u_3 + \varphi_4(x)u_4$$

avec

$$\varphi_0(x) = \begin{cases} \dfrac{x - x_1}{x_0 - x_1}, & x \in [x_0, x_1] \\ 0 & \text{ailleurs} \end{cases}$$

$$\varphi_4(x) = \begin{cases} \dfrac{x - x_3}{x_4 - x_3}, & x \in [x_3, x_4] \\ 0 & \text{ailleurs} \end{cases}$$

Alors l'équation (2.5.13) devient :

$$\sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} u_h'(x)v_h'(x)dx = \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} f(x)v_h(x)dx \tag{2.5.14}$$

où

$$u_h \in V_h = \left\{ v_h \in C^0\left([a,b]\right), v_h = P_1 sur I_i \right\}$$

On appelle V_h l'espace des éléments finis de degré 1. et

$$v_h \in V_h^0 = \{v_h \in V_h, v_h(a) = v_h(b) = 0\}$$

Les fonctions V_h^0 sont affines par morceaux. Toute fonction v_h de V_h^0 admet la représentation

$$v_h = \sum_{i=0}^{N} y_i \varphi_i(x)$$

où

$$\varphi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & \text{si } x \in I_{i-1} \\ \dfrac{x - x_{i+1}}{x_i - x_{i+1}}, & \text{si } x \in I_i \\ 0 & \text{sinon} \end{cases}$$

$\varphi_i(x)$ est une famille génératrice de V_h^0 car

$$\lambda_1\varphi_1(x) + \lambda_2\varphi_2(x) + \lambda_3\varphi_3(x) + \ldots + \lambda_N\varphi_N(x) = 0$$

$$\Rightarrow \begin{cases} \lambda_1 = 0 \text{ pour } x = x_1 \\ \lambda_2 = 0 \text{ pour } x = x_2 \\ \lambda_3 = 0 \text{ pour } x = x_3 \\ \vdots \\ \lambda_N = 0 \text{ pour } x = x_N \end{cases}$$

donc elle est libre, alors $\{\varphi_i\}$ est une base de V_h^0. Donc on peut se contenter de satisfaire (2.5.14) seulement pour les fonctions de base φ_i, $i = 1 \ldots, N$. En utilisant le fait que φ_i s'annule en dehors des intervalles I_{i-1} et I_i, (2.5.14) donne

$$\int_{I_{i-1}\cup I_i} u_h'(x)\varphi_i'(x)dx = \int_{I_{i-1}\cup I_i} f(x)\varphi_i(x)dx, \qquad i = 1, \ldots, N$$

$$\Leftrightarrow \int_{x_{i-1}}^{x_i} u_h'(x)\varphi_i'(x)dx + \int_{x_i}^{x_{i+1}} u_h'(x)\varphi_i'(x)dx = \int_{x_{i-1}}^{x_i} f(x)\varphi_i(x)dx + \int_{x_i}^{x_{i+1}} f(x)\varphi_i(x)dx \tag{2.5.15}$$

On peut de plus écrire

$$u_h = \sum_{i=1}^{N} u_i(x)\varphi_i(x) + \alpha\varphi_0(x) + \beta\varphi_{N+1}$$

où

$$\begin{cases} u_i = u_h(x_i) \\ \varphi_0(x) = \dfrac{x - x_1}{a - x_1}, x \in [a, x_1] \\ \varphi_{N+1}(x) = \dfrac{x - x_N}{b - x_N}, x \in [x_N, b] \end{cases}$$

On insérant ces expressions dans (2.5.15), on obtient

$$u_1\int_{x_0}^{x_1} \varphi_1'(x)\varphi_1'(x)dx + u_1\int_{x_1}^{x_2} \varphi_1'(x)\varphi_1'(x)dx + u_1\int_{x_1}^{x_2} \varphi_1'(x)\varphi_2'(x)dx = \int_{x_0}^{x_2} f(x)\varphi_1(x)dx + \frac{\alpha}{x_1 - a}$$

$$u_1\int_{x_1}^{x_2} \varphi_1'(x)\varphi_2'(x)dx + u_2\int_{x_1}^{x_3} \varphi_2'(x)\varphi_2'(x)dx + u_3\int_{x_2}^{x_3} \varphi_2'(x)\varphi_3'(x)dx = \int_{x_1}^{x_3} f(x)\varphi_2(x)dx$$

$$u_{i-1}\int_{I_{i-1}} \varphi_{i-1}'(x)\varphi_i'(x)dx + u_i\int_{I_{i-1}\cup I_i} \varphi_i'(x)\varphi_i'(x)dx + u_{i+1}\int_{I_i} \varphi_{i+1}'(x)\varphi_i'(x)dx = \int_{I_{i-1}\cup I_i} f(x)\varphi_i(x)dx,$$

$$u_{N-1}\int_{I_{N-1}} \varphi_{N-1}'(x)\varphi_N'(x)dx + u_N\int_{I_{N-1}\cup I_N} \varphi_N'(x)\varphi_N'(x)dx + u_{N+1}\int_{I_N} \varphi_{N+1}'(x)\varphi_N'(x)dx$$

$$= \int_{I_{N-1} \cup I_N} f(x)\varphi_N(x)dx + \frac{\beta}{b - x_N}$$

Dans le cas particulier où tous les intervalles ont la même longueur h, on a

$$\begin{cases} \varphi'_{i-1} = -\dfrac{1}{h} \text{ dans } I_{i-1} \\ \varphi'_i = \dfrac{1}{h} \text{ dans } I_{i-1} \\ \varphi'_i = -\dfrac{1}{h} \text{ dans } I_i \\ \varphi'_{i+1} = \dfrac{1}{h} \text{ dans } I_i \end{cases}$$

On obtient donc

$$2u_1 - u_2 = h \int_{x_0}^{x_2} f(x)\varphi_1(x)dx + \frac{\alpha}{h}$$

$$-u_{i-1} + 2u_i - u_{i+1} = h \int_{I_{i-1} \cup I_i} f(x)\varphi_i(x)dx, \qquad i = 2, \dots N-1$$

$$-u_{N-1} + 2u_N = h \int_{I_{N-1} \cup I_N} f(x)\varphi_N(x)dx + \frac{\beta}{h}$$

on obtient le système linéaire

$$Au_h = hF$$

$$\Leftrightarrow \begin{pmatrix} 2 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & -1 & & 0 \\ \vdots & & \ddots & \ddots & \ddots & \\ & & \ddots & -1 & 2 & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix} = h \begin{pmatrix} \int_{x_0}^{x_2} f(x)\varphi_1(x)dx + \dfrac{\alpha}{h} \\ \int_{x_1}^{x_3} f(x)\varphi_2(x)dx \\ \int_{x_2}^{x_4} f(x)\varphi_3(x)dx \\ \vdots \\ \int_{x_{N-2}}^{x_N} f(x)\varphi_{N-1}(x)dx \\ \int_{x_{N-1}}^{x_{N+1}} \varphi_N(x)dx + \dfrac{\beta}{h} \end{pmatrix}$$

Le système linéaire obtenu a la même matrice que celle obtenue dans les différences finies mais le second membre ainsi que la solution sont différents.

Chapitre 3

Techniques d'optimisation

3.1 Introduction

L'optimisation est une branche des mathématiques cherchant à modéliser, à analyser et à résoudre analytiquement ou numériquement les problèmes qui consistent à minimiser ou maximiser une fonction sur un ensemble.L'optimisation intervient dans de nombreux domaines :

- En recherche opérationnelle (problème de transport, économie, gestion de stocks...).
- En analyse numérique (approximation/résolution de systèmes linéaires, non linéaires...).
- En automatique (modélisation de systèmes, filtrage...).
- En ingénierie (dimensionnement de structures, conception optimale de systèmes (réseaux, ordinateurs...)).
- Dans la théorie des jeux pour la recherche de stratégies.
- En théorie du contrôle et de la commande.

Beaucoup de systèmes susceptibles d'être décrits par un modèle mathématique sont optimisés. La qualité des résultats et des prédictions dépend de la pertinence du modèle, du bon choix des variables que l'on cherche à optimiser, de l'efficacité de l'algorithme et des moyens pour le traitement numérique.

3.2 Problèmes d'optimisation

L'optimisation est l'étude des problèmes qui s'expriment de la manière suivante.

Minimisation

Définition 12 *Étant donné une fonction $f : A \to \mathbb{R}$, trouver un élément $\bar{x}$ de A tel que $f(\bar{x}) \leq f(x)$ pour tout $x \in A$. On dit que l'on cherche à minimiser la fonction f sur l'ensemble A.*

- La fonction f peut s'appeler : fonction coût (ou coût), fonction objectif (ou objectif), critère, etc.
- L'ensemble A est appelé ensemble admissible, et les points de A sont appelés les points admissibles du problème à optimiser.
- le point $\bar{x}$ est appelé solution ou minimum ou minimiseur du problème.
- On dit que le problème est réalisable si l'ensemble A est non vide.

On peut écrire ce problème de différentes manières :

$$\min_{x \in A} f(x)$$

$$\min \{f(x) \mid x \in A\}$$

$$\min f(A)$$

ou

$$\begin{cases} \min f(x) \\ x \in A \end{cases}$$

Maximisation
Un problème de maximisation d'une fonction f est équivalent au problème de minimisation de $-f$. On peut le définir comme suit

$$f(x) \leq f(\bar{x}), \forall x \in A \Leftrightarrow -f(x) \geq -f(\bar{x})$$

L'équivalence veut dire ici que les solutions sont les mêmes et que les valeurs optimales sont opposées. En particulier, une méthode pour analyser et résoudre un problème de minimisation pourra être utilisée pour analyser et résoudre un problème de maximisation.
Certaines propriétés de f sont maintenues par passage de f en $-f$ (continuité, différentiabilité ou linéarité), alors que d'autres, sont détruites par cette transformation , par exemple : minimiser f sur A avec f et A convexes, peut être considérer comme un problème facile, alors que : maximiser f sur A avec f et A convexes, est de fait un problème d'optimisation très difficile. La convexité est un exemple de propriété tournée davantage vers la minimisation que vers la maximisation.

3.2.1 Région admissible

N'importe quel point x satisfaisant les contraintes d'égalités et d'inégalités est dit point admissible du problème. L'ensemble de point satisfaisant ces contraintes est dit région admissible de $f(x)$. Ainsi la région admissible peut être définie par l'ensemble

$$A = \{x \mid a_i(x) = 0, i = 1, \ldots, p \text{ et } c_j(x), j = 1, \ldots, q\}$$

N'importe quel point x qui n'appartient pas à A est dit point non admissible.

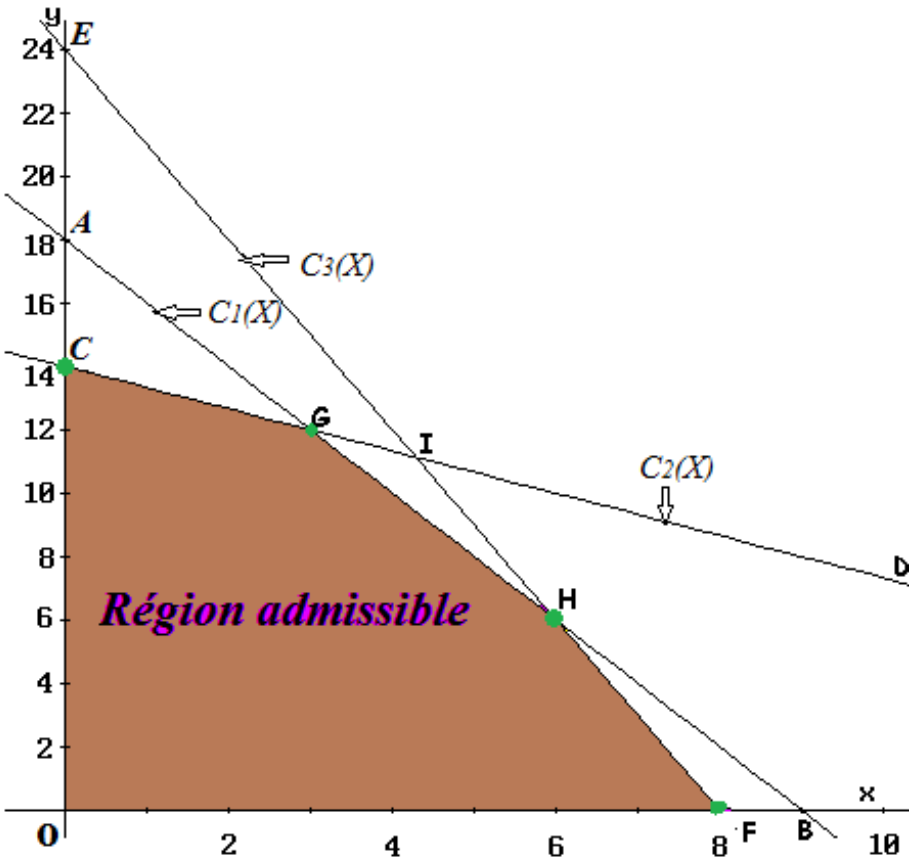
- Si les contraintes du problème à optimiser étaient toutes des inégalités, les contraintes se divisent en trois parties :
  1. Point intérieur est un point pour lequel $c_j(x) > 0$ pour tout j. Ceci est un point admissible.
  2. Point au bord est un point pour lequel au moins un $c_j(x) = 0$. Ceci peut être ou ne pas être un point admissible.
  3. Point extérieurs est un point pour lequel au moins un $c_j(x) < 0$. Ceci est un point non admissible.
- Si une contrainte $c_k(x)$ est nulle durant une itération spécifique, elle est dite active.
- Si $c_k(\bar{x})$ est nulle la convergence est atteinte, l'optimum $\bar{x}$ est localisé au bord. Dans ce cas là, le point optimal est dit contraint.

Exemple 5 *Résoudre en utilisant la méthode graphique le problème d'optimisation suivant :*

$$\max f(X) = 3x + 2y$$
$$S.C : C_1(X) = 2x + y \leq 18$$
$$C_2(X) = 2x + 3y \leq 42$$
$$C_3(X) = 3x + y \leq 24$$
$$C_4(X) = x \geq 0$$
$$C_5(X) = y \geq 0$$

— *On trace le système de coordonnées. On représente la variable'x en abscisse et y en ordonnée, qu'on montre dans la figure (??).*

math et num 3ELT/exemple1.png



— *On représente les contraintes. On commence par la première, on trace la droite qu'on obtient si on considère la contrainte comme égale. Elle apparaît comme le segment qui met en relation* ***A*** *et* ***B****. On reproduit le processus avec les autres contraintes.*

— *La région réalisable est l'intersection des régions délimitées aussi par l'ensemble des contraintes, que par les conditions de non-négativité des variables, c'est-à-dire, par les deux axes de coordonnés. Cette région est représentée par le polygone* ***O-F-H-G-C****, en couleur marron.*

— *Finalement, on évalue la fonction objectif* $(3x + 2y)$ *dans chacun des points (résultat qu'on recueilli dans le tableau suivant). Comme le point* ***G*** *fournit la plus grande valeur à la fonction* $f(X)$ *et l'objectif c'est de maximiser, ce point représente la solution optimale :* $f(X) = 33$ *avec* $x = 3$ *et* $y = 12$.

Sommet	Coordonnées(x, y)	$f(X)$
O	$(0, 0)$	0
C	$(0, 14)$	28
G	$(3, 12)$	33
H	$(6, 6)$	30
F	$(8, 0)$	24

3.3 Outils mathématiques

3.3.1 Formes quadratiques

Définition 13 *Soit A une matrice symétrique de taille $n \times n$ et b un vecteur de $\mathbb{R}^n$. La forme quadratique $q : \mathbb{R}^n \to \mathbb{R}$ est définie par*

$$q(x) = \frac{1}{2} x^T A x - b^T x$$

Définition 14 *Soit A une matrice symétrique de taille $n \times n$. On dit que A est semi-définie positive (SDP) et on note $A \geq 0$, lorsqu'on a*

$$x^T A x \geq 0, \forall x \in \mathbb{R}^n$$

A est définie positive (DP) et on note $A > 0$, lorsqu'on a

$$x^T A x > 0, \forall x \in \mathbb{R}^n, x \neq 0$$

On peut relier cette définition avec les valeurs propres de la matrice A.

Proposition 14 *Soit A une matrice symétrique de taille $n \times n$. On note $\lambda_i, i = 1, \ldots, n$ ses valeurs propres (réelles). On a les équivalences suivantes :*

$$A \geq 0 \Leftrightarrow \lambda_i \geq 0, i = 1, \ldots, n$$

$$A > 0 \Leftrightarrow \lambda_i > 0, i = 1, \ldots, n$$

Remarque 6 *Lorsque la matrice A est définie positive (resp. semi-définie positive), on dira que $q(x)$ est une forme quadratique définie positive (resp. semi-définie positive).*

3.3.2 Différentiabilité

Définition 15 *Soit une fonction $f : \mathbb{R}^n \to \mathbb{R}^m$ représentée dans la base canonique de $\mathbb{R}^m$ par le vecteur*

$$f(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix}$$

continue en a. On dit que f est différentiable en a s'il existe une application linéaire , notée $f'(a)$, telle que pour tout $d \in \mathbb{R}^n$ on ait

$$f(a + d) = f(a) + f'(a)d + \|d\| \epsilon(d)$$

où ϵ est une fonction continue en 0 qui vérifie $\lim_{d \to 0} \epsilon(d) = 0$ et $\|d\| = (\sum_{k=1}^{n} d_k)^{\frac{1}{2}}$. On appel $f'(a)$ la dérivée de f au point a.

Proposition 15 *Soit $f : \mathbb{R}^n \to \mathbb{R}^m$ différentiable en a. Alors*

$$f'(a)d = \lim_{t \to 0} \frac{f(a+td) - f(a)}{t}$$

Cette méthode d'estimation du gradient est souvent appelée différences finies qu'on a déjà vu dans le deuxième chapitre. La quantité $f'(a)d$ est la dérivée directionnelle de f au point a dans la direction d.
La proposition suivante fait le lien entre la matrice de $f'(a)$ et les dérivées partielles de f au point a.

1. **Calcul de la dérivée première :**

   Proposition 16 *Soit $f : \mathbb{R}^n \to \mathbb{R}^m$ différentiable en a. Alors on peut représenter la matrice de $f'(a)$ dans les bases canoniques de $\mathbb{R}^n$ et de $\mathbb{R}^m$ et on a*

   $$(f'(a)_{ij}) = \frac{\partial f_i}{\partial x_j}(a)$$

   On appelle souvent $f'(a)$ la matrice jacobienne de f au point a. Lorsque $m = 1$ on adopte une notation et un nom particuliers :
   le gradient est le vecteur noté $\nabla f(a)$ et défini par

   $$f'(a) = \nabla f(a)^T$$

   On a alors

   $$f(a+d) = f(a) + \nabla f(a)^T d + \|d\| \, \epsilon(d)$$

2. **Calcul de la dérivée seconde :** Dans le cas $m = 1$, $f : \mathbb{R}^n \to \mathbb{R}$

   Définition 16 *L'application $f : \mathbb{R}^n \to \mathbb{R}$ est dite deux fois différentiable s'il existe une matrice symétrique $\nabla^2 f(a)$ telle que*

   $$f(a+d) = f(a) + \nabla f(a)^T d + d^T \nabla^2 f(a) d + \|d\|^2 \, \epsilon(d)$$

   On appelle $\nabla^2 f(a)$ matrice hessienne de f au point a. Comme l'énonce le théorème suivant, cette matrice s'obtient à partir des dérivées secondes de f :

   Théorème 14 *Soit $f : \mathbb{R}^n \to \mathbb{R}$ deux fois différentiable en un point a. Si on note $g(x) = \nabla f(x)$, alors la matrice hessienne est définie par $\nabla^2 f(a) = g'(a)$, soit*

   $$\left(\nabla^2 f(a)_{ij}\right) = \frac{\partial^2 f_i}{\partial x_i \partial x_j}(a)$$

3.3.3 Notions de convexité

En mathématiques, le mot « convexe » est utilisé dans la désignation de deux notions bien distinctes :

- Lorsqu'il se rapporte à une forme géométrique, un ensemble de points, il renvoie au concept d'ensemble convexe .
- Lorsqu'il se rapporte à une fonction, il renvoie au concept de fonction convexe.

En économie, la convexité est un indicateur de risque de taux directement lié au concept mathématique de fonction convexe.
Un objet géométrique est dit convexe lorsque, chaque fois qu'on y prend deux points x et y, le segment [x,y] qui les joint y est entièrement contenu. Ainsi un cube plein, un disque ou une boule sont convexes (voir figure 3.1), mais un objet creux ou bosselé ne l'est pas (voir figure 3.2).

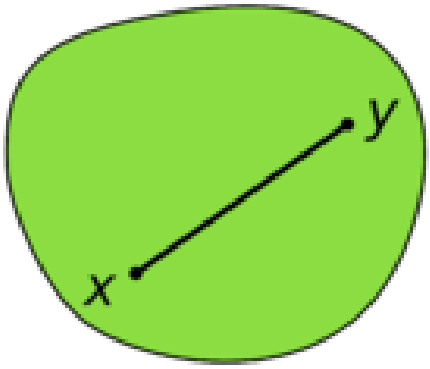


FIGURE 3.1 – ensemble convexe

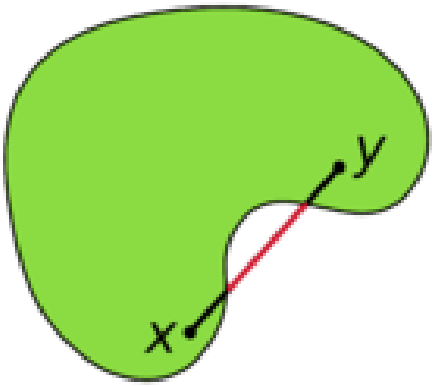


FIGURE 3.2 – ensemble non convexe

Ensemble convexe

Définition 17 *Soit x_1 et x_2 deux vecteurs de $\mathbb{R}^n$. Le segment de droite rejoignant l'extrémité de ces vecteurs, l'ensemble des points :*

$$D = \{x \in \mathbb{R}^n / x = \alpha x_1 + (1-\alpha)x_2, 0 \leq \alpha \leq 1\}$$

Exemple 6 *Considérons les deux vecteurs $x_1 = \begin{pmatrix} 5 \\ 2 \end{pmatrix}$ et $x_2 = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$.*
Soit $[S_1, S_2]$ le segment de droite reliant les extrémités de x_1 et x_2.

$$[S_1, S_2] = \left\{ x \in \mathbb{R}^2 / x = \alpha \begin{pmatrix} 5 \\ 2 \end{pmatrix} + (1-\alpha) \begin{pmatrix} 1 \\ 4 \end{pmatrix}, 0 \leq \alpha \leq 1 \right\}$$

En faisant varier la valeur de α entre 0 et 1, on obtient les points de segment :

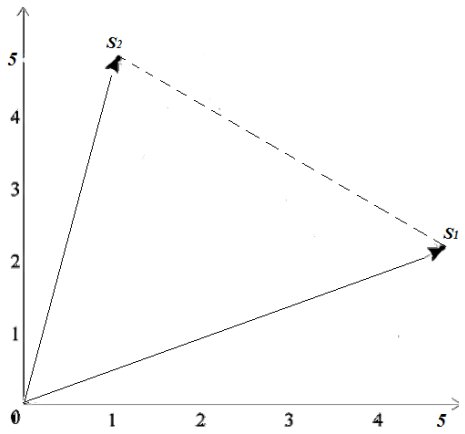
$$\alpha = 0 \qquad x = \begin{pmatrix} 1 \\ 4 \end{pmatrix} = S_2$$

$$\alpha = \frac{1}{4} \qquad x = \frac{1}{4}\begin{pmatrix} 5 \\ 2 \end{pmatrix} + \frac{3}{4}\begin{pmatrix} 1 \\ 4 \end{pmatrix} = \begin{pmatrix} 2 \\ \\ \frac{7}{2} \end{pmatrix}$$

$$\alpha = \frac{1}{2} \qquad x = \frac{1}{2}\begin{pmatrix} 5 \\ 2 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 1 \\ 4 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$$

$$\alpha = \frac{3}{4} \qquad x = \frac{3}{4}\begin{pmatrix} 5 \\ 2 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 1 \\ 4 \end{pmatrix} = \begin{pmatrix} 4 \\ \\ \frac{5}{2} \end{pmatrix}$$

$$\alpha = 1 \qquad x = \begin{pmatrix} 5 \\ 2 \end{pmatrix} = S_1$$



Fonction convexe

Définition 18 *Soit S un ensemble convexe de $\mathbb{R}^n$, $f : S \to \mathbb{R}$ est dite convexe sur s si :*

$$\forall (x, y) \in S^2, \forall \alpha \in [0, 1], f\left(\alpha x + (1 - \alpha)y\right) \leq \alpha f(x) + (1 - \alpha) f(y)$$

On dit que f est strictement convexe sur S si pour $x \neq y$:

$$\forall (x, y) \in S^2, \forall \alpha \in [0, 1], f\left(\alpha x + (1 - \alpha)y\right) < \alpha f(x) + (1 - \alpha) f(y)$$

Caractérisation de la convexité en terme de hessien

Proposition 17 *Si $f : \mathbb{R} \to \mathbb{R}$ est deux fois continûment dérivable sur un ensemble S convexe, alors f est convexe si et seulement si $f''(x) \geq 0, \forall x \in S$ et f est strictement convexe si et seulement si $f''(x) > 0, \forall x \in S$*

Ce résultat se généralise pour $n > 1$: le résultat suivant fait le lien entre le hessien et la propriété de convexité.

Théorème 15 *Si $f : \mathbb{R}^n \to \mathbb{R}^n$ est deux fois continûment dérivable sur un ensemble S convexe, alors f est convexe si et seulement si $\nabla^2 f(x) \geq 0, \forall x \in S$ et f est strictement convexe si et seulement si $\nabla^2 f(x) > 0, \forall x \in S$*

Corollaire 2 *Soit f une forme quadratique définie par*

$$f(x) = \frac{1}{2} x^T H x - b^T x$$

Alors f est convexe si et seulement si $H \geq 0$, et strictement convexe si et seulement si $H > 0$.

Cela provient du fait que $\nabla^2 f(x) = H$.
Dans le cas où la fonction f n'est supposée qu'une fois différentiable, on a le résultat suivant :

Théorème 16 *Si $f : S \subset \mathbb{R}^n \to \mathbb{R}$ une fonction une fois différentiable, alors f est convexe si et seulement si*

$$f(y) \geq f(x) + \nabla^T f(x)(y - x), \forall (x, y) \in S^2$$

La fonction f est strictement convexe si et seulement si

$$f(y) > f(x) + \nabla^T f(x)(y - x), \forall (x, y) \in S^2$$

3.3.4 Types d'extremum

Définition 19 *Un point $\bar{x} \in A$ est dit un minimiseur local de $f(x)$ s'il existe un $\epsilon > 0$ tel que*

$$f(\bar{x}) \leq f(x)$$

si

$$x \in A \text{ et } \|x - \bar{x}\| < \epsilon$$

Définition 20 *Un point $\bar{x} \in A$ est dit un minimiseur global de $f(x)$ tel que $\forall x \in A$*

$$f(\bar{x}) \leq f(x)$$

Un minimiseur global est un minimiseur local.

Exemple 7 *Soit $f(x) = x^2 + y^2 + xy + 1$*

1. *Déterminer les points critiques de f.*
2. *Étudier les extremums locaux de f.*

Solution

1. On calcul les dérivées partielles de f :

$$\frac{\partial f}{\partial x}(x, y) = 2x + y, \frac{\partial f}{\partial y}(x, y) = x + 2y$$

   Si (x, y) est un point critique de f, il vérifie donc le système

$$\begin{cases} 2x + y = 0 \\ x + 2y = 0 \end{cases}$$

   $(0, 0)$ est donc le seul point critique de f.
2. Les extremums locaux d'une fonction différentiable ne pouvant être atteints qu'en un point critique, il suffit d'étudier si $(0, 0)$ est un extremum local. En faisant un développement carré de f on obtient :

$$f(x, y) = \left(x + \frac{1}{2}y\right)^2 + \frac{3}{4}y^2 + 1 \geq 1 = f(0, 0)$$

   Ainsi, $(0, 0)$ est minimum local et même global de f.

3.3.5 Conditions nécessaires pour un minimum local

Le gradient $g(x)$ et le Hessien $H(x)$ doivent satisfaire certaines conditions en $\bar{x}$ (minimiseur local). Deux conditions vont être présentées :

1. Conditions qui vont être satisfaite en $\bar{x}$ Ce sont des conditions nécessaires.
2. Conditions qui vont garantir que $\bar{x}$ est un minimiseur local. Ce sont des conditions suffisantes.

La direction

Définition 21 *Soit $\delta = \alpha d$ une variation de x où $\alpha > 0$ et d le vecteur direction. Si A est la région admissible et une constante $\tilde{\alpha}$ existe telle que*

$$x + \alpha d \in A$$

pour tout α avec $0 \leq \alpha \leq \tilde{\alpha}$, alors d est dite direction admissible au point x.

Exemple 8 *La région admissible d'un problème d'optimisation est donnée par :*

$$A = \{x / x_1 \geq 2, x_2 \geq 0\}$$

Lequel des vecteurs $d_1 = (-2\ \ 2)^T$, $d_2 = (0\ \ 2)^T$, $d_3 = (2\ \ 0)^T$ est une direction admissible aux points $x_1 = (4\ \ 1)^T$, $x_2 = (2\ \ 3)^T$ et $x_3 = (1\ \ 4)^T$?

Solution

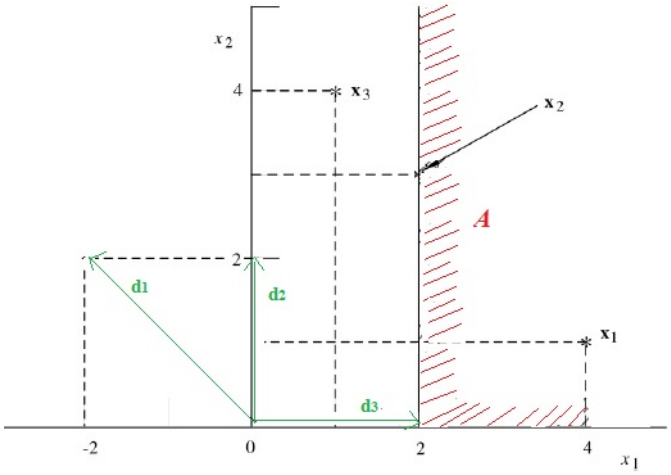


FIGURE 3.3 – les vecteurs direction

D'après la figure (3.3), on les résultats suivants :

— *d_1 est une direction admissible au point x_1 pour tout $0 \leq \alpha \leq 1$. car on a*

$$x_1 + \alpha d_1 \in A$$

$$\Leftrightarrow \begin{cases} 4 - 2\alpha \geq 2 \\ 1 + 2\alpha \geq 2 \end{cases} \Rightarrow \begin{cases} \alpha \leq 1 \\ \alpha \geq \dfrac{1}{2} \end{cases}$$

— *d_1 n'est pas une direction admissible au point x_2.*

— *d_2 est une direction admissible aux points x_1 et x_2 pour un $\tilde{\alpha} > 0$.*

— *d_3 est une direction admissible aux points x_1 et x_2 pour un $\tilde{\alpha} > 0$.*
— *d_1, d_2 et d_3 ne sont pas des directions admissibles au point x_3, car x_3 n'appartient pas à la région admissible.*

Conditions nécessaires du premier ordre

La fonction objectif doit satisfaire deux types de conditions dans le but d'avoir un minimum, dites, conditions du premier et du second ordre. Les conditions du premier ordre utilisent le gradient.

Théorème 17 *1. Si $f(x) \in C^1$ et $\bar{x}$ un minimiseur local, alors*

$$g(\bar{x})^T d \geq 0$$

pur toute direction admissible d au point $\bar{x}$.

2. Si $\bar{x}$ est localisé à l'intérieur de A, alors

$$g(\bar{x}) = 0$$

Conditions nécessaires du second ordre

Les conditions nécessaire du second ordre utilisent non seulement le gradient mais aussi le Hessien. Soit d est une direction arbitraire au point x. Rappelons que,la forme quadratique $d^T H(x)d$ est dite D.P,si $d^T H(x)d > 0$ S.D.P, si $d^T H(x)d \geq 0$, S.D.N si $d^T H(x)d \leq 0$ et D.N si $d^T H(x)d < 0$, pour tout $d \neq 0$ au point x. Si $d^T H(x)d$ admet des valeurs positives et négatives est dite indéfinie.

Théorème 18 — *Si $f(x) \in C^2$ et $\bar{x}$ un minimiseur local, alors pour toute direction admissible d au point $\bar{x}$*

1. $g(\bar{x})^T d \geq 0$

2. Si $g(\bar{x})^T d = 0$,$d^T H(\bar{x})d \geq 0$

— *Si $\bar{x}$ est localisé à l'intérieur de A, alors*

1.

$$g(\bar{x}) = 0$$

2. $d^T H(\bar{x})d \geq 0$, $\forall d \neq 0$.

Exemple 9 *Soit $\bar{x} = (\frac{1}{2}\ 0)^T$ un minimiseur local du problème*

$$\min f(X) = x^2 - x + y + xy$$

$$S.C : x \geq 2, y \geq 0$$

Montrer que les conditions nécessaires du second ordre sont vérifiées.

solution

Les dérivées partielles premières de f sont :

$$\frac{\partial f}{\partial x}(x, y) = 2x_+ y - 1, \frac{\partial f}{\partial y}(x, y) = x + 1$$

Si $d = (d_1 \ d_2)^T$ est une direction admissible, on obtient

$$g(x)^T d = (2x_+ y - 1)d_1 + (x + 1)d_2$$

au point $x = \bar{x}$

$$g(\bar{x})^T d = \frac{3}{2} d_2$$

Si $d_2 \geq 0$, on a

$$g(\bar{x})^T d \geq 0$$

Alors les conditions nécessaires du premier ordre sont vérifiées.
Si $d_2 = 0$

$$g(\bar{x})^T d = 0$$

le hessien est

$$H(\bar{x}) = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}$$

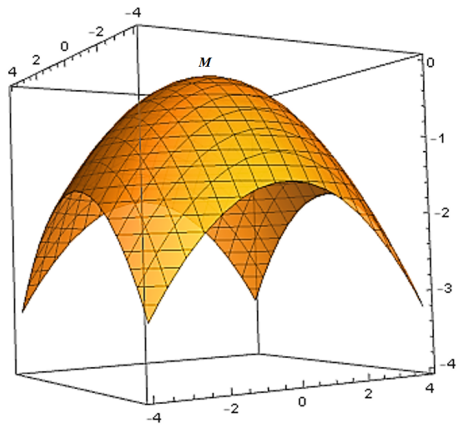
et

$$d^T H(\bar{x}) d = 2d_1^2 \geq 0$$

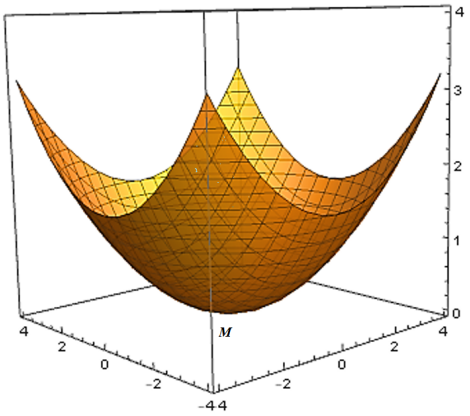
pour toute valeur de d_1, les conditions nécessaires du second ordre sont vérifiées.

3.3.6 Classification des points stationnaires

Si les points extremums appelés minimiseurs et maximiseurs, sont localisés à l'intérieur de la région admissible, ils sont appelés points stationnaires lorsque $g(x) = 0$ en ces points. Un autre type de points stationnaires est le point selle ou point col.
Dans le cas d'une fonction à deux variables $z = f(x, y)$, le graphe est une surface dans l'espace à trois dimensions. Une telle fonction présente un maximum au point $M(x_0, y_0, f(x_0, y_0))$ si $f(x_0, y_0))$ atteint une valeur supérieure à toutes celles qui prend $f(x, y)$ au voisinage de $x = x_0$ et $y = y_0$.



De même, $f(x, y)$ possède un minimum au point $M(x_0, y_0, f(x_0, y_0))$ si $f(x_0, y_0)$ atteint une valeur inférieure à toutes celles que prend $f(x, y)$ au voisinage de $x = x0$ et $y = y0$.



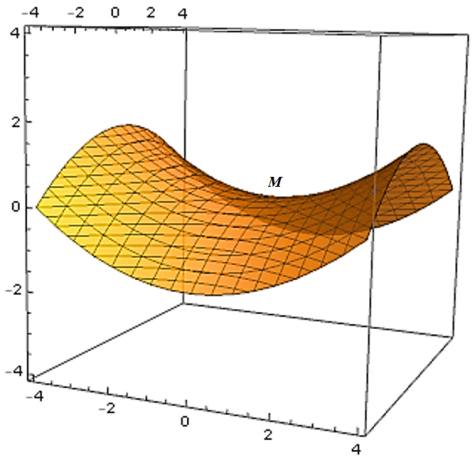
Il en résulte qu'au point $M(x_0, y_0, f(x0, y0))$, il existe un plan tangent horizontal. Ce plan tangent est engendré par deux tangentes, elles-mêmes déterminées par :

$$\frac{\partial f}{\partial x} \text{ et } \frac{\partial f}{\partial y}$$

Ainsi, la condition nécessaire à l'existence d'un extremum est la suivante

$$\left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) = (0, 0)$$

Cette condition est nécessaire mais pas suffisante. En effet, il existe des fonctions pour lesquelles $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} = 0$ sans qu'il existe un extremum en ce point. Dans ce cas, on parle de point-selle.
Il est toujours possible de trouver un point situé au-dessus du point-selle et un autre au-dessous, ceci quelque soit le voisinage du point-selle considéré. Notons encore, qu'en un point-selle la fonction présente un minimum pour l'une des variables et un maximum pour l'autre variable.



Il faut donc remplir une condition suffisante qui est la suivante :

$$D = \frac{\partial^2 f}{\partial x^2}\frac{\partial^2 f}{\partial y^2} - \left(\frac{\partial^2 f}{\partial x \partial y}\right)^2 > 0$$

Ainsi on obtient le résultat suivant :

Résulat :
Soit $M(x_0, y_0, f(x_0, y_0))$ le point en lequel :

$$\left(\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y}\right) = 0$$

Alors, si on ce point

1. $\frac{\partial^2 f}{\partial x^2} > 0$ et $D > 0$, f possède un minimum au point M.
2. $\frac{\partial^2 f}{\partial x^2} < 0$ et $D > 0$, f possède un maximum au point M.
3. Si $D < 0$, f possède ni un minimum ni un maximum au point M, mais un point-selle.
4. Si $D = 0$, on ne peut rien conclure.

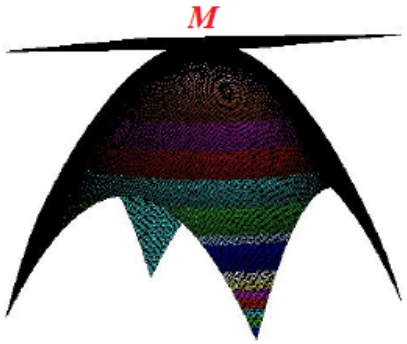
Exemple 10 *Soit la fonction* $f(x, y) = -x^2 - y^2$ *et son plan tangent d'équation* $z = x + y$. *Résolvant le système d'équations suivant :*

$$\begin{cases} \frac{\partial f}{\partial x} = 0 \\ \frac{\partial f}{\partial y} = 0 \end{cases} \Leftrightarrow \begin{cases} -2x = 0 \\ -2y = 0 \end{cases} \Leftrightarrow \begin{cases} x = 0 \\ y = 0 \end{cases}$$

Donc, il existe un point stationnaire qui $M = (0, 0, 0)$. *Pour connaître la nature de ce point, il suffit d'appliquer le résultat précédent.On trouve alors :*

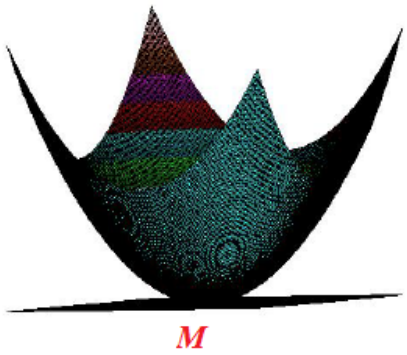
$$\frac{\partial^2 f}{\partial x^2} = -2 < 0 \text{ et } D = 4 > 0$$

Donc, $M = (0, 0, 0)$ *est un maximum comme le montre la figure suivante*



Exemple 11 *Soit la fonction* $f(x, y) = x^2 + y^2$ *et son plan tangent d'équation* $z = x + y$. *Résolvant le système d'équations suivant :*

$$\begin{cases} \frac{\partial f}{\partial x} = 0 \\ \frac{\partial f}{\partial y} = 0 \end{cases} \Leftrightarrow \begin{cases} 2x = 0 \\ 2y = 0 \end{cases} \Leftrightarrow \begin{cases} x = 0 \\ y = 0 \end{cases}$$



Donc, il existe un point stationnaire qui $M = (0, 0, 0)$. *Pour connaître la nature de ce point, il suffit d'appliquer le résultat précédent.On trouve alors :*

$$\frac{\partial^2 f}{\partial x^2} = 2 > 0 \text{ et } D = 4 > 0$$

Donc, $M = (0, 0, 0)$ *est un minimum comme le montre la figure suivante*

Exemple 12 *Soit la fonction* $f(x, y) = -x^2 + y^2$ *et son plan tangent d'équation* $z = x + y$. *Résolvant le système d'équations suivant :*

$$\begin{cases} \dfrac{\partial f}{\partial x} = 0 \\ \dfrac{\partial f}{\partial y} = 0 \end{cases} \Leftrightarrow \begin{cases} -2x = 0 \\ 2y = 0 \end{cases} \Leftrightarrow \begin{cases} x = 0 \\ y = 0 \end{cases}$$

Donc, il existe un point stationnaire qui $M = (0, 0, 0)$. *Pour connaître la nature de ce point, il suffit d'appliquer le résultat précédent.On trouve alors :*

$$\frac{\partial^2 f}{\partial x^2} = -2 < 0 \text{ et } D = -4 < 0$$

Donc, $M = (0, 0, 0)$ *n'est ni un maximum ni un minimum, c'est un point-selle comme le montre la figure suivante*

3.4 Méthodes d'optimisation unidimensionnelles

Les méthodes unidimensionnelles sont utilisées dans l'optimisation de fonctions à une seule variable. Elles peuvent être classées en deux groupes :

1. **Les méthodes de subdivision d'intervalles :**
   Elles consistent à se rapprocher de l'optimum par réductions successives de l'intervalle de recherche complet, en tenant compte de la valeur de la fonction aux extrémités de chaque sous-intervalle exploré séquentiellement.

   — La méthode la plus simple de dichotomie classique utilise un facteur de réduction constant de $\frac{1}{2}$.
   — Dans la méthode de Fibonacci, le facteur de réduction est adapté au cours de la recherche : il est défini comme le rapport de deux termes consécutifs de la suite de Fibonacci (ce facteur tend vers le nombre d'or égal à $\frac{1+\sqrt{5}}{2}$). L'intérêt de la méthode réside dans le fait qu'à chaque itération, la valeur de la fonction est évaluée en un seul point pour la détermination de l'intervalle suivant. L'autre extrémité du nouvel intervalle de recherche est déduit des points testés lors des itérations précédentes
   — La section dorée ou méthode du nombre d'or est très similaire à la méthode de Fibonacci. Elle utilise un facteur de réduction égal au nombre d'or.

   Les techniques de subdivision d'intervalles peuvent s'appliquer à des fonctions éventuellement discontinues. Elles nécessitent que ces dernières soient unimodales. Dans le cas contraire, elles ne garantissent pas la convergence vers l'optimum global. Par ailleurs, l'optimum ne peut être localisé "exactement". Il est seulement possible d'obtenir sa valeur dans un certain intervalle d'incertitude (l'intervalle obtenu à la dernière itération à la suite des subdivisions successives). La précision de la recherche est fonction de l'intervalle de départ et du nombre d'itérations. Elle peut être améliorée en augmentant le nombre de subdivisions.
2. **Les méthodes d'interpolation :**

   — Les méthodes Lagrangiennes s'efforcent de se rapprocher de l'optimum par réductions successives de l'intervalle de recherche, en interpolant la fonction par un polynôme d'ordre n.
   — La technique d'interpolation quadratique (parabolique) de la fonction par un polynôme d'ordre 2 est la plus populaire.

Dans cette section on s'intéresse à la méthode de la section dorée et la méthode d'interpolation parabolique.

3.4.1 Méthode de la section dorée

Un problème d'optimisation unidimensionnel est définit par

$$\min F = f(x)$$

où $f(x)$ est une fonction à une seule variable. Ce problème a une solution si $f(x)$ possède un seul minimum dans un intervalle considéré $[a, b]$, où a et b sont les limites inférieure et supérieure respectivement du minimiseur $\bar{x}$.
Dans les méthodes de Recherche Linéaire, $\bar{x}$ appartient à un intervalle $[a, b]$ appelé intervalle d'incertitude, le but est de réduire d'une manière itérative l'intervalle jusqu'à obtenir le plus petit intervalle $[a_n, b_n]$ qui contient $\bar{x}$.

Définition 22 *Soit f une fonction continue sur $[a,b]$. On dit que f est unimodale s'il existe un $x_* \in]a,b[$ tel que f soit strictement décroissante sur $[a,x_*]$ et strictement croissante sur $[x_*,b]$. On a donc un minimum local strict en x_* (c'est même l'unique minimum global sur $[a,b]$).*

Définition 23 *Soit f une fonction continue sur un intervalle I, et trois réels $a<c<b$ de I. On dit que le triplet (a,c,b) est admissible pour le problème de minimisation de f si on a $f(a)\geq f(c)$ et $f(b)\geq f(c)$.*

Si $f(a)=f(c)$ ou $f(b)=f(c)$ alors là le minimum il est dans a ou dans b.

Objectif de la méthode

La méthode de la section dorée consiste à s'arranger pour que la taille du triplet soit divisée d'un facteur constant à chaque étape. On s'aperçoit alors que cela contraint ce facteur à être $\varphi = \frac{1+\sqrt{5}}{2}$.
On se donne une fonction f continue sur l'intervalle $[a_0,b_0]$. On pose

$$\alpha = \frac{1}{\varphi}$$

$$c_0 = a_0 + (1-\alpha)(b_0-a_0)$$

et

$$d_0 = a_0 + \alpha(b_0-a_0)$$

On calcul $f(a_0)$,$f(b_0)$,$f(c_0)$ et $f(d_0)$ et on suppose qu'un des triplets (a_0,c_0,b_0) ou (a_0,d_0,b_0) est admissible. On définit les suites par récurrence :

— Si $f(c_n)<f(d_n)$, alors le triplet (a_n,c_n,d_n) est admissible et on pose

$$(a_{n+1},d_{n+1},b_{n+1}) = (a_n,c_n,d_n)$$

et

$$c_{n+1} = a_{n+1} + (1-\alpha)(b_{n+1}-a_{n+1})$$

On a simplement besoin de calculer $f(c_{n+1})$, puisque $f(a_{n+1})$, $f(d_{n+1})$ et $f(b_{n+1})$ sont déjà connues. Si $f(c_n)\geq f(d_n)$, alors le triplet (c_n,d_n,b_n) est admissible et on pose

$$(a_{n+1},d_{n+1},b_{n+1}) = (c_n,d_n,b_n)$$

et

$$d_{n+1} = a_{n+1} + \alpha(b_{n+1}-a_{n+1})$$

On a simplement besoin de calculer $f(d_{n+1})$, puisque $f(a_{n+1})$, $f(c_{n+1})$ et $f(b_{n+1})$ sont déjà connues.

Algorithme

poser $\varphi = \dfrac{1+\sqrt{5}}{2}$
poser $a_0 = a$
poser $b_0 = b$
pour $i = 0,\ldots,N_{max}$

poser $c' = a_i + \frac{1}{\varphi^2}(b_i - a_i)$
poser $d' = a_i + \frac{1}{\varphi}(b_i - a_i)$
Si $(f(c') < f(d'))$ alors
poser $a_{i+1} = a_i$
poser $b_{i+1} = d'$
Sinon si $(f(c') > f(d'))$ alors
poser $a_{i+1} = c'$
poser $b_{i+1} = b_i$
Sinon si $(f(c') = f(d'))$ alors
poser $a_{i+1} = c'$
poser $b_{i+1} = d'$
fin pour si
fin pour i

Ici, le N_{max} est le nombre maximal d'itérations que l'on se fixe. A cette fin, on doit valider un critère d'arrêt de la forme : $|b_{i+1} - a_{i+1}| < \epsilon$, où ϵ est l'erreur (ou tolérance) que l'on se permet sur la solution $\bar{x}$ du problème.

3.4.2 Interpolation parabolique (quadratique)

La méthode d'interpolation quadratique consiste à approximer l'expression de la fonction objectif par un polynôme du second ordre

$$p(x) = a_0 + a_1 x + a_2 x^2$$

où a_0, b_0 et c_0 sont des constantes. Soit

$$p(x_i) = f(x_i) = f_i \tag{3.4.1}$$

pour $i = 1, 2, 3$ où $[x_1, x_3]$ est l'intervalle qui contient le minimiseur de $f(x)$.
Considérons que les valeurs de f_i sont connues, ainsi a_0, a_1 et a_2 peuvent être déduite par la solution du système (3.4.1).

Interpolation en un point

La première dérivée de $p(x)$ est donnée par :

$$p'(x) = a_1 + 2a_2 x$$

Si $p'(x) = 0$ et $a_2 \neq 0$ alors le minimiseur de $p(x)$ est déduit par

$$x_* = \frac{-a_1}{2a_2}$$

En résolvant simultanément les équations du système (3.4.1), on trouve

$$a_1 = -\frac{(x_2^2 - x_3^2)f_1 + (x_3^2 - x_1^2)f_2 + (x_1^2 - x_2^2)f_3}{(x_1 - x_2)(x_1 - x_3)(x_2 - x_3)}$$

$$a_2 = \frac{(x_2 - x_3)f_1 + (x_3 - x_1)f_2 + (x_1 - x_2)f_3}{(x_1 - x_2)(x_1 - x_3)(x_2 - x_3)}$$

Ainsi

$$x_* = -\frac{(x_2^2 - x_3^2)f_1 + (x_3^2 - x_1^2)f_2 + (x_1^2 - x_2^2)f_3}{2\left[(x_2 - x_3)f_1 + (x_3 - x_1)f_2 + (x_1 - x_2)f_3\right]}$$

Si $p(x)$ est une bonne approximation de $f(x)$, alors x_* sera une bonne estimé de $\bar{x}$.

Interpolation en deux points

Ici, on considère que les valeurs de $f(x)$ et de ces premières dérivées sont connues en deux points distincts. On peut écrire :

$$p(x_1) = a_0 + a_1 x_1 + a_2 x_1^2 = f(x_1) = f_1$$

$$p(x_2) = a_0 + a_1 x_2 + a_2 x_2^2 = f(x_2) = f_2$$

$$p'(x_1) = a_1 + 2a_2 x_1 = f_1'$$

La solution de ces équations donne

$$a_1 = f_1' - \frac{2x_1\left[f_1'(x_1 - x_2) - (f_1 - f_2)\right]}{(x_1 - x_2)^2}$$

$$a_2 = \frac{f_1'(x_1 - x_2) - (f_1 - f_2)}{(x_1 - x_2)^2}$$

d'où

$$x_* = x_1 + \frac{f_1'(x_2 - x_1)^2}{2\left[f_1 - f_2 + f_1'(x_2 - x_1)\right]}$$

Maintenant si la dérivée est connue en deux points x_1 et x_2 alors

$$x_* = x_2 + \frac{f_2'(x_2 - x_1)}{f_2' - f_1'}$$

3.5 Méthodes d'optimisation multidimensionnelles

Les méthodes multidimensionnelles sont consacrées à l'optimisation de fonction à un paramètre ou plus. On peut les classés de la manière suivante :

- Elles sont dites d'ordre 0, si elles n'utilisent que la valeur de la fonction. Ces méthodes ont pour avantage de se passer du calcul du gradient surtout lorsque la fonction n'est pas différentiable ou lorsque le calcul de son gradient est complexe ou représente un coût important. Leur inconvénient et qu'elles sont peu précise et convergent très lentement vers l'optimum.
- Elles sont dites d'ordre 1, si elles nécessitent en plus le gradient de la fonction. Elles ont pour avantage d'accélérer la localisation de l'optimum, car le gradient donne une information sur la direction de recherche de la solution par contre, elles ne sont applicables qu'aux problèmes dans lesquels la fonction est continûment différentiable.
- Elles sont dites d'ordre 2, si elles utilisent le gradient et le héssien de la fonction.

Les méthodes multidimensionnelles peuvent être divisées en deux groupes :

- **les méthodes analytiques ou de descente :**se basent sur la connaissance d'une direction de recherche, souvent donnée par le gradient de la fonction. Comme exemple il y a les méthodes de descente : gradient à pas fixe ou à pas optimal, la méthode du Gradient Conjugué et les méthodes Newton et Quasi-Newton.

— **Les méthodes heuristiques (géométriques) :** elles explorent l'espace par essais successifs en recherchant les directions les plus favorables. On emploie le plus souvent la stratégie de Hooke et Jeeves, la méthode de Rosenbrock, ou la méthode du simplex de Nelder et Mead. Toutes ces techniques sont déterministes et locales mais elles sont beaucoup plus robustes que les méthodes analytiques classiques, en particulier si la fonction objectif est discontinue ou bruitée. Par contre, elles deviennent contre indiqué lorsque le nombre de paramètres est élevé.

3.5.1 Méthodes de descente

Dans cette section, on va s'intéresser aux algorithmes de calcul de minimum et plus particulièrement aux algorithmes de descente. Partant d'un point x_0 arbitrairement choisi, un algorithme de descente va chercher à générer une suite d'itérés $(x_k)_{k\in\mathbb{N}}$ telle que

$$\forall k \in \mathbb{N}, f(x_{k+1}) \leq f(x_k)$$

Définition 24 *Soit $f : \mathbb{R}^n \to \mathbb{R}$ une application continue. Soit $x \in \mathbb{R}^n$ et $d \in \mathbb{R}^n$. La dérivée directionnelle en x dans la direction de d est définie par*

$$df(x;d) := \lim_{t\to 0^+} \frac{f(x+td) - f(x)}{t}$$

si cette limite existe.

Proposition 18 *Si f est différentiable en un point $x \in \mathbb{R}^n$ alors pour tout $d \neq 0$, f admet une dérivée dans la direction d en x et*

$$df(x;d) = Df(x)(d) = \nabla f(x)^T d$$

La dérivée directionnelle donne des informations sur la pente de la fonction dans la direction d, tout comme la dérivée donne des informations sur la pente des fonctions à une variable. En particulier,

— *Si $df(x;d) > 0$, alors f est croissante dans la direction d.*
— *Si $df(x;d) < 0$, alors f est décroissante dans la direction d.*

Dans ce dernier cas, on dira que d est une direction de descente de f.

Direction de descente

Définition 25 *Soit $f : \mathbb{R}^n \to \mathbb{R}$ et $x \in \mathbb{R}^n$ et $d \in \mathbb{R}^n$. Le vecteur $d \in \mathbb{R}^n$ est une direction de descente pour f à partir du point x si $t \to f(x+td)$ est décroissante en $t = 0$, c'est-à-dire il existe un $\alpha > 0$ tel que*

$$\forall 0 < t < \alpha, f(x+td) < f(x)$$

$$df(x;d) := \lim_{t\to 0^+} \frac{f(x+td) - f(x)}{t}$$

si cette limite existe.

Parmi toutes les directions de descente existantes en un point x donné, il est naturel de s'intéresser à celle où la pente est la plus forte. Un résultat remarquable montre que cette direction est donnée par l'opposé du gradient.

Proposition 19 *Soit f une fonction différentiable en un point $x \in \mathbb{R}^n$ alors pour toute direction $d \neq 0$ de norme constante égale à $\|d\| = \|\nabla f(x)\|$, on a*

$$(-\nabla f(x))^T \nabla f(x) \leq d^T \nabla f(x)$$

Algorithme général pour les méthodes de descente

Données : f supposé au moins différentiable, x_0 point initial arbitrairement choisi. Sortie : une approximation de la solution du problème :

$$\min_{x \in \mathbb{R}^\kappa} f(x)$$

1. $k = 0$
2. tant que le test de convergence n'est pas satisfait
   — Trouver une direction de descente d_k telle que
   $$\nabla f(x_k)^T d < 0$$
   — Choisir un pas $\alpha_k > 0$ à faire dans la direction d_k tel que :
   $$f(x_k + \alpha_k d_k) \leq f(x_k)$$
   — On pose $x_{k+1} = x_k + \alpha_k d_k$; $k = k + 1$
3. Retourner x_k

Tests d'arrêts

Soit $\bar{x}$ un point de minimum local du critère f à optimiser. En pratique, un test d'arrêt devra être choisi pour garantir que l'algorithme s'arrête toujours après un nombre fini d'itérations et que le dernier point calculé soit suffisamment proche de $\bar{x}$.
Soit $\epsilon > 0$ la précision demandée. Plusieurs critères sont à notre disposition : tout d'abord un critère d'optimalité basé sur les conditions nécessaires d'optimalité du premier ordre : on teste si

$$\|\nabla f(x_k)\| < \epsilon$$

auquel l'algorithme s'arrête et fournit l'itéré courant x_k comme solution.
En pratique, le test d'optimalité n'est pas toujours satisfait et on devra faire appel à d'autres critères comme :

$$\|x_{k+1} - x_k\| < \epsilon \|x_k\|$$

$$|f(x_{k+1}) - f(x_k)| < \epsilon |x_k|$$

ou fixé le seuil du nombre minimal d'itération à l'avance ($k < k_{\max}$).

La convergence

Il est important de garantir la convergence d'un algorithme sous certaines hypothèses. Étudier la convergence d'un algorithme, c'est étudier la convergence de la suite des itérés générés par l'algorithme.

Définition 26 *Soit un algorithme itératif qui génère une suite* $(x_k)_{k\in\mathbb{N}}$ *dans* $\mathbb{R}^n$ *afin de résoudre le problème :*

$$\min_{x\in\mathbb{R}^\varkappa} = f(x)$$

où $f : \mathbb{R}^n \to \mathbb{R}$ *est une application de classe* C^1*. L'algorithme est dit globalement convergent si quel que soit le point initial* $x_0 \in \mathbb{R}^n$

$$\lim_{k\to+\infty} \|\nabla f(x_k)\| = 0$$

Cette propriété garantit que le critère d'arrêt $\|\nabla f(x_k)\| \leq \epsilon$ sera satisfait à partir d'un certain rang quelle que soit la précision $\epsilon > 0$ demandée.

Méthode du gradient

La méthode du gradient fait partie des classes de méthodes dites de descente. Soit $x_k \in \mathbb{R}^n$ l'itéré courant. Étant donnés la valeur $f(x_k)$ et le gradient $\nabla f(x_k)$, on remplace f au voisinage de x_k par son développement de Taylor au premier ordre :

$$f(x_k + d) \approx f(x_k) + \nabla f(x_k)^T d$$

On voudrait que la dérivée directionnelle $\nabla f(x_k)^T d$ d soit la plus petite possible dans un voisinage de $d = 0$. On cherche donc à résoudre :

$$\min_{d\in\mathbb{R}^\varkappa} \nabla f(x_k)^T d, \quad S.C\, \|d\| = \|\nabla f(x_k)\|$$

dont la solution nous est donné par

$$d_k = -\nabla f(x_k)$$

Le choix de la direction de plus forte descente définit une famille d'algorithmes appelés algorithmes de descente de gradient dont le schéma est le suivant :

Données : f, x_0 première approximation de la solution cherchée, $\epsilon > 0$ précision demandée. Sortie : une approximation x_* de la solution du problème $\nabla f(x) = 0$

1. $k = 0$
2. tant que le test de convergence n'est pas satisfait
   - Trouver une direction de descente d_k telle que
   $$d_k = -\nabla f(x_k)$$
   - Choisir un pas $\alpha_k > 0$ à faire dans la direction d_k tel que :
   $$f(x_k + \alpha_k d_k) \leq f(x_k)$$
   - On pose $x_{k+1} = x_k + \alpha_k d_k$; $k = k + 1$
3. Retourner x_k

Il reste maintenant à définir une stratégie de calcul du pas. Nous étudions ici en première approche une méthode à pas optimal, puis une à pas fixe.

1. **Gradient à pas optimal :**
   Une idée naturelle consiste à suivre la direction de plus forte descente et à

faire un pas qui rende la fonction à minimiser la plus petite possible dans cette direction. Cette méthode est appelée méthode de gradient à pas optimal ou encore méthode de plus forte pente.
On remplace dans l'algorithme de descente du gradient l'étape : Trouver une direction de descente d_k telle que $d_k = -\nabla f(x_k)$, par Calculer un pas optimal α_k solution de :

$$\min_{\alpha>0} f(x_k + \alpha d_k)$$

la résolution du problème de minimisation unidimensionnel de cette étape, même de façon approchée, coûte cher en temps de calcul. Pour ces raisons, on peut lui préférer parfois l'algorithme de gradient à pas constant (ou à pas fixe).

2. **Gradient à pas fixe :**
L'idée est très simple : on impose une fois pour toutes, la taille du pas effectué selon la direction de descente calculée à chaque itération. Les étapes : Choisir un pas $\alpha_k > 0$ à faire dans la direction d_k tel que : $f(x_k + \alpha_k d_k) \leq f(x_k)$ et $x_{k+1} = x_k + \alpha_k d_k$; $k = k + 1$ de l'algorithme de descente de gradient sont alors remplacées par :

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$

3. **Méthode de la plus forte pente avec Hessien :**
Si le Hessien de f existe et peut être calculé, la valeur α qui minimise $f(x_k+\alpha d_k)$ peut être déterminée en utilisant le développement de Taylor à l'ordre 2 et en posant $\nabla f(x_k) = g_k$, on obtient

$$f(x_k + \alpha d_k) \approx g_k + (\alpha d_k)^T g_k + \frac{1}{2}(\alpha d_k)^T H(x_k)(\alpha d_k)$$

et si d_k est la direction de la plus forte pente c'est-à-dire : $d_k = -g_k$, on obtient

$$f(x_k - \alpha g_k) \approx g_k - \alpha g_k^T g_k + \frac{1}{2}\alpha^2 g_k^T H(x_k) g_k$$

En dérivant par rapport à α, on obtient

$$\frac{df(x_k + \alpha d_k)}{d\alpha} \approx -g_k^T g_k + \alpha g_k^T H(x_k) g_k$$

En posant ce résultat égal à zéro, on obtient

$$\alpha = \alpha_k \approx \frac{g_k^T g_k}{g_k^T H(x_k) g_k}$$

Alors

$$x_{k+1} = x_k - \frac{g_k^T g_k}{g_k^T H(x_k) g_k} g_k$$

Méthode du gradient conjugué

1. **Méthode du gradient conjugué linéaire :**
Soit $f(x) = \frac{1}{2}x^T Ax - bx$ avec A une matrice symétrique définie positive. On sait alors qu'il existe un unique minimum sur $\mathbb{R}^n$ donné par $x_* = A^{-1}b$.
Dans la suite, on note $r(x) = Ax - b = \nabla f(x)$ le résidu.

Définition 27 *Les vecteurs non nuls* $\{d_1, \cdots, d_p\}$ *sont dits conjugués par rapport à la matrice* A *si*

$$\text{pour tout } i \neq j \in \{1, \cdots, p\}, d_i^T A d_j = 0$$

Lemme 1 *Soient* $A \in \mathcal{M}_n(\mathbb{R})$ *symétrique définie positive,* $k \in \mathbb{N}$ *et* $\mathcal{F} = \{d_i\}_{1 \leq i \leq k}$ *une famille de vecteurs A-conjuguée. alors*

$$\begin{cases} \mathcal{F} \text{ une famille libre si } k \leq n \\ \text{Si } k = n, \mathcal{F} \text{ est une base} \end{cases}$$

Définition 28 *Soit* $A \in \mathcal{M}_n(\mathbb{R})$ *symétrique définie positive et* b *un vecteur de* $\mathbb{R}^n$. *La méthode du gradient conjugué linéaire consiste à construire une famille de directions de desecnte A-conjuguées pour résoudre le système linéaire* $Ax - b = 0$ *en minimisant la forme quadratique*

$$q(x) = \frac{1}{2} x^T A x - bx$$

Principe de La méthode :
Le but est de construire une suite d'itérés $(x^{(k)})$ qui converge vers la solution du problème d'optimisation

$$\min q(x)$$

avec $q(x) = \frac{1}{2} x^T A x - bx$,$A \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive et b un vecteur de $\mathbb{R}^n$.
Notons que la suite $x^{(k)}$ est définie par

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}$$

$\alpha^{(k)}$ est une suite de pas qui pour chaque itéré k

$$\alpha^{(k)} = \frac{r^{(k)} d^{(k)}}{d^{(k)^T} A d^{(k)}}$$

avec

$$r^{(k)} = -\nabla q(x^{(k)}) = b - A x^{(k)}$$

et $d^{(k)}$ une suite de directions de descente A-conjuguées définies en chaque itération par

$$d^{(k+1)} = r^{(k+1)} - \beta^{(k+1)} d^{(k)}$$

avec

$$\beta^{(k+1)} = \frac{r^{(k)^T} A d^{(k)}}{d^{(k)^T} A d^{(k)}}$$

Algorithme de la méthode du gradient conjugué linéaire :

Données : A matrice symétrique D. P, b vecteur, x_0 point initial arbitrairement choisi.
Sortie : une approximation de la solution du problème : $Ax = b$

(a) $k = 0$
$r_0 = b - Ax_0$

(b) tant que $\left\|x^{(k)}\right\| > \epsilon$

— Calculer
$\alpha^{(k+1)} = \dfrac{r^{(k)}d^{(k)}}{d^{(k)^T}Ad^{(k)}}$
$r^{(k+1)} = r^{(k)} - \alpha^{(k+1)}Ad^{(k)}$
$\beta^{(k+1)} = \dfrac{r^{(k)^T}Ad^{(k)}}{d^{(k)^T}Ad^{(k)}}$
$d^{(k+1)} = r^{(k+1)} - \beta^{(k+1)}d^{(k)}$
— On pose $x^{(k+1)} = x^{(k)} + \alpha^{(k)}d^{(k)}$; $k = k + 1$
fin

(c) **Retourner** $x^{(k)}$

2. **Méthode du gradient conjugué non-linéaire :**

Définition 29 *La méthode du gradient conjugué non-linéaire consiste à construire une famille de directions de descente conjuguée pour résoudre le problème d'optimisation.*

Théorème 19 *Soit $k \in \mathbb{N}$. alors*

(a) Le résidu $r^{(k)}$ est orthogonale aux directions $d^{(i)}$ pour tout $i \in \{0, 1, \cdots, k-1\}$

(b) Le résidu $r^{(k)}$ est orthogonale à la famille $\left\{r^{(0)}, r^{(1)}, \cdots, r^{(k-1)}\right\}$

(c)

$$r^{(k)^T}d^{(k)} = -r^{(k)^T}r^{(k)}$$

Principe de La méthode :
On reprend l'algorithme du gradient conjugué linéaire avec les changements suivants :

— La suite des résidus $r^{(k)}$ sera définie à l'itération k par :

$$r^{(k)} = -\nabla f(x^{(k)})$$

— En utilisant le théorème (19) on remarque que le terme général de la suite $\beta^{(k)}$ s'écrit aussi :

$$\beta^{(k+1)} = \frac{r^{(k+1)^T}r^{(k+1)}}{r^{(k)^T}r^{(k)}}$$

— Un autre changement a été proposé par **Polak-Ribière**, pour le calcul des termes de la suite $\beta^{(k)}$

$$\beta^{(k+1)} = \frac{\nabla f(x^{(k+1)^T})\left(\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})\right)}{\nabla f(x^{(k)})\nabla f(x^{(k)})}$$

Algorithme de la méthode du gradient conjugué non-linéaire :

Données : f supposée au moins différentiable, x_0 point initial arbitrairement choisi.
Sortie : une approximation de la solution du problème :

$$\min_{x\in\mathbb{R}^n} f(x)$$

(a) $k = 0$

(b) tant que $\left\|\nabla f(x^{(k)})\right\| > \epsilon$

— Calculer
$\alpha^{(k)} = \min\left(f(x^{(k)}) + \alpha d^{(k)}\right)$
$r^{(k+1)} = r^{(k)} - \alpha^{(k+1)} A d^{(k)}$
$\beta^{(k+1)} = \dfrac{\nabla f(x^{(k+1)^T})\left(\nabla f(x^{(k+1)}) - \nabla f(x^{(k)})\right)}{\nabla f(x^{(k)})\nabla f(x^{(k)})}$
$d^{(k+1)} = -\nabla f(x^{(k+1)}) + \beta^{(k+1)} d^{(k)}$
— On pose $x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}$; $k = k + 1$
fin

(c) **Retourner** $x^{(k)}$

Méthode de Newton

La méthode repose sur le fait que f soit de classe C^2, admet un minimum en un point x_* et que $\nabla^2 f = H(x)$ (la matrice hessienne) soit définie positive dans un voisinage V de x_*. Alors en utilisant un développement de Taylor d'ordre 2 de la fonction au voisinage de son minimum x_*, on obtient :

$$f(x_*) \approx f(x_*) + \nabla f(x^{(k)})^T (x - x^{(k)}) + \frac{1}{2}(x_* - x^{(k)})\nabla^2 f(x^{(k)})(x_* - x^{(k)})$$
$$= q_k(x_*), \forall k \in \mathbb{N}, \text{ tel que } x^{(k)} \in V(x_*)$$

f est de classe C^2 donc x_* est un point critique de f par conséquent il est un point critique pour la forme quadratique q_k, d'où

$$\nabla q_k(x_*) = \nabla f(x^{(k)}) + \nabla^2 f(x^{(k)})(x_* - x^{(k)}) = 0$$

ce qui donne

$$x_* = x^{(k)} - \left(\nabla^2 f(x^{(k)})\right)^{-1} \nabla f(x^{(k)})$$

Proposition 20 *Soit f une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$ de classe C^2 et soit $x \in D_f$ tel que $\nabla^2 f(x) > 0$. Alors $d = -\left(\nabla^2 f(x^{(k)})\right)^{-1} \nabla f(x^{(k)})$ est une direction de descente.*
D'où la suite des itérés $(x^{(k)})$ définissant l'algorithme de Newton sera donnée par :

$$x^{(k+1)} = x^{(k)} + d^{(k)}$$

avec $\alpha^{(k)} = 1$ et $d^{(k)}$ est l'unique solution de l'équation

$$\nabla^2 f(x^{(k)}) d^{(k)} = -\nabla f(x^{(k)})$$

Remarque 7 — *Le pas de la méthode de Newton peut être pris égal à 1 ou peut être choisi par la recherche linéaire.*
— *Lorsque le hessien $H(x^{(k)})$ n'est pas défini positif, la direction de déplacement $d^{(k)}$ dans la méthode de Newton peut ne pas être une direction de descente.*

Algorithme de la méthode de Newton :

Données : f supposée au moins différentiable, x_0 point initial arbitrairement choisi.
Sortie : une approximation de la solution du problème :

$$\min_{x\in\mathbb{R}^n} f(x)$$

1. $k = 0$
   $\alpha_0 > 0$
2. tant que $\left\|x^{(k+1)} - x^{(k)})\right\| > \epsilon$

   — Calculer
   $d^{(k)} = -\left(\nabla^2 f(x^{(k)})\right)^{-1} \nabla f(x^{(k)})$
   — Recherche linéaire : choix du pas $\alpha_k > 0$
   — On pose $x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}$; $k = k + 1$
   fin
3. **Retourner** $x^{(k)}$