Première partie

Optimisation

Chapitre I

Premiers exemples de problèmes d'optimisation et vocabulaire.

I.1 Motivations.

Les problèmes d'optimisation apparaîssent dans de nombreux domaines. Un ingénieur qui doit construire un bâtiment se doit de choisir les meilleurs matériaux au meilleur rapport qualité-prix afin d'avoir une structure aussi sûre que possible le moins cher possible. Un courtier doit choisir les bons investissements qui génèreront les meilleurs rentes tout en gardant le risque de perte sous contrôle (à son minimum). Un système physique va avoir tendance à se mettre dans un état d'énergie minimale tout en respectant les contraintes physiques qui peuvent lui être données. On pourrait même imaginer un étudiant qui cherche à minimiser le nombre d'heures à passer sur son cours pour avoir une bonne note [1].

Pour toutes ces situations, on peut trouver des points communs :

— Il y a un but ou un **objectif** à atteindre pour l'activité concernée (rendre meilleur), via un **coût** à minimiser.

— On cherche à atteindre cet objectif au regard potentiellement de toutes une série de **contraintes** ou d'attendus qu'il faut atteindre et/ou satisfaire.

— Implicitement, on suppose qu'il y a un ou des choix possibles de paramètres pour satisfaire aux deux items précédents. Ces paramètres seront appelés **variables d'optimisation ou de design**. On devra sélectionner les variables importantes d'un problème d'optimisation à considérer. Ainsi, alors que le temps qu'il fait peut influer sur la construction d'un bâtiment (ou les révisions d'un cours...), ce n'est pas forcément le cas pour le banquier. La variable "temps qu'il fait" peut donc ou non influencer l'objectif.

En résumé : **la formulation d'un problème d'optimisation** implique de définir et **comprendre** correctement le problème posé et de savoir le **formaliser** en une série de problèmes mathématiques pour espérer le **résoudre**. Plus précisément, la formalisation d'un problème d'optimisation va mettre en jeu :

— Le choix d'une ou plusieurs *variables d'optimisation* vivant dans un espace général à déterminer,

— Définir une *fonction objectif* (ou *fonction coût*),

1. même si la solution miracle n'existe pas, il faut travailler son cours et ses TDs/TPs !

— Le cas échéant, identifier l'ensemble des *contraintes*.

Dans la suite, donnons quelques exemples de tels problèmes.

I.2 Premiers exemples simples

I.2.1 Révision des examens

Un étudiant doit réviser pour ses examens et il cherche à optimiser le nombre d'heures passées à réviser pour obtenir la meilleure note possible. Il a 4 UE à passer et a une semaine pour réviser, ce qui lui donne 42 heures de travail (6 jours et 7h de travail par jour). Les variables d'optimisation sont le nombre d'heures de travail par matières (il y en a donc 4) : pour $i \in \{1, .., 4\}$, on note x_i, le nombre d'heures de révisions pour la matière numéro i. L'étudiant fait face à des contraintes : la somme des heures travaillées doit être au plus égale à 42 et il doit passer un nombre d'heures positive sur chaque UE.

L'ensemble des contraintes peut ainsi se formaliser à l'aide d'un ensemble K :

$$K = \left\{ x = (x_1, x_2, x_3, x_4) \in \mathbb{R}^4, \text{ tel que } \forall i \in \{1, .., 4\}, x_i \geq 0, \text{ et } \sum_{i=1}^{4} x_i \leq 42 \right\}.$$

Pour $x \in \mathbb{R}^4$, on note $M(x)$ la moyenne des notes (sur 20) obtenues par l'étudiant après avoir révisé x_i heures sur la matière i. L'objectif est de maximiser M ou encore de minimiser $20 - M$.

Le problème d'optimisation peut donc s'écrire :

$$\text{Trouver } x^* \in K \text{ tel que } 20 - M(x^*) = \inf_{x \in K}(20 - M(x)).$$

I.2.2 Consommation des ménages.

On considère un ménage qui peut consommer n types de marchandises (avec $n \in \mathbb{N}^*$) dont les prix forment un vecteur $p \in \mathbb{R}^n_+$. Son revenu à dépenser est un réel $b > 0$. Ce ménage cherche à optimiser la quantité de chaque marchandise achetée pour en tirer le maximum de contentement sans dépasser son budget. Ses choix de consommation sont supposés être modélisés par une fonction d'utilité u de $\mathbb{R}^n_+$ dans $\mathbb{R}$ (croissante et concave), qui mesure le bénéfice que le ménage tire de la consommation de la quantité $x = (x_1, ..., x_n)$ des n marchandises. La consommation du ménage sera le vecteur x^* qui réalisera le maximum de

$$\max_{x \in \mathbb{R}^n_+, \langle x, p \rangle \leq b} u(x),$$

où $\langle x, p \rangle = \sum_{i=1}^n x_i p_i$ (ici $\langle ., . \rangle$ représente le produit scalaire canonique de $\mathbb{R}^n$).

Autrement dit c'est le vecteur qui maximise l'utilité sous une contrainte de budget maximal.

I.2.3 Problème du toboggan

On se donne deux points A et B (avec le point A plus haut que le point B qui est au sol), et on se demande quelle serait la forme d'un toboggan qui permettrait d'atteindre le point B en partant du

point A, le plus rapidement possible. Mathématiquement, on cherche donc à déterminer une courbe du plan reliant A et B, telle qu'un point matériel de poids donné (donc sous l'action de la gravité) glissant le long de cette courbe partant de A arrive en B en un temps minimal. Le tobbogan est donc représenté par le graphe d'une fonction $f : [0,1] \to \mathbb{R}$, avec $f(0) = h > 0$ (hauteur du toboggan donnée) et $f(1) = 0$. On prend donc $A = (0,h)$ et $B = (1,0)$.

En écrivant la conservation de l'énergie, on trouve que, pour une fonction f donnée, le temps de descente $T(f)$ du toboggan est donné par la formule

$$T(f) = \int_0^1 \frac{\sqrt{1+f'(x)^2}}{2g(h-f(x))}dx,$$

avec g la constante de gravitation.

Le problème revient donc à minimiser T avec f dans un certain espace de fonctions $\mathcal{F}$: $\min_{f\in\mathcal{F}} T(f)$.

I.2.4 Identification de paramètres

On se donne un signal $f : \mathbb{R} \to \mathbb{R}$ dont on sait qu'il dépend de 3 paramètres $(a,b,c) \in \mathbb{R}^3$ et qui a la forme :

$$f(t) = a\cos(bt+c). \tag{I.1}$$

Pour identifier ces paramètres, on ne dispose que d'un nombre fini $m \in \mathbb{N}^*$ de mesures $(t_i, y_i)_{i\in\{1,..,m\}}$. On choisit alors de minimiser une fonctionnelle $\mathcal{J}$ représentant une certaine distance aux mesures :

$$\mathcal{J}(a,b,c) = \sum_{i=1}^{m}(y_i - f(t_i))^2.$$

C'est ce qui s'appelle calibrer les paramètres.

I.3 Mise en forme mathématique générale et éléments de vocabulaire

I.3.1 Forme générale d'un problème d'optimisation

Tous les exemples ci-dessus rentrent dans un cadre assez général sous lequel on peut écrire un problème d'optimisation. On notera :

- V l'espace dans lequel le problème est posé et où vont vivre les variables d'optimisation (i.e. où seront à chercher les variables d'optimisation). On cherchera V comme un espace vectoriel normé muni d'une norme que l'on notera $\|\cdot\|$.
- On se donne aussi K qui définira l'ensemble des contraintes qui s'appliqueront sur le système.
- Enfin, on se donne $\mathcal{J}$ le critère à minimiser que l'on appellera fonction objectif ou fonction coût ou fonctionnelle, avec $\mathcal{J} : K \subset V \to \mathbb{R}$.

Ce sont des ***données du problème***.

Remarque importante : *Ensuite, on peut toujours se ramener à un problème de minimisation. En effet, si le problème consiste à maximiser une certaine fonction coût $\mathcal{J}$, alors étudier le problème de maximisation de $\mathcal{J}$ revient à étudier le problème de minimisation de $-\mathcal{J}$).*

Ainsi, dans la suite, on ne considèrera que des problèmes de minimisation pour développer les stratégies de résolution numérique.

Un problème d'optimisation *Opt* peut alors se mettre sous la forme générale très simple suivante : on cherche à savoir s'il existe un u^* dans K qui réalise le minimum de $\mathcal{J}$ sur K, $\min_{u\in K} \mathcal{J}(u)$.

Cela peut encore s'écrire : *Trouver $u^* \in K$ tel que*

$$\mathcal{J}(u^*) = \inf_{u\in K} \mathcal{J}(u).$$

L'inconnue du problème est donc u à chercher dans l'ensemble K.

Le point $u^* \in K$, s'il existe, est alors appelé *point de minimum* de $\mathcal{J}$ sur K. On note encore :

$$\mathcal{J}(u^*) = \min_{u\in K} \mathcal{J}(u).$$

Il se pose alors plusieurs questions mathématiques qui se traduisent sur le problème concret de départ.

(A) Parle-t-on de min ou d'inf ? Y a-t-il existence d'au moins une solution u^* à ce problème ?
(B) Si une solution existe est-elle unique ?
(C) A-t-on une caractérisation du ou des minima, s'ils existent,
(D) Peut-on trouver cette (ces) solution(s) minimale(s) quand elle(s) existe(nt) ? Est-on capable de donner l'expression de cette solution ?

Suivant les cas, la réponse est plus ou moins simple, positive ou négative, voire même parfois inconnue !

Pour les exemples précédents, on peut identifier les données V, K, $\mathcal{J}$.

I.3.2 Définitions et éléments de vocabulaire.

Soit V un espace vectoriel muni d'une norme $\|\cdot\|$ et K un sous-ensemble de V ($V \subset K$). On considère $\mathcal{J} : K \subset V \to \mathbb{R}$. On s'intéresse au problème de minimisation associé à $\mathcal{J}$.

On distingue les minima locaux des minima globaux.

Definition I.3.1 *On dit que u est un point de minimum local de $\mathcal{J}$ sur K si*

$$u \in K \text{ et } \exists \delta > 0,\ \forall v \in K,\ \|v - u\| < \delta \Rightarrow \mathcal{J}(v) \geq \mathcal{J}(u).$$

On dit alors que la valeur $\mathcal{J}(u)$ est un minimum local de $\mathcal{J}$ sur K.

Definition I.3.2 *On dit que u est un point de minimum global de $\mathcal{J}$ sur K si*

$$u \in K \text{ et } \forall v \in K, \mathcal{J}(v) \geq \mathcal{J}(u).$$

On dit alors que la valeur $\mathcal{J}(u)$ est un minimum global de $\mathcal{J}$ sur K.

Definition I.3.3 *On appelle infimum de $\mathcal{J}$ sur K et on note $\inf_{u\in K}(\mathcal{J}(u))$:*
— *la borne inférieure dans $\mathbb{R}$ de l'ensemble $\{\mathcal{J}(u),\ u \in K\}$, si K est non vide et $\mathcal{J}$ est minorée.*
— *Si $\mathcal{J}$ n'est pas minorée sur K et K est non vide, alors l'infimum est $-\infty$.*
Si K est vide, par convention, l'infimum est $+\infty$.

Definition I.3.4 *Supposons que K est non vide. Une suite minimisante de $\mathcal{J}$ dans K est une suite $(u_n)_{n\in\mathbb{N}}$ telle que*

$$\forall n \in \mathbb{N},\ u_n \in K \text{ et } \lim_{n\to+\infty} \mathcal{J}(u_n) = \inf_{v\in K} \mathcal{J}(v).$$

Par définition même de l'infimum, **une suite minimisante existe toujours**.

En effet, commençons par le cas où $\mathcal{J}$ est minorée et donc $\inf_{v\in K} \mathcal{J}(v)$ est un réel donné. Soit $n \in \mathbb{N}^*$. Par définition de l'infimum, $\inf_{u\in K} \mathcal{J}(u) + \frac{1}{n}$ n'est plus un minorant de $\mathcal{J}$ sur K (puisque $\inf_{u\in K} \mathcal{J}(u) + \frac{1}{n} > \inf_{u\in K} \mathcal{J}(u)$ et que $\inf_{u\in K} \mathcal{J}(u)$ est la borne supérieure des minorants de $\mathcal{J}$ sur K). Donc il existe $u_n \in K$ tel que $\mathcal{J}(u_n) < \inf_{u\in K} \mathcal{J}(u) + \frac{1}{n}$. On a donc pour tout $n \in \mathbb{N}$,

$$\inf_{u\in K} \mathcal{J}(u) \leq \mathcal{J}(u_n) < \inf_{u\in K} \mathcal{J}(u) + \frac{1}{n}.$$

On en déduit le résultat par encadrement en passant à la limite lorsque $n \to +\infty$: la suite $(u_n)_{n\in\mathbb{N}^*}$ est bien une suite minimisante.

Si $\mathcal{J}$ n'est pas minorée, alors $\inf_{v\in K} \mathcal{J}(v) = -\infty$. Et on sait donc que pour tout $m \in \mathbb{R}$, il existe $v \in K$ tel que $\mathcal{J}(v) < m$. Soit alors $n \in \mathbb{N}$. En appliquant ce qui précède à $m = -n$, on en déduit qu'il existe un point $v_n \in K$ tel que $\mathcal{J}(v_n) < -n$. En passant à la limite $n \to +\infty$, on a donc bien le résultat voulu : la suite $(v_n)_{n\in\mathbb{N}}$ est bien une suite minimisante.

I.3.3 Quelques points ou caractéristiques importants.

L'espace K et type de contraintes.

Si K est égal à V, on parlera de problème d'*optimisation sans contraintes* (ou problème d'*optimisation libre*). Sinon, on parlera de problème d'*optimisation avec contraintes.*

Très souvent l'espace K sera donné par une série d'équations ou d'inéquations et de la forme :

— *Contraintes d'égalités.*
$$K := \{u \in V,\, h_i(u) = 0, i = 1, .., p\}\,,$$
pour $p \in \mathbb{N}^*$ donné et $h_i : V \to \mathbb{R}$ données, $i = 1, .., p$. On parlera alors d'un cas de problème d'*optimisation avec contraintes d'égalité.* Il y a alors $p \in \mathbb{N}^*$ contraintes données par les fonctions h_i, $i = 1, .., p$.

— *Contraintes d'inégalités.*
$$K := \{u \in V,\, g_i(u) \leq 0, i = 1, .., q\}\,,$$
pour $q \in \mathbb{N}^*$ donné et $g_i : V \to \mathbb{R}$ données, $i = 1, .., q$. On parlera alors d'un cas de problème d'*optimisation avec contraintes d'inégalité.* Il y a alors $q \in \mathbb{N}^*$ contraintes données par les fonctions g_i, $i = 1, .., q$.

— *Contraintes mixtes.*
$$K := \{u \in V,\, h_i(u) = 0, i = 1, .., p,\, g_i(u) \leq 0, i = 1, .., q\}\,,$$
pour $p \in \mathbb{N}^*$ et $q \in \mathbb{N}^*$ donnés.

Si V est un espace de dimension finie, on parlera de problème d'*optimisation en dimension finie.*

Cas particulier où $V = \mathbb{R}^d$.

On se place dans le cadre qui sera **le cadre majoritaire du cours** ; c'est le cas où $V = \mathbb{R}^d$ pour $d \in \mathbb{N}$. On notera $\langle \cdot, \cdot \rangle$ le produit scalaire euclidien de $\mathbb{R}^d$, i.e. pour tout $(u, v) \in \mathbb{R}^d$, $\langle u, v \rangle = \sum_{i=1}^d u_i v_i$, si $u = (u_i)_{i \in \{1,...,d\}}$ et $v = (v_i)_{i \in \{1,...,d\}}$.

Une classe de fonctions $\mathcal{J}$ importante.

Definition I.3.5 *On dit que $\mathcal{J} : \mathbb{R}^d \to \mathbb{R}$ est une fonctionnelle quadratique si il existe $A \in M_d(\mathbb{R})$, $b \in \mathbb{R}^d$ et $c \in \mathbb{R}$ tels que, pour tout $x \in \mathbb{R}^d$, $\mathcal{J}(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle + c$.*

Remarque I.3.6 *On peut tout aussi bien écrire $\mathcal{J}(x) = \frac{1}{2}\langle Ax, x\rangle + \langle b, x\rangle + c$, avec $b \in \mathbb{R}^d$ (on choisit $-b$ à la place de b dans la définition).*

Ici, on fait donc la distinction entre **fonctionnelle** et **forme** quadratique. La différence entre les deux se situe dans le terme linéaire $\langle b, x\rangle$ et le terme constant c que l'on ajoute à la forme quadratique.

Definition I.3.7 *On dit que $\mathcal{J} : \mathbb{R}^d \to \mathbb{R}$ est une forme quadratique si il existe $A \in M_d(\mathbb{R})$ telle que, pour tout $x \in \mathbb{R}^d$, $\mathcal{J}(x) = \frac{1}{2}\langle Ax, x\rangle$.*

Les divers types de problème d'optimisation.

Donnons dans ce cas précis les divers types de problèmes d'optimisation que l'on va distinguer suivant la nature des contraintes.

1. Problèmes sans contraintes
   (a) Problèmes linéaires si $\mathcal{J}$ est **affine** [2],
   (b) Problèmes quadratiques si $\mathcal{J}$ est quadratique,
   (c) Problèmes non linéaires, sinon.
2. Problèmes avec contraintes linéaires (h_i et g_i **affines**)
   (a) Problèmes avec contraintes d'égalité uniquement
      i. Problème linéaire, si $\mathcal{J}$ est **affine**
      ii. Problèmes linéaires-quadratiques, si $\mathcal{J}$ est quadratique
      iii. Problèmes non-linéaires, sinon
   (b) Problèmes avec contraintes d'inégalité et mixtes.
      i. Programmation linéaire si $\mathcal{J}$ est **affine**,
      ii. Problèmes linéaires quadratiques si $\mathcal{J}$ est quadratique,
      iii. Problèmes non linéaires avec contraintes linéaires, sinon.
3. Problèmes de programmation non linéaire.

Cette classification peut varier légèrement suivant les personnes et on peut donner un analogue de classification dans des cas plus généraux que ceux de ce cours.

Dans la première partie de ce cours, on étudiera en détails le cas où $V = \mathbb{R}^d$ avec $d \in \mathbb{N}^*$ donné. C'est ce que l'on appelle de **l'optimisation en dimension finie** .

2. une fonction affine s'écrit $\mathcal{J}(x) = a + f(x)$ où $f : \mathbb{R}^d \to \mathbb{R}$ est une application linéaire, et $a \in \mathbb{R}$.

Chapitre II

Résultats d'existence et unicité en dimension finie ($V = \mathbb{R}^n$)

On se place dans le cadre d'un problème d'optimisation en dimension finie, i.e. $V = \mathbb{R}^n$ pour $n \in \mathbb{N}^*$, et on se demande si l'on sait dire si on a existence et unicité d'un point de minimum (on dit aussi *minimiseur*). On va voir que les réponses à cette question dépendent grandement du problème considéré (normal...). On commence par donner des exemples assez simples de situations où la réponse n'est pas oui !

Soit $V = \mathbb{R}^n$, $K \subset V$ et $\mathcal{J} : V \to \mathbb{R}$ pour prendre le cas le plus simple [1]. Dans ce chapître, on s'intéresse au problème de minimisation : *Trouver* $u^* \in K$ *tel que*

$$\mathcal{J}(u^*) = \inf_{u \in K} \mathcal{J}(u),$$

et on essaie de savoir si ce problème :

— admet au moins une solution ("existence"),

— admet au plus une solution ("unicité").

1. On verra que, suivant les cas, on peut aussi considérer que $\mathcal{J}$ n'est définie que sur un ouvert contenant K ou sur K lui même.

II.1 Quelques exemples et contre-exemples simples en dimension 1.

On se place sur $V = [a, b] \subset \mathbb{R}$.

1. Vous pouvez tracer une fonction $\mathcal{J}$ continue n'ayant pas de min ni de max sur $\mathbb{R}$. Penser par exemple à une droite affine sur $\mathbb{R}$. Elle n'a ni min ni max sur $\mathbb{R}$. Par contre elle possède un min et un max sur tout intervalle borné du type $[a, b]$ à l'extrémité des intervalles ! Voir illustration en Figure II.1.

   Ce qu'on voit avec cet exemple très simple, c'est que **les espaces choisis comptent énormément !** Il y a un gros rôle joué par la topologie suivant si l'espace est ouvert, fermé voire compact.

2. Essayez d'imaginer une fonction qui admet un minimum global, mais que le point de minimum n'est pas unique au sens où il y a plusieurs points où le minimum global est atteint. Solution [2]. Voir illustration en Figure II.2.
   On peut avoir existence du point de minimum global, mais pas unicité.

3. On peut avoir le cas d'un infimum non atteint [3]. Voir illustration en Figure II.3.

4. On peut avoir plein de points de minima locaux, mais aucun global. Voir illustration en Figure II.4.

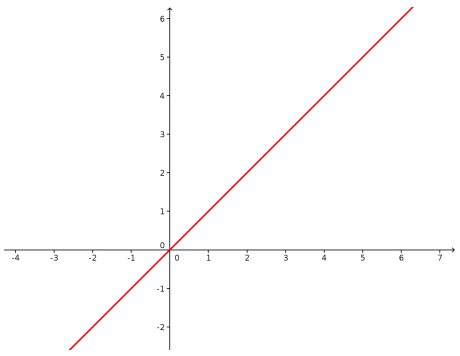

FIGURE II.1 – Exemple de fonction n'admettant pas de minimum sur $\mathbb{R}$, $f : x \mapsto x$.

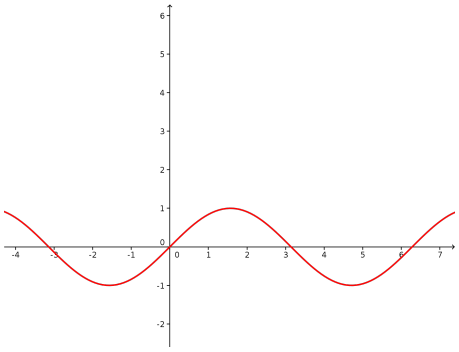

FIGURE II.2 – Exemple de fonction ayant plusieurs points de minima et maxima globaux sur $\mathbb{R}$, $f : x \mapsto \sin(x)$

2. Pensez par exemple à la fonction sinus sur $\mathbb{R}$. Il y a une seule valeur de minimum, mais les points de minimum sont localisés en les points du type $-\frac{\pi}{2} + 2k\pi$, avec $k \in \mathbb{Z}$.
3. $x \mapsto \frac{1}{x}$ sur $[1, +\infty[$.

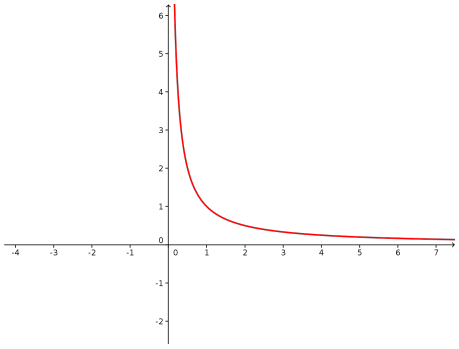

FIGURE II.3 – Exemple de fonction admettant un infimum non atteint sur $\mathbb{R}^{+*}$, $f : x \mapsto \frac{1}{x}$.

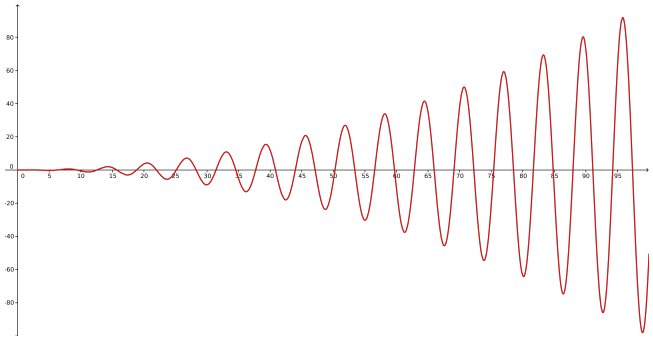

FIGURE II.4 – Exemple de fonction admettant plein de points de minima locaux, mais aucun global sur $\mathbb{R}$, $f : x \mapsto 10^{-2}x^2 \sin(x)$

Bref, les situations peuvent être très variées et compliquées.

II.2 Résultats d'existence dans le cas général.

II.2.1 Sur un espace métrique compact.

On a notamment le théorème suivant si on minimise sur un compact une fonction continue. Vous connaissez normalement ce résultat depuis un moment...

Théorème II.2.1 *Toute fonction continue sur un espace métrique compact est bornée et atteint ses bornes.*

Dans le cas considéré dans ce cours, on a donc le résultat suivant.

Corollaire II.2.2 *Soit $K \subset V$ compact et $\mathcal{J} : V \to \mathbb{R}$ continue. Le problème de minimisation de $\mathcal{J}$ sur K admet au moins une solution.*

II.2.2 Existence d'un minimum en dimension finie dans le cas général.

On cherche à avoir un résultat un peu plus général lorsque l'on n'est pas forcément sur un compact. Le résultat suivant est valable si on est en dimension finie.

Théorème II.2.3 *Soit K un ensemble fermé et non vide de $\mathbb{R}^n$ ($n \in \mathbb{N}^*$) et $\mathcal{J}$ une fonction continue sur K à valeur dans $\mathbb{R}$ vérifiant la propriété dite "infinie à l'infini"* [4] *:*

$$\forall (u_n)_{n\in\mathbb{N}} \text{ suite de } K, \lim_{n\to+\infty} \|u_n\| = +\infty \Rightarrow \lim_{n\to+\infty} \mathcal{J}(u_n) = +\infty.$$

Alors il existe au moins un point de minimum de $\mathcal{J}$ sur K. De plus, on peut extraire de toute suite minimisante de $\mathcal{J}$ sur K une sous-suite convergeant vers un point de minimum sur K.

Preuve : Soit $(u_n)_{n\in\mathbb{N}}$ une suite minimisante de $\mathcal{J}$ sur K (on a montré en section I.3.2 qu'une suite minimisante existe toujours). Par définition de cette suite minimisante, on sait que $(\mathcal{J}(u_n))_{n\in\mathbb{N}}$ converge vers $\inf_K \mathcal{J}$. Si $\inf_K \mathcal{J} \in \mathbb{R}$, alors $(\mathcal{J}(u_n))_{n\in\mathbb{N}}$ est une suite convergente et donc $(\mathcal{J}(u_n))_{n\in\mathbb{N}}$ est une suite bornée. Si $\inf_K \mathcal{J} = -\infty$, alors $\mathcal{J}(u_n) \to -\infty$ lorsque $n \to +\infty$.

Mais comme $\mathcal{J}$ est "infinie à l'infini", on en déduit que $(u_n)_{n\in\mathbb{N}}$ est bornée. Montrons cela par l'absurde. Supposons donc $(u_n)_{n\in\mathbb{N}}$ n'est pas bornée. Alors [5]

$$\forall M \geq 0, \exists n(M) \in \mathbb{N} \text{ tel que } \|u_{n(M)}\| > M. \tag{II.1}$$

On va construire une sous-suite extraite de $(u_n)_{n\in\mathbb{N}}$ qui converge vers $+\infty$. On note pour cela $A_0 := \{n \in \mathbb{N}, \|u_n\| > 0\}$. A_0 est un sous-ensemble de $\mathbb{N}$. De plus, en appliquant (II.1) avec $M = 0$, on sait que $n(0) \in A_0$. Donc A_0 est non vide. C'est donc un ensemble non vide et minoré (par 0) de $\mathbb{N}$, il admet donc un plus petit élément. Notons le $\varphi(0)$. Définissons alors l'espace A_p par récurrence avec pour $p \in \mathbb{N}^*$, $A_p := \{n \in \mathbb{N}, n \geq \varphi(p-1) + 1 \text{ et } \|u_n\| > p\}$. L'ensemble A_p est un sous-ensemble de $\mathbb{N}$. De plus, A_p est non vide. En effet, sinon, pour tout $n \in \mathbb{N}$, on a soit $n < \varphi(p-1) + 1$ ou $\|u_n\| \leq p$,

4. On dit alors aussi que f est *infinie à l'infini*.
5. On écrit la négation de $(u_n)_{n\in\mathbb{N}}$ bornée : $\exists M > 0$, tel que $\forall n \in \mathbb{N}$, $\|u_n\| \leq M$.

mais dans ce cas, cela signifie que pour tout $n \in \mathbb{N}$, $\|u_n\| \leq \max((\|u_k\|)_{k\in\{0,\varphi(p-1)\}}, p)$. Ce qui n'est pas possible puisqu'on a supposé $(u_n)_{n\in\mathbb{N}}$ non bornée. L'ensemble A_p est donc un ensemble non vide et minoré de $\mathbb{N}$, il admet donc un plus petit élément que l'on note $\varphi(p)$. On vient donc de construire par récurrence une application $\varphi : \mathbb{N} \to \mathbb{N}$. De plus, elle est strictement croissante, en effet, pour $p \in \mathbb{N}$, on a par définition, $\varphi(p+1) > \varphi(p)$.

On en déduit que $(u_{\varphi(p)})_{p\in\mathbb{N}}$ est une sous-suite extraite de $(u_n)_{n\in\mathbb{N}}$ et que de plus $\|u_{\varphi(p)}\| > p$. Posons alors pour tout $p \in \mathbb{N}$, $v_p = u_{\varphi(p)}$. On a alors pour tout $p \in \mathbb{N}$, $\|v_p\| \geq p$ et donc $\|v_p\| \to +\infty$ lorsque $p \to +\infty$. Comme $\mathcal{J}$ est infinie à l'infini, on sait que $\mathcal{J}(v_p) \to +\infty$, lorsque $p \to +\infty$. Et donc :

— si $\inf_K \mathcal{J} \in \mathbb{R}$, cela est en contradiction avec $(\mathcal{J}(u_n))_{n\in\mathbb{N}}$ et donc $(\mathcal{J}(v_p))_{p\in\mathbb{N}}$ est bornée,

— si $\inf_K \mathcal{J} = -\infty$, cela est en contradiction avec $\mathcal{J}(v_p) \to \inf_K \mathcal{J}$ lorsque $p \to +\infty$ comme $(v_p)_{p\in\mathbb{N}}$ est une suite extraite de $(u_n)_{n\in\mathbb{N}}$.

En conclusion on a bien montré que $(u_n)_{n\in\mathbb{N}}$ est une suite bornée. Par le théorème de Bolzano-Weierstrass, on sait donc que l'on peut en extraire de $(u_n)_{n\in\mathbb{N}}$ une sous-suite qui converge dans $\mathbb{R}^n$ vers un point $u^* \in \mathbb{R}^n$. Mais comme K est fermé, on en déduit que $u^* \in K$.

Enfin, vu la définition de $(u_n)_{n\in\mathbb{N}}$ (c'est une suite minimisante), on sait donc, en utilisant que $\mathcal{J}$ est continue, que

$$\mathcal{J}(u^*) = \inf_{u\in K} \mathcal{J}(u). \tag{II.2}$$

Le point u^* est donc un point de minimum de $\mathcal{J}$ sur K.

□

On peut donner quelques exemples de fonctions qui sont "infinies à l'infini".

Exemple 1 : $\mathcal{J} : \mathbb{R} \to \mathbb{R}, x \mapsto \|x\|^2$.

Exemple 2 : $f : K \subset \mathbb{R}^2 \to \mathbb{R}, (x, y) \mapsto x^4 + y^4 - x^2$, et $K = \{(x, y) \in \mathbb{R}^2, x + y \leq 4\}$.

Et en dimension infinie ? Attention, **ce théorème n'est plus valable en dimension infinie**, puisque les fermés bornés ne sont plus forcément compacts. On fait alors appel à l'analyse Hilbertienne [6] (notion convergence faible et semi-continuité inférieure de $\mathcal{J}$ pour pouvoir conclure, mais ceci est hors programme pour ce cours).

Pour ce qui est du problème de l'unicité et du caractère global ou local du minimum, on peut avoir plein de situations différentes. **La convexité** est une notion qui aide pour les problèmes d'unicité des minimiseurs, mais aussi sur le caractère global ou local du minimiseur. C'est l'objet de la section suivante.

II.3 Existence et unicité dans le cas d'une fonction coût convexe

II.3.1 Définitions, quelques rappels

Definition II.3.1 *On dit qu'un ensemble C est convexe, si $\forall (u, v) \in C \times C$, $\forall t \in [0, 1]$, $tu + (1-t)v \in C$.*

Definition II.3.2 *Soit E un $\mathbb{R}$-ev, $C \subset E$ un ensemble convexe et $\mathcal{J} : C \to \mathbb{R}$.*

6. Pour ceux qui ont déjà eu un cours d'analyse Hilbertienne.

$\mathcal{J}$ est une ***fonction convexe*** *si elle vérifie pour tout $(x, y) \in C^2$, $t \in [0, 1]$,*

$$\mathcal{J}(tx + (1-t)y) \leq t\mathcal{J}(x) + (1-t)\mathcal{J}(y).$$

$\mathcal{J}$ est une ***fonction strictement convexe*** *si elle vérifie pour tout $(x, y) \in C^2$ avec $x \neq y$, $t \in]0, 1[$,*

$$\mathcal{J}(tx + (1-t)y) < t\mathcal{J}(x) + (1-t)\mathcal{J}(y).$$

$\mathcal{J}$ est une ***fonction α-convexe*** *si pour tout $(x, y) \in C^2$, pour tout $t \in [0, 1]$, on a*

$$\mathcal{J}(tx + (1-t)y) \leq t\mathcal{J}(x) + (1-t)\mathcal{J}(y) - \frac{\alpha}{2}t(1-t)\|x - y\|^2.$$

On parle de problème d'*optimisation convexe*, si $\mathcal{J}$ est convexe et K, l'ensemble sur lequel on minimise $\mathcal{J}$, est convexe.
On peut faire quelques remarques importantes :

— Si $\mathcal{J}$ est une fonction strictement convexe, alors $\mathcal{J}$ est convexe,

— Si $\mathcal{J}$ est une fonction α-convexe avec $\alpha > 0$, alors $\mathcal{J}$ est une fonction strictement convexe.

Par contre les implications réciproques sont fausses.

De plus,

— Une fonction 0-convexe est une fonction convexe.

— Une somme de fonctions convexes est convexe,

— Une composition de deux fonctions convexes f et g ($f \circ g$) **n'est pas forcément convexe** ! Par contre si f est convexe et g est linéaire, la composition $f \circ g$ est convexe.

— Une fonction convexe **doit** être définie sur un ensemble convexe !

On rappelle également le résultat suivant valable en dimension finie qui dit que toute fonction convexe sur un ouvert inclus dans $\mathbb{R}^d$, $d \in \mathbb{N}^*$, est continue.

Théorème II.3.3 *Soit U un ouvert convexe de $\mathbb{R}^d$, $d \in \mathbb{N}^*$, et $\mathcal{J} : U \to \mathbb{R}$ une fonction convexe sur l'ouvert U. Alors $\mathcal{J}$ est continue sur U.*

II.3.2 Existence d'un minimiseur sur $\mathbb{R}^d$

Proposition II.3.4 *Soit $\mathcal{J} : \mathbb{R}^d \to \mathbb{R}$ une fonction α-convexe (avec $\alpha > 0$). Alors $\mathcal{J}$ est "infinie à l'infini".*

Preuve : En appliquant l'inégalité d'α-convexité, avec $u = 0 \in \mathbb{R}^d$, on obtient, $\forall y \in \mathbb{R}^d$, $\forall \delta \in [0, 1]$,

$$\mathcal{J}(\delta y) \leq (1-\delta)\mathcal{J}(0) + \delta\mathcal{J}(y) - \frac{\alpha}{2}\delta(1-\delta)\|y\|^2.$$

Ce qui se réécrit,

$$\forall y \in \mathbb{R}^d, \forall \delta \in]0, 1], \mathcal{J}(y) \geq \mathcal{J}(0) + \frac{\mathcal{J}(\delta y) - \mathcal{J}(0)}{\delta} + \frac{\alpha}{2}(1-\delta)\|y\|^2.$$

On peut alors appliquer cette inégalité aux $y \in \mathbb{R}^d$ tels que $\|y\| \geq 2$, avec $\delta = \frac{1}{\|y\|}$, on obtient

$$\mathcal{J}(y) \geq \mathcal{J}(0) + \left[\mathcal{J}(\frac{y}{\|y\|}) - \mathcal{J}(0)\right] \|y\| + \frac{\alpha}{2}\left(1 - \frac{1}{\|y\|}\right)\|y\|^2 \geq \mathcal{J}(0) + K\|y\| + \frac{\alpha}{4}\|y\|^2, \qquad \text{(II.3)}$$

où K est un minorant de $u \mapsto \mathcal{J}(u) - \mathcal{J}(0)$ sur la sphère unité (K existe car la sphère unité est compacte et $\mathcal{J}$ est continue sur $\mathbb{R}^d$ cf. théorème II.3.3) et où on a utilisé que $(1 - \frac{1}{\|y\|}) \geq \frac{1}{2}$.

On déduit de (II.3) que $\mathcal{J}$ est infinie à l'infini.

□

On en déduit en particulier que l'on peut appliquer le théorème II.2.3 si $\mathcal{J}$ est α-convexe avec $\alpha > 0$ et K fermé non vide, pour en déduire l'existence d'au moins un point de minimum.

II.3.3 Unicité et caractère local ou global du minimum.

Théorème II.3.5 *On suppose que $\mathcal{J} : K \subset \mathbb{R}^d \to \mathbb{R}$ et K sont convexes. On a alors :*

(i) Tout minimum local de $\mathcal{J}$ sur K est un minimum global de $\mathcal{J}$ sur K,

(ii) Si $\mathcal{J}$ est strictement convexe, alors le problème a au plus une solution, i.e. il y a au plus un point de minimum (0 *ou* 1 *point de minimum).*

Preuve : Montrons *(i)*. Soit u^* un point de minimum local de $\mathcal{J}$ sur K. Si $K = \{u^*\}$, alors *(i)* est vrai (il n'y a qu'une seule valeur de $\mathcal{J}$ sur K, $\mathcal{J}(u^*)$). Supposons alors que $K \neq \{u^*\}$. Soit $v \in K \setminus \{u^*\}$ un élément quelconque de K différent de u^*. On a, comme K est convexe, pour tout $t \in [0,1]$,

$$tu^* + (1-t)v \in K.$$

Comme u^* est un point de minimum local de $\mathcal{J}$, on sait qu'il existe $\delta > 0$ tel que si $w \in K$ et $\|w - u^*\| \leq \delta$, alors $\mathcal{J}(u^*) \leq \mathcal{J}(w)$.

Choisissons alors $t^* \in [0,1]$ tel que $1 > t^* > \max(0, 1 - \frac{\delta}{\|u^*-v\|})$ (possible puisque $\|u^* - v\| \neq 0$ et $\max(0, 1 - \frac{\delta}{\|u^*-v\|}) < 1$) et posons $w^* = t^*u^* + (1-t^*)v$. On a $t^* \in [0,1]$ et donc $w^* \in K$. De plus $\|w^* - u^*\| = (1-t^*)\|u^* - v\| \leq \frac{\delta}{\|u^*-v\|}\|u^* - v\| \leq \delta$, puisque par définition de t^*, on a $1 - t^* < \min(1, \frac{\delta}{\|u^*-v\|})$. Donc $\mathcal{J}(w^*) \geq \mathcal{J}(u^*)$. Or puisque $\mathcal{J}$ est convexe,

$$\mathcal{J}(w^*) \leq t^*\mathcal{J}(u^*) + (1-t^*)\mathcal{J}(v).$$

En combinant les deux dernières inégalités, on trouve

$$\mathcal{J}(u^*) \leq \mathcal{J}(w^*) \leq t^*\mathcal{J}(u^*) + (1-t^*)\mathcal{J}(v).$$

Et donc les deux inégalités extrèmes donnent

$$(1-t^*)\mathcal{J}(u^*) \leq (1-t^*)\mathcal{J}(v),$$

Finalement, puisque

$$t^* < 1,$$

on a donc montré que $\forall v \in K$,

$$\mathcal{J}(v) \geq \mathcal{J}(u^*).$$

Cela signifie que u^* est bien un point de minimum global.

Montrons *(ii)* par l'absurde. Supposons qu'il existe deux points de minimum global dans K, notés $u_1 \in K$ et $u_2 \in K$ distincts. Alors dans ce cas, puisque pour tout $t \in]0,1[$, $tu_1 + (1-t)u_2 \in K$, on en déduit par stricte convexité de $\mathcal{J}$ et définition de u_1, que

$$\mathcal{J}(u_1) \leq \mathcal{J}(tu_1 + (1-t)u_2) < t\mathcal{J}(u_1) + (1-t)\mathcal{J}(u_2).$$

En combinant les deux inégalités extrèmes, on trouve

$$(1-t)\mathcal{J}(u_1) < (1-t)\mathcal{J}(u_2).$$

Donc $\mathcal{J}(u_1) < \mathcal{J}(u_2)$ (puisque $t < 1$) : contradiction avec la définition de u_2. □

En combinant tous ces résultats, on arrive à avoir des informations sur l'existence et l'unicité d'une solution au problème de minimisation dans le cas α-convexe avec $\alpha > 0$.

II.4 Apport de la régularité de $\mathcal{J}$.

Pour pouvoir répondre à une des questions que l'on s'est posée au départ : "peut-on avoir une caractérisation du point de minimum si il existe ?", si $\mathcal{J}$ est suffisamment régulière ($\mathcal{C}^1$, $\mathcal{C}^2$), on aura à disposition des conditions à vérifier sur les différentielles première (gradient) et seconde (Hessienne) de la fonction que l'on cherche à minimiser. On verra ces conditions dans le chapître suivant. Dans cette section, on fait les rappels nécessaires pour pouvoir établir les résultats de caractérisation et avoir des façons plus simple de montrer qu'une fonctionnelle est α-convexe par exemple dans le cas régulier.

II.4.1 Rappel de la notion de Gradient et Hessienne.

Gradient

Soit U un ouvert de $\mathbb{R}^d$, $a \in U$ et $F : U \to \mathbb{R}$ différentiable en a (si vous n'êtes pas à l'aise avec "différentiable", remplacez par $\mathcal{C}^1$). On appelle **gradient** de F en un point a de $\mathbb{R}^d$, le vecteur de $\mathbb{R}^d$ dont les coordonnées dans la base canonique de $\mathbb{R}^d$ sont données par les dérivée partielles, autrement dit :

$$\nabla F(a) = \begin{pmatrix} \dfrac{\partial F}{\partial x_1}(a) \\ \dfrac{\partial F}{\partial x_2}(a) \\ \vdots \\ \dfrac{\partial F}{\partial x_d}(a) \end{pmatrix}$$

Remarque II.4.1 *Pour les personnes à l'aise avec la notion de différentielle. On a donc en particulier pour tout $h \in \mathbb{R}^d$:*

$$DF(a).h = \langle \nabla F(a), h \rangle = {}^t \nabla F(a) h.$$

Hessienne

Si F est $\mathcal{C}^2$ en $a \in U$, on peut définir la différentielle seconde de F en a, $D^2F(a)$. Dans ce cas, la matrice $A = \left(\dfrac{\partial^2 F}{\partial x_i \partial x_j}(a) \right)_{(i,j)\in\{1,..,d\}^2} \in M_d(\mathbb{R})$ est appelée **matrice Hessienne** de F en a, on pourra la noter $Hess_a(F)$. On remarque que comme $\dfrac{\partial^2 F}{\partial x_i \partial x_j}(a) = \dfrac{\partial^2 F}{\partial x_j \partial x_i}(a)$ (théorème de Schwartz), la matrice est symétrique.

De plus, pour tout $(h, l) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$D^2F(a).(h, l) = \sum_{i,j} (Hess_a(F))_{i,j} h_i l_j = \langle Hess_a(F)h, l \rangle.$$

II.4.2 Conséquences de la régularité dans le cas convexe : caractérisation des fonctions α-convexes.

En utilisant les notions de calcul différentiel dans le cas où la fonction considérée est régulière, on a accès à des résultats pratiques pour voir si une fonction est convexe, strictement convexe, α-convexe. Ce qui peut s'avérer pratique dans les exercices notamment !

Proposition II.4.2 *Soit $\mathcal{J} : \mathbb{R}^d \to \mathbb{R}$ et $\mathcal{C}^1$ et $\alpha \geq 0$ donné. Alors les trois assertions suivantes sont équivalentes :*

(i) $\mathcal{J}$ une fonctionnelle α-convexe sur $\mathbb{R}^d$

(ii) Pour tout $(u, v) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\mathcal{J}(v) \geq \mathcal{J}(u) + (\nabla \mathcal{J}(u), v - u) + \frac{\alpha}{2}\|v - u\|^2. \tag{II.4}$$

(iii) Pour tout $(u, v) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$(\nabla \mathcal{J}(v) - \nabla \mathcal{J}(u), v - u) \geq \alpha \|v - u\|^2.$$

Preuve : (i)⇒(ii). Soit $(u, v) \in \mathbb{R}^d$, on a pour $t \in [0, 1]$,

$$\mathcal{J}((1-t)u + tv) \leq (1-t)\mathcal{J}(u) + t\mathcal{J}(v) - \frac{\alpha}{2}t(1-t)\|u - v\|^2. \tag{II.5}$$

Mais de plus en utilisant la formule de Taylor Young, lorsque $t \to 0$:

$$\mathcal{J}((1-t)u + tv) = \mathcal{J}(u + t(v-u)) = \mathcal{J}(u) + t\underbrace{D\mathcal{J}(u) \cdot (v-u)}_{=\langle \nabla \mathcal{J}(u), v-u\rangle} + \|u - v\| o(t). \tag{II.6}$$

En rassemblant (II.5) et (II.6), on trouve

$$tD\mathcal{J}(u) \cdot (v-u) + \|u - v\| o(t) \leq t(\mathcal{J}(v) - \mathcal{J}(u)) - \frac{\alpha}{2}t(1-t)\|u - v\|^2.$$

En choisissant $t \in]0, 1]$, on peut donc écrire, lorsque $t \to 0^+$,

$$\mathcal{J}(v) \geq \mathcal{J}(u) + D\mathcal{J}(u) \cdot (v-u) + \frac{\alpha}{2}(1-t)\|u - v\|^2 + \|u - v\| o(1).$$

En faisant tendre $t \to 0^+$, on obtient donc

$$\mathcal{J}(v) \geq \mathcal{J}(u) + D\mathcal{J}(u) \cdot (v-u) + \frac{\alpha}{2}\|u - v\|^2.$$

On rappelle que $D\mathcal{J}(u) \cdot (v-u)$ peut aussi s'écrire $\langle \nabla \mathcal{J}(u), v-u\rangle$. Si vous n'êtes pas à l'aise avec la notation $D\mathcal{J}(u)$, vous pouvez ré-écrire toute la preuve avec $\nabla \mathcal{J}(u)$

(ii)⇒(iii) On échange le rôle de u et v dans (ii), et on trouve une inégalité (ii)'. On additionne ces deux inégalités (ii) et (ii)', et on obtient le résultat.

(iii)⇒(i) Soit $\varphi : t \mapsto \mathcal{J}(u + t(v-u))$. Vu les hypothèses sur $\mathcal{J}$, φ est continue et même $\mathcal{C}^1$ sur $\mathbb{R}$ et $\forall t \in \mathbb{R}, \forall (u,v) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\varphi'(t) = D\mathcal{J}(u + t(v-u)) \cdot (v-u).$$

On a ici utilisé la différentiation de fonctions composées.

De plus, par (iii) (appliqué à $u + t(v-u)$ et $u + s(v-u)$), on a $\forall (t,s) \in \mathbb{R} \times \mathbb{R}$,

$$\varphi'(t) - \varphi'(s) \geq \alpha(t-s)\|v-u\|^2, \text{ si } t > s.$$

Soit alors $\theta \in]0,1[$, en intégrant sur $(t,s) \in [\theta, 1] \times [0, \theta[$, on obtient

$$\int_0^\theta \int_\theta^1 \varphi'(t)dtds - \int_\theta^1 \int_0^\theta \varphi'(s)dsdt \geq \alpha(\int_0^\theta \int_\theta^1 tdtds - \int_0^\theta \int_\theta^1 sdtds)\|v-u\|^2.$$

Ce qui donne

$$\theta \int_\theta^1 \varphi'(t)dt - (1-\theta)\int_0^\theta \varphi'(s)ds \geq \frac{\alpha}{2}\theta(1-\theta)\|v-u\|^2.$$

Autrement dit

$$\theta\varphi(1) + (1-\theta)\varphi(0) - \varphi(\theta) \geq \frac{\alpha}{2}\theta(1-\theta)\|v-u\|^2. \tag{II.7}$$

En remplaçant par l'expression de φ en fonction de $\mathcal{J}$, on trouve exactement l'expression de l'α-convexité. □

Remarque II.4.3 *Ce théorème peut être généralisé au cas où $\mathcal{J}$ est une fonction définie sur un ouvert $\Omega \subset \mathbb{R}^d$ et (toujours) $\mathcal{C}^1$. Dans ce cas là dans les énoncés des assertions on remplace le "Pour tout $(u,v) \in \mathbb{R}^d \times \mathbb{R}^d$" par "Pour tout $(u,v) \in U \times U$", avec $U \subset \Omega$ un ensemble convexe. Et on dit que $\mathcal{J}$ est α-convexe sur U.*

Proposition II.4.4 *Soit $\mathcal{J} : \mathbb{R}^d \to \mathbb{R}$, $\mathcal{C}^2$ et $\alpha \geq 0$ donné. Alors $\mathcal{J}$ est α-convexe sur $\mathbb{R}^d$ si et seulement si*

$$D^2\mathcal{J}(u).(w,w) \geq \alpha\|w\|^2 \text{ pour tout } (u,w) \in \mathbb{R}^d \times \mathbb{R}^d,$$

ce qui s'écrit également

$$(Hess\mathcal{J}(u)w, w) \geq \alpha\|w\|^2 \text{ pour tout } (u,w) \in \mathbb{R}^d \times \mathbb{R}^d,$$

.

Preuve : Supposons que $\mathcal{J}$ est α-convexe. Soient $(u,v) \in \mathbb{R}^d \times \mathbb{R}^d$. Alors $\varphi : t \mapsto \mathcal{J}(u + t(v-u))$ est $\mathcal{C}^2$ et $\varphi'(t) = D\mathcal{J}(u + t(v-u)) \cdot (v-u) = \langle \nabla\mathcal{J}(u + t(v-u)), v-u\rangle$, et $\varphi''(t) = D^2\mathcal{J}(u + t(v-u)) \cdot (v-u, v-u) = \langle Hess_{u+t(v-u)}(\mathcal{J})(v-u), v-u\rangle$. On dispose des équivalences de la proposition précédente. On peut aussi reprendre la preuve de (iii)⇒(i) et utiliser que $\forall (t,s) \in \mathbb{R} \times \mathbb{R}$, si $t > s$,

$$\varphi'(t) - \varphi'(s) \geq \alpha(t-s)\|v-u\|^2.$$

Choisissons $s \in \mathbb{R}$ et $h \in \mathbb{R}_+^*$ et $t = s + h$, on a alors

$$\varphi'(s+h) - \varphi'(s) \geq \alpha h \|v-u\|^2.$$

Et donc comme φ est $\mathcal{C}^2$ sur $\mathbb{R}$, on en déduit, en divisant par $h > 0$ et en faisant tendre h vers 0^+, que

$$\varphi''(t) \geq \alpha \|v-u\|^2.$$

On a donc pour tout $(u, v) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$D^2\mathcal{J}(u + t(v-u)) \cdot (v-u, v-u) \geq \alpha \|v-u\|^2.$$

Soit alors $(w, u) \in \mathbb{R}^d \times \mathbb{R}^d$. On choisit $v = w + u$ dans l'inégalité précédente, on obtient

$$D^2\mathcal{J}(u + tw) \cdot (w, w) \geq \alpha \|w\|^2.$$

En prenant $t = 0$, on obtient donc le résultat voulu.

Supposons maintenant que

$$D^2\mathcal{J}(u).(w, w) \geq \alpha \|w\|^2 \text{ pour tout } (u, w) \in \mathbb{R}^d.$$

Soit $(u, v) \in \mathbb{R}^d \times \mathbb{R}^d$ et $t \in \mathbb{R}$. En appliquant cette inégalité avec $(u + t(v-u))$ et $(v-u)$, on obtient

$$D^2\mathcal{J}(u + t(v-u)) \cdot (v-u, v-u) \geq \alpha \|v-u\|^2.$$

Et comme précédemment, on déduit donc que

$$\varphi''(t) \geq \alpha \|(u-v)\|^2.$$

En intégrant entre 0 et 1, on trouve donc que

$$\varphi'(1) - \varphi'(0) \geq \alpha \|v-u\|^2.$$

Et donc

$$(D\mathcal{J}(v) - D\mathcal{J}(u)) \cdot (v-u) \geq \alpha \|v-u\|^2.$$

En conclusion $\mathcal{J}$ est α-convexe. □

Remarque II.4.5 *On retrouve le fait bien connu que si la Hessienne est positive alors $\mathcal{J}$ est convexe (c'est le cas $\alpha = 0$ du théorème).*

Remarque II.4.6 *Ce théorème peut être généralisé au cas où $\mathcal{J}$ est une fonction définie sur un ouvert $\Omega \subset \mathbb{R}^d$ et (toujours) $\mathcal{C}^2$. Dans ce cas là, on remplace la conclusion*

$$D^2\mathcal{J}(u).(w, w) \geq \alpha \|w\|^2 \text{ pour tout } (u, w) \in \mathbb{R}^d \times \mathbb{R}^d,$$

par

$$D^2\mathcal{J}(u).(v-u, v-u) \geq \alpha \|v-u\|^2 \text{ pour tout } (u, v) \in U \times U,$$

avec $U \subset \Omega$ un ensemble convexe. Et on dit que $\mathcal{J}$ est α-convexe sur U.

II.4.3 Conséquences sur un problème d'optimisation.

Généralisation du théorème II.3.4 à un ensemble convexe

On peut aussi montrer la proposition suivante que l'on admettra.

Proposition II.4.7 ***(admis)*** *Soit Ω un ouvert de $\mathbb{R}^d$ et $U \subset \Omega$ un ensemble convexe et non borné. Soit alors $\mathcal{J} : \Omega \to \mathbb{R}$ une fonction α-convexe (avec $\alpha > 0$) sur U et $\mathcal{C}^1$ sur Ω. Alors $\mathcal{J}$ est infinie à l'infini sur U.*

Cas particulier

On rappelle ici un résultat d'algèbre linéaire qui nous dit que si $A \in M_d(\mathbb{R})$ est une matrice symétrique, alors

$$\langle Aw, w\rangle \geq \lambda_{min}(A)\|w\|^2,$$

si $\lambda_{min}(A)$ est la plus petite des valeurs propres de A. On en déduit donc que si on étudie la Hessienne au point u et qu'on est capable de trouver ses valeurs propres, cela peut éventuellement nous donner un moyen de prouver l'α-convexité en utilisant la proposition II.4.4.

Chapitre III

Problèmes sans contraintes en dimension finie.

On parlera de problèmes d'optimisation sans contraintes lorsque $K = V$. On se place par défaut dans le cas où $V = \mathbb{R}^d$, avec $d \in \mathbb{N}^*$ donné. Si ce n'est pas le cas, on le précisera explicitement. Dans ce cas particulier, on a toujours à disposition les résultats d'existence et/ou unicité du chapître précédent. On précisera cependant parfois ces résultats au besoin.

III.1 Caractérisations du(des) point(s) de minimum local dans un ouvert

On va voir que les caractérisations que l'on a valent généralement pour un minimum local.

III.1.1 Conditions nécessaires et/ou suffisantes d'optimalité.

On se place dans le cas où $V = \mathbb{R}^d$, avec $d \geq 1$. On a accès là aussi à des conditions nécessaires ou suffisantes d'optimalité.
On suppose que $U \subset \mathbb{R}^d$ est un ouvert de $\mathbb{R}^d$. Commençons par la condition nécessaire.

Condition nécessaire.

Proposition III.1.1 ***Si*** *$F : U \subset \mathbb{R}^d \to \mathbb{R}$ admet un minimum local en un point a de U et si F est différentiable en a,* ***alors*** *$\nabla F(a) = 0$ (autrement dit $\dfrac{\partial F}{\partial x_i}(a) \equiv 0$, pour tout $i \in \{1, .., d\}$).*

Remarque III.1.2 *Cette équation vérifiée au point de minimum, s'appelle aussi équation d'Euler.*

Preuve : Voir annexe B. □

Conditions nécessaires et/ou suffisantes

On continue avec un résultat qui porte sur la différentielle seconde et qui contient une condition suffisante d'optimalité si on a un point critique.

Théorème III.1.3 *Soit $F : U \subset \mathbb{R}^d \to \mathbb{R}$ une fonction de classe $\mathcal{C}^2$ et supposons qu'il existe $a \in U$, tel que $\nabla F(a) = 0$. Alors :*

(i) Si F admet un minimum local en a, alors $Hess_a(F)$ est une matrice symétrique positive (condition nécessaire).

(ii) Si

$$Hess_a(F)$$

est une matrice symétrique définie positive, alors F admet un minimum local en a (condition suffisante).

Preuve : Voir annexe B. □

⚠Si la matrice Hessienne n'est que positive (et pas **définie** positive) au point critique, on n'est pas assuré d'avoir un minimum (exemple classique de x^3 en dimension 1).

III.1.2 Cas de l'optimisation convexe : condition nécessaire et suffisante.

Théorème III.1.4 *Extrema de fonctions convexes. Si $\mathcal{J} : \mathbb{R}^d \to \mathbb{R}$, est convexe et $\mathcal{C}^1$, alors tout point $u^* \in \mathbb{R}^d$ tel que*

$$\forall v \in \mathbb{R}^d, \underbrace{D\mathcal{J}(u^*) \cdot v}_{\langle \nabla \mathcal{J}(u^*), v \rangle} = 0,$$

est un point de minimum global de $\mathcal{J}$ sur $\mathbb{R}^d$ et réciproquement.

Preuve : Supposons que $u^* \in \mathbb{R}^d$ est un point de minimum global de $\mathcal{J}$ sur $\mathbb{R}^d$. Soit alors $v \in \mathbb{R}^d$. On a $(1-t)u^* + tv$ est dans $\mathbb{R}^d$ pour tout $t \in [0,1]$. Par formule de Taylor, on déduit que lorsque $t \to 0^+$, $\mathcal{J}((1-t)u^* + tv) - \mathcal{J}(u^*) = tD\mathcal{J}(u^*)(v-u^*) + o(t)$.
Si on avait $D\mathcal{J}(u^*).(v-u^*) < 0$, on en déduirait que pour t suffisamment petit $\mathcal{J}((1-t)u^* + tv) - \mathcal{J}(u^*) < 0$, ce qui est une contradiction avec la définition de u^*. On a donc $\forall v \in \mathbb{R}^d$, $D\mathcal{J}(u^*).(v-u^*) \geq 0$. Cela donne $\forall w \in \mathbb{R}^d$, $D\mathcal{J}(u^*).w \geq 0$ (il suffit de poser $v = w + u^*$ dans l'inégalité). En appliquant maintenant ce résultat à $-w \in \mathbb{R}^d$, on obtient

$$D\mathcal{J}(u^*).w \leq 0.$$

On a donc finalement pour tout $v \in \mathbb{R}^d$, $D\mathcal{J}(u^*) \cdot v = 0$.

Supposons maintenant que pour tout $v \in \mathbb{R}^d$, $D\mathcal{J}(u^*).v = 0$. On a pour tout $(u,v) \in \mathbb{R}^d$, $t \in [0,1]$, $(1-t)u + tv \in \mathbb{R}^d$ et, comme $\mathcal{J}$ est convexe, pour tout $t \in]0,1]$,

$$\frac{\mathcal{J}((1-t)u + tv) - \mathcal{J}(u)}{t} \leq \mathcal{J}(v) - \mathcal{J}(u).$$

Donc en faisant tendre t vers 0^+, on obtient

$$D\mathcal{J}(u).(v-u) \leq \mathcal{J}(v) - \mathcal{J}(u).$$

Donc en particulier pour tout $v \in \mathbb{R}^d$, en utilisant l'inégalité précédente avec $u = u^*$,

$$\mathcal{J}(v) - \mathcal{J}(u^*) \geq 0,$$

et donc u^* est bien un point de minimum global. □

III.2 Approximation numérique.

Ici, on cherche à avoir une réponse à une autre des questions que l'on s'est posées au début. On avait dit qu'une fois que l'on sait que le point de minimum existe et est éventuellement unique, le point intéressant est de savoir le donner ou de l'estimer. Cela rejoint donc la question : "peut on approcher le minimum et le point de minimum d'une fonctionnelle donnée ? ". Pour faire cela, il est assez rare d'avoir accès analytiquement à ce point (sauf dans des cas très simples). On va donc avoir recours à des méthodes numériques qui vont nous permettre d'approcher ces valeurs.

Dans le cadre de ce cours d'optimisation, on ne considèrera que des problèmes d'optimisation en dimension finie, et on choisit de travailler avec $V = \mathbb{R}^d$. De plus, on considèrera par défaut, une fonctionnelle $\mathcal{J}$ qui sera définie sur $\mathbb{R}^d$ à valeurs dans $\mathbb{R}$. Si ce n'est pas le cas, on précisera le domaine de définition de $\mathcal{J}$. Bien souvent un ouvert $\mathcal{U}$ de $\mathbb{R}^d$.

III.2.1 Problématiques et principes généraux

Lorsque l'on ne sait pas trouver facilement et/ou explicitement la solution du problème de minimisation, on va faire appel à une méthode numérique qui va permettre d'approcher la solution du problème.

Supposons que l'on sache qu'un minimum global existe, notons u^* un point de minimum. Les algorithmes que nous allons étudier dans ce cours sont de nature itérative. En partant d'une donnée initiale u^0, on construit itérativement une suite $(u_n)_{n\in\mathbb{N}}$ et l'on espère que plus n est grand plus u_n se rapproche de u^*. Autrement dit, on espère la convergence asymptotique de $(u_n)_{n\in\mathbb{N}}$ vers u^* lorsque $n \to +\infty$.

En théorie, l'algorithme permet de construire une suite (dénombrable) dont on va essayer de montrer *la convergence* vers le minimum u^*. En pratique, on devra arrêter les itérations de l'algorithme à une itération donnée N suffisamment grande pour espérer avoir une erreur entre la valeur calculée et la valeur approchée suffisamment petite.

Il se pose alors plusieurs questions :

(α) Quelle stratégie adopter pour trouver un algorithme de calcul ?

(β) Peut-on s'assurer de la convergence de l'algorithme vers u^* ? Vitesse de convergence éventuelle ?

(γ) Comment arrêter les itérations de l'algorithme ?

Pour obtenir des résultats de convergence, on va être amenés à faire des hypothèses assez fortes sur la fonction coût $\mathcal{J}$, qui ne seront parfois pas nécessairement vérifiées en pratique (ou l'on aura des difficultés à les vérifier en pratique). C'est toujours un problème qui se pose entre théorie et pratique.

Cela peut également poser des problèmes de non convergence de l'algorithme ou de convergence vers un minimum local au lieu du minimum recherché...

Dans ce qui suit, on supposera que $\mathcal{J}$ est au moins $\mathcal{C}^2$ sur le domaine sur lequel elle est définie ($\mathcal{U}$ ouvert ou $\mathbb{R}^d$), si ce n'est pas le cas, on précisera exactement les hypothèses. On se place en général également dans le cadre d'hypothèses permettant d'assurer existence et unicité du minimum (*voir sections et chapîtres précédents pour ces hypothèses*). On notera u^* ce point de minimum.

III.2.2 Principe des méthodes de descente et gradient

Dans cette section, nous allons donner une famille d'algorithmes qui permettent d'approcher un point de minimum. Ces algorithmes sont itératifs et se basent sur le principe de suivre, au fur et à mesure des itérations, une *direction de descente*. Cette direction est en fait un vecteur qui permet de garantir que si on suit cette direction "pas trop longtemps" on diminue la valeur de $\mathcal{J}$.

Principe général ; direction de descente.

Commençons par formaliser ce principe de la *direction de descente*. On se donne $\mathcal{U} \subset \mathbb{R}^d$ un ouvert de $\mathbb{R}^d$.

Definition III.2.1 *Soit $\mathcal{J} : \mathcal{U} \to \mathbb{R}$ et $u_0 \in \mathcal{U} \subset \mathbb{R}^d$. On appelle direction de descente pour $\mathcal{J}$ en u_0 tout vecteur w de $\mathcal{U}$, telle qu'il existe $\eta > 0$ tel que $\forall t \in [0, \eta]$, $u_0 + tw \in \mathcal{U}$ et $\mathcal{J}(u_0 + tw) \leq \mathcal{J}(u_0)$.*

On retrouve dans cette définition l'intuition initiale. Le vecteur $u_0 + tw$ peut s'interpréter comme le vecteur obtenu en "partant" de u_0 et en suivant la direction w sur une longueur t. Cette longueur n'est pas trop grande : on ne le fait que sur une longueur au plus η. Vous pouvez faire un dessin pour vous aider. Et on voit que si on évalue $\mathcal{J}$ en ce vecteur $u_0 + tw$, la valeur est plus petite que celle obtenue en u_0 : on est donc bien "descendu".

Grâce à cette définition, on va pouvoir préciser maintenant la famille d'algorithmes que l'on va regarder. On s'intéresse aux méthodes dites de *descente* où l'on cherche à approcher la solution u^* par une suite $(u_k)_{k \in \mathbb{N}^*}$ suivant un algorithme type suivant :

$\mathcal{D}$

- Initialisation : u_0 donnée.
- Itération : $k \geq 0$. Pour d_k, ρ_k, u_k donnés,

$$u_{k+1} = u_k + \rho_k d_k,$$

avec d_k une *direction de descente* pour u_k.

On voit là que vu la définition de *direction descente*, on peut raisonnablement espérer que si ρ_k est suffisamment petit, à chaque itération, on trouve une nouvelle itération u_{k+1} telle que $\mathcal{J}(u_{k+1}) \leq \mathcal{J}(u_k)$. Et on paraît bien partis pour aller chercher le point de minimum !

Toute la difficulté alors est de bien choisir la direction de descente d_k à chaque itération k et la "longueur" (on dira plutôt *pas*) de descente ρ_k. Ces choix donnent lieu à différentes méthodes. Une grosse famille de méthodes repose sur un choix particulier de direction de descente relié au gradient de la fonctionnelle $\mathcal{J}$ elle-même.

Méthodes de gradients

On se base sur la constatation simple suivante. Supposons que $\mathcal{J} : \mathbb{R}^d \to \mathbb{R}$. Soit $u^0 \in \mathbb{R}^d$ tel que $\nabla\mathcal{J}(u^0)$ ne soit pas le vecteur nul. On approche $\mathcal{J}$ au voisinage de u^0 par son approximation affine de Taylor à l'ordre 1, $\mathcal{L}(u) := \mathcal{J}(u^0) + \langle \nabla\mathcal{J}(u^0), (u - u^0)\rangle$.

Si on choisit u sur la droite passant par u^0 et de coefficient directeur $\nabla\mathcal{J}(u^0)$, alors il existe $\alpha \in \mathbb{R}$ tel que u peut s'écrire $u(\alpha) = u^0 - \alpha\nabla\mathcal{J}(u^0)$ (on l'appelle alors $u(\alpha)$ pour insister sur la dépendance en α). Si on choisit $\alpha > 0$, alors $\mathcal{L}(u(\alpha)) = \mathcal{J}(u^0) - \alpha\|\nabla\mathcal{J}(u^0)\|^2 < \mathcal{J}(u^0)$.

Et comme par formule de Taylor, on sait que $\mathcal{J}(u(\alpha)) = \mathcal{L}(u(\alpha)) + o(\alpha\|\nabla\mathcal{J}(u^0)\|)$, alors pour α suffisamment petit, on aura $\mathcal{J}(u(\alpha)) < \mathcal{J}(u^0)$.

En résumé, on voit donc que **la direction opposée à celle du gradient est une direction de descente**.

Dans le cas des *méthodes de gradients*, on choisit $d_k = -\nabla\mathcal{J}(u_k)$ (ou une combinaison de tels gradients).

Reste à choisir le coefficient de descente ρ_k (que l'on appelle le **pas de descente**) pour les itérations. En général, on se concentre sur 3 choix possibles :

(a) $\rho_k = \rho$ une valeur constante donnée : c'est la *méthode de gradient à pas constant.*
(b) ρ_k variable à fixer au cours des itérations : c'est la *méthode de gradient à pas variable.*
(c) on peut choisir ρ_k de telle sorte à minimiser à chaque itération la fonction $f_k : \mathbb{R} \to \mathbb{R}$, $\rho \mapsto \mathcal{J}(u_k + \rho d_k)$: c'est la *méthode de gradient à pas optimal.*

L'algorithme de gradient à pas variable (b) s'écrit

— *Initialisation :* u^0 donné. On se donne $(\rho_k)_{k\in\mathbb{N}}$ une suite de $\mathbb{R}^+$,
— *Itération :* $k \in \mathbb{N}$, $u_{k+1} = u_k - \rho_k\nabla\mathcal{J}(u_k)$.

L'algorithme de gradient à pas variable, inclus la méthode de gradient à pas constant (a). C'est le cas particulier où on choisit $\rho > 0$ et que $\rho_k = \rho$ pour tout $k \in \mathbb{N}$.

L'algorithme de gradient à pas optimal (c) s'écrit

— *Initialisation :* u^0 donné.
— *Itération :* $k \in \mathbb{N}$, $u_{k+1} = u_k - \rho_k \nabla \mathcal{J}(u_k)$ *où* ρ_k *solution de* $\mathcal{J}(\rho_k) = \min_{\rho \in \mathbb{R}} \mathcal{J}(u_k - \rho \nabla \mathcal{J}(u_k))$.

Dans la suite $\langle \cdot, \cdot \rangle$ désignera toujours le produit scalaire euclidien sur $\mathbb{R}^d$ et $\| \cdot \|$ la norme associée.

Dans la suite du chapître on se concentrera sur ces trois types d'algorithmes. Vous aurez à les mettre en oeuvre en TP. Avant de se concentrer sur les théorèmes de convergence pour ces algorithmes, on va regarder des aspects pratiques de test d'arrêt des algorithmes.

III.2.3 Tests d'arrêt des algorithmes.

Un premier aspect pratique important est, lorsque l'on implémente les algorithmes (cf. TP), de savoir arrêter les algorithmes. Il faut donc des critères pratiques qui nous permettraient de savoir si on a obtenu au bout d'un certain nombre d'itérations une approximation u_k qui est satisfaisante au sens où elle est proche de u^*.

Il faut donc choisir des tests d'arrêts. Ici, on peut décider par exemple d'arrêter les itérations :

1. si
$$\|\nabla \mathcal{J}(u_k)\| < \tau_2,$$
avec $\tau_2 > 0$ petit, représentant un seuil que l'on fixera à l'avance. Typiquement, on peut prendre $\tau_2 = \varepsilon\|\nabla \mathcal{J}(u_0)\|$, avec ε petit. Il restera à vérifier que u_k trouvé est bien un minimum, car on pourrait arriver sur un point selle. On sait que le fait que le gradient s'annule n'est qu'une condition nécessaire !

Ne pas oublier d'ajouter le tests d'arrêt sur critère itération maximale.

III.2.4 Résultats de convergence des méthodes de gradients

Convergence de la méthode de gradient à pas variable et de la méthode de gradient à pas constant

On rappelle l'algorithme :

— *Initialisation :* u^0 donné. On se donne $(\rho_k)_{k\in\mathbb{N}}$ une suite de $\mathbb{R}^+$,
— *Itération :* $k \in \mathbb{N}$, $u_{k+1} = u_k - \rho_k \nabla \mathcal{J}(u_k)$.

Remarque III.2.2 *La méthode de gradient à pas variable, contient la méthode de gradient à pas constant. C'est le cas particulier où on choisit $\rho > 0$ et que $\rho_k = \rho$ pour tout $k \in \mathbb{N}$.*

Théorème III.2.3 *Soit $\alpha > 0$. Soit $\mathcal{J} : \mathbb{R}^d \to \mathbb{R}$ une fonctionnelle $\mathcal{C}^1$. On suppose que $\mathcal{J}$ est α-convexe et qu'il existe $M > 0$ tel que*

$$\|\nabla \mathcal{J}(v) - \nabla \mathcal{J}(u)\| \leq M\|v - u\|, \text{ pour tout } (u, v) \in \mathbb{R}^d \times \mathbb{R}^d. \tag{III.1}$$

Soit $u_0 \in \mathbb{R}^d$ et $(\rho_k)_{k\in\mathbb{N}}$ une suite de réels positifs donnés. On considère la méthode du gradient à pas variable pour $\mathcal{J}$ définie par l'initialisation u_0 et $(\rho_k)_{k\in\mathbb{N}}$.
S'il existe deux nombres réels a et b tels que $0 < a \leq \rho_k \leq b < \dfrac{2\alpha}{M^2}$ pour tout entier $k \geq 0$, alors la méthode du gradient à pas variable converge et il existe $0 \leq \beta < 1$ tel que pour tout $k \in \mathbb{N}$, $\|u_k - u^*\| \leq \beta^k \|u_0 - u^*\|$, où u^* est le point de minimum de la fonction $\mathcal{J}$ sur $\mathbb{R}^d$.

Preuve : cf. cours d'amphi.

Convergence de la méthode de gradient à pas optimal

L'algorithme de gradient à pas optimal s'écrit :

— *Initialisation :* u^0 donné.
— *Itération :* $k \in \mathbb{N}$, $u_{k+1} = u_k - \rho_k \nabla \mathcal{J}(u_k)$ *où* ρ_k *solution de* $\mathcal{J}(\rho_k) = \min_{\rho \in \mathbb{R}} \mathcal{J}(u_k - \rho \nabla \mathcal{J}(u_k))$.

Tout d'abord, l'algorithme de gradient à pas optimal a une propriété notable :
À chaque itération, les **deux directions successives de descente (**$d_k := \nabla \mathcal{J}(u_k)$ **et** $d_{k+1} := \nabla \mathcal{J}(u_{k+1})$ **pour un** $k \in \mathbb{N}$ **donné) sont orthogonales.** En effet, dans le calcul de ρ_k, on va chercher à minimiser la fonction d'une variable réelle $\rho \mapsto \mathcal{J}(u_k - \rho \nabla \mathcal{J}(u_k))$. La condition nécessaire d'optimalité nous dit donc que si ρ est un minimum, alors $\left(\frac{d}{d\rho}\mathcal{J}(u_k - \rho \nabla \mathcal{J}(u_k))\right)_{/\rho=\rho_k} = -{}^t\nabla \mathcal{J}(u_{k+1}) \nabla \mathcal{J}(u_k) = -\langle \nabla \mathcal{J}(u_{k+1}), \nabla \mathcal{J}(u_k) \rangle = 0$.

On peut établir comme pour le cas des deux autres méthodes un théorème de convergence.

Théorème III.2.4 *On suppose que* $\mathcal{J} : \mathbb{R}^d \to \mathbb{R}$ *est* $\mathcal{C}^1$ *et* α*-convexe, avec* $\alpha > 0$. *On suppose de plus que* $\nabla \mathcal{J}$ *est lipschitzien sur* $\mathbb{R}^d$, *i.e.*

$$\exists M > 0, \text{ tel que } \forall (v, w) \in \mathbb{R}^d \times \mathbb{R}^d, \|\nabla \mathcal{J}(v) - \nabla \mathcal{J}(w)\| \leq M \|v - w\|.$$

Soit $u_0 \in \mathbb{R}^d$ *donné. On définit la suite* $(u_k)_{k \in \mathbb{N}}$ *par l'algorithme de gradient à pas optimal avec initialisation* u_0. *Alors la méthode du gradient à pas optimal converge, i.e.* $\|u_k - u^*\| \to 0$ *lorsque* $k \to +\infty$, *si* u^* *est le point de minimum de* $\mathcal{J}$.

Preuve : cf. cours d'amphi.

Remarque III.2.5 *Ce résultat reste aussi valable lorque* $\nabla \mathcal{J}$ *est lipschitzien sur tout borné de* $\mathbb{R}^d$, *i.e.*

$$\forall M > 0, \exists C_M > 0, \|v\| + \|w\| \leq M \text{ implique } \|\mathcal{J}(v) - \mathcal{J}(w)\| \leq C_M \|v - w\|.$$

Il suffit avant d'utiliser l'hypothèse de Lipschitziannité dans la preuve d'utiliser que, comme $\mathcal{J}$ *est* α*-convexe, on sait qu'elle est "infinie à l'infinie", et donc* $(\mathcal{J}(u_k))_{k \in \mathbb{N}}$ *bornée implique* $(u_k)_{k \in \mathbb{N}}$ *bornée. Et la preuve se poursuit.*

III.2.5 Cas des fonctionnelles quadratiques.

On s'intéresse au cas où $\mathcal{J}(u) = \frac{1}{2}(Au, u) - (b, u) + c$, où A est une matrice symétrique, b un vecteur donné et c un réel donné. Dans ce cas, $\mathcal{J}$ est $\mathcal{C}^2$ sur $\mathbb{R}^d$ et [1] $\nabla \mathcal{J}(u) = Au - b$. Chercher un point critique revient donc à résoudre le système linéaire $Au = b$.

1. Faites le calcul !

Dans le cas où A une matrice symétrique définie et positive, on peut calculer explicitement le paramètre ρ_k dans la méthode de gradient à pas optimal. On se sert du fait que pour tout vecteur $u \in \mathbb{R}^d$, $\nabla \mathcal{J}(u) = Au - b$. En écrivant la relation d'orthogonalité entre les directions de descente, on trouve

$$^t(Au_{k+1} - b)(Au_k - b) =^t (A(u_k - \rho_k(Au_k - b)) - b)(Au_k - b).$$

Donc si on utilise la notation pour $u \in \mathbb{R}^d$,

$$\|u\|_A =^t uAu = \langle u, Au \rangle,$$

on trouve

$$\rho_k = \frac{\|Au_k - b\|^2}{\|Au_k - b\|_A^2}.$$

Dans ce cas, on peut un peu aussi préciser les estimations obtenues pour les méthodes de gradients. De plus, on peut développer une autre méthode : *la méthode du gradient conjugué* qui converge en au plus n itérations.

III.2.6 Newton ou Newton amélioré

Un façon de rechercher le minimum d'une fonction peut également être de rechercher ses points critiques (i.e. trouver les points $u \in \mathbb{R}^d$ qui vérifient $\nabla \mathcal{J}(u) = 0$.). Cela revient donc à chercher le zéro d'une fonction et on pense tout naturellement à utiliser une méthode de recherche de zéros de fonctions comme la méthode de Newton. Par contre, cela implique d'avoir pas mal d'hypothèses sur la fonctionnelle $\mathcal{J}$. En particulier, on aimerait être capable de lui appliquer le théorème classique de convergence de la méthode de Newton. On le rappelle ci-dessous.

Proposition III.2.6 *Soit F une fonction de classe $\mathcal{C}^2$ de $\mathbb{R}^d$ dans $\mathbb{R}^d$ et u un zéro de F tel que $DF(u)$ inversible. Alors il existe $\varepsilon > 0$, tel que, si $u^0 \in \mathbb{R}^d$, avec $\|u - u^0\| \leq \varepsilon$, alors la méthode de Newton converge, c'est à dire que la suite $(u^k)_{k\in\mathbb{N}}$ définie par*

— *Initialisation : u^0 donnée,*

— *Itération $k \in \mathbb{N}$: $u^{k+1} = u^k - DF(u^k)^{-1}F(u^k)$,*

converge vers u et il existe une constante $C > 0$ telle que pour tout $k \in \mathbb{N}$,

$$\|u^{k+1} - u\| \leq C\|u^k - u\|^2.$$

On cherche donc à appliquer l'algorithme de Newton à $F = \nabla \mathcal{J}$. Il est donc nécessaire de calculer la Hessienne et de l'inverser à chaque itération. Quand l'algorithme converge, il converge quadratiquement, donc très vite, par contre, il faut pouvoir partir proche de la solution au départ. Il se peut aussi que l'on converge, mais que l'on converge vers un maximum ou un point selle de $\mathcal{J}$, puisque l'on ne cherche "que" les points critiques de F. La méthode n'est donc pas la solution à tout, mais converge quand même quadratiquement lorsqu'elle converge.

La problématique de calcul de Hessien est aussi grande. Ce qui fait que certains algorithmes sont basés sur une approximation de la Hessienne à chaque étape. C'est ce que l'on appelle des *algorithmes de Newton modifiés* (*BFGS* notamment).

Chapitre IV

Problèmes avec contraintes en dimension finie

On cherche maintenant à résoudre le problème avec contraintes (i.e. $K \subsetneq \mathbb{R}^d$) : *Trouver* $u^* \in K$, *tel que*

$$\mathcal{J}(u^*) = \min_{u \in K} \mathcal{J}(u).$$

Le cadre est différent du chapître précédent puisqu'on impose des contraintes et qu'en particulier K n'est pas forcément un ouvert (et non nécessairement convexe).

On va se reposer les mêmes questions que pour le cas sans contraintes. Mais les réponses se compliquent... On va chercher à généraliser les équations de caractérisation déjà vues, elles se transforment en inéquations.

IV.1 Quelques conditions d'optimalité

Première remarque : Si K est ouvert, on dispose de pas mal de résultats des chapîtres précédents.

IV.1.1 Petits rappels préliminaires

(a) Soit U un ouvert de $\mathbb{R}^d$. Si $\mathcal{J} : U \to \mathbb{R}$ est une fonction différentiable en $u \in U$ alors pour tout $w \in \mathbb{R}^d$, il existe h_0 tel que si $h \leq h_0$, alors $u + hw \in U$ et

$$\frac{\mathcal{J}(u + hw) - \mathcal{J}(u)}{h} \tag{IV.1}$$

converge vers

$$\langle \nabla \mathcal{J}(u), w \rangle$$

lorsque $h \to 0$.

(b) On définit l'orthogonal d'un ensemble $A \subset \mathbb{R}^d$, et on le note $A^\perp = \{w \in \mathbb{R}^d, \text{ tel que } \langle w, a \rangle = 0, \forall a \in A\}$. C'est un sous-espace vectoriel de $\mathbb{R}^d$ (il est donc aussi fermé). De plus si A est un sous espace vectoriel de $\mathbb{R}^d$, alors $(A^\perp)^\perp = A$.

(c) On note pour une famille $(w_i)_{i\in\{1,...,l\}}$ de $l \in \mathbb{N}^*$ vecteurs de $\mathbb{R}^d$, $Vect((w_i)_{i\in\{1,...,l\}})$ l'espace vectoriel engendré par ces l vecteurs, i.e.

$$Vect((w_i)_{i\in\{1,...,l\}}) := \left\{ \sum_{i=1}^{l} \lambda_i w_i, \text{ tels que } \lambda_i \in \mathbb{R}, \forall i \in \{1, ..., l\} \right\} \subset \mathbb{R}^d.$$

IV.1.2 Cas d'un ensemble de contraintes convexe.

Commençons par regarder ce qu'il se passe si l'ensemble des contraintes est convexe.

Théorème IV.1.1 *Soit $u \in K$ convexe. On suppose que $\mathcal{J} : \mathbb{R}^d \to \mathbb{R}$ est différentiable en u. Si u est un point de minimum local de $\mathcal{J}$ sur K, alors :*

$$\langle \nabla \mathcal{J}(u), v - u \rangle \geq 0, \forall v \in K. \tag{IV.2}$$

Si $u \in K$ vérifie la précédente inégalité (IV.2) *et si $\mathcal{J}$ est convexe, alors u est un minimum global de $\mathcal{J}$ sur K.*

Remarque IV.1.2 *Cette inéquation s'appelle aussi inéquation d'Euler.*

Preuve : Soit $v \in K$ et $h \in]0, 1]$, on a $u + h(v - u) = (1 - h)u + hv \in K$, comme K est convexe. De plus, on sait que u est un minimum local de $\mathcal{J}$ sur K donc, il existe $\eta > 0$ tel que si $w \in K$ et $\|u - w\| \leq \eta$, alors $\mathcal{J}(w) \geq \mathcal{J}(u)$. En choisissant $h < \frac{\eta}{\|v-u\|}$ (ce qui est possible puisque $u \neq v$, comme $h > 0$), on en déduit que $\|u + h(v - u) - u\| = h\|v - u\| < \eta$. Ainsi, pour tout $0 < h < \frac{\eta}{\|v-u\|}$, on a

$$\mathcal{J}(u + h(v - u)) \geq \mathcal{J}(u).$$

Et donc pour tout $0 < h < \frac{\eta}{\|v-u\|}$:

$$\frac{\mathcal{J}(u + h(v - u)) - \mathcal{J}(u)}{h} \geq 0.$$

En faisant tendre h vers 0, on obtient le résultat (voir le rappel préliminaire).
Démontrons maintenant la deuxième partie du théorème. Supposons donc que $u \in K$ vérifie (IV.2) et que $\mathcal{J}$ est convexe. On utilise une des caractérisation la convexité pour en déduire directement le résultat. En effet, on sait par la proposition II.4.2 et la remarque II.4.3 que, si $\mathcal{J}$ est convexe et que K est convexe, alors pour tout $v \in K$:

$$\mathcal{J}(v) \geq \mathcal{J}(u) + (\nabla \mathcal{J}(u), v - u), \forall v \in K.$$

Le résultat est donc immédiat avec cette inégalité. □

Remarque IV.1.3 *Si $K = \mathbb{R}^d$, on retrouve une proposition déjà vue où on a même l'égalité. En effet dans ce cas là, donnons nous $w \in \mathbb{R}^d$, en posant $v = w + u \in \mathbb{R}^d$, on a $\langle \nabla \mathcal{J}(u), w \rangle \geq 0$. Et donc on obtient pour tout $w \in \mathbb{R}^d$, $\langle \nabla \mathcal{J}(u), w \rangle \geq 0$. En appliquant l'inégalité précédente à $-w \in \mathbb{R}^d$ (si $w \in \mathbb{R}^d$), on a aussi pour tout $w \in \mathbb{R}^d$, $\langle \nabla \mathcal{J}(u), w \rangle \leq 0$. Et donc en conclusion, pour tout $w \in \mathbb{R}^d$, $\langle \nabla \mathcal{J}(u), w \rangle = 0$, ce qui donne $\nabla \mathcal{J}(u) = 0$.*

IV.1.3 Cas de contraintes d'égalités.

On suppose que $\mathcal{J} : \mathbb{R}^d \to \mathbb{R}$ ($d \in \mathbb{N}^*$), continue et $g : \mathbb{R}^d \to \mathbb{R}^p$ ($p \in \mathbb{N}^*$) et $p \leq d$ telle que pour tout $u \in \mathbb{R}^d$, $g(u) = (g_1(u), g_2(u), ..., g_p(u)) \in \mathbb{R}^p$ et pour tout $i \in \{1, ..., p\}$, $g_i : \mathbb{R}^d \to \mathbb{R}$. On suppose dans cette section que l'espace des contraintes K s'écrit :

$$K := \{u \in \mathbb{R}^d, \quad \text{tels que } g_1(u) = ... = g_p(u) = 0\}. \tag{IV.3}$$

⚠ *K n'est pas alors forcément convexe ni ouvert !*

Souvenez vous de la démonstration de la proposition B.2.1 ou du théorème précédent IV.1.1. On aimerait pouvoir répéter la même stratégie ici. Mais le problème dans cette démonstration est que contrairement au cas ouvert ou convexe, on ne peut pas assurer que si $u \in K$ et $w \in \mathbb{R}^d$ (la direction), alors $u + tw$ est encore dans K pour $t \in \mathbb{R}$ suffisamment petit [1]. Pour récupérer ce genre de résultat, on va être amené à limiter les directions w dans lesquelles on se déplace. On va seulement avoir le droit de le faire dans ce qui s'appelle un *cône de directions admissibles*. Ces directions admissibles n'assureront pas forcément que $u + tw$ est dans K pour t suffisamment petit, mais qu'on n'en est pas loin.
Pour formaliser un peu, en gros, l'idée est alors de considérer les directions $w \in \mathbb{R}^d$ telles que :

$$u + t(w + \varepsilon(t)) \in K, \text{ pour } t \text{ suffisamment petit,}$$

avec

$$\lim_{t \to 0} \varepsilon(t) = 0, \text{ avec } \varepsilon(t) \in \mathbb{R}^d.$$

Si on considère une telle direction, et si u est un point de minimum local de $\mathcal{J}$ sur K, alors on sait que pour t assez petit, on a

$$\frac{\mathcal{J}(u + t(w + \varepsilon(t)) - \mathcal{J}(u)}{t} \geq 0. \tag{IV.4}$$

et donc si $\mathcal{J}$ est différentiable

$$\langle \nabla \mathcal{J}(u), w \rangle \geq 0, \tag{IV.5}$$

si w est dans cet ensemble de directions admissibles. Et on aura récupéré une des étapes de la stratégie de démonstration.

Le problème est donc ensuite d'identifier ces directions admissibles. En posant $\varphi : t \mapsto u + t(w + \varepsilon(t))$, on obtient $\varphi(0) = u$ et $\varphi'(0) = w$ et pour t assez petit $\varphi(t) \in K$. La fonction φ représente une courbe paramétrée tracée sur K. Un autre façon d'interpréter ce qui précède est donc de dire qu'on peut donc construire une courbe paramétrée au voisinage de $u \in K$ qui est une courbe incluse dans K. Dans cette idée, on donne la définition suivante.

Definition IV.1.4 *Soit $u \in K$. On introduit l'ensemble $T_K(u)$ donné par*

$$T_K(u) = \{w \in \mathbb{R}^d \text{ tels qu'il existe un intervalle ouvert } I \text{ contenant } 0 \text{ et une fonction } \varphi : I \to K, \mathcal{C}^1, \text{ telle que } \varphi(0) = u \text{ et } \varphi'(0) = w\}. \tag{IV.6}$$

1. on peut donner une "image" vecteur $u + tv$, en disant que l'on part de u et on se déplace dans la direction w sur une distance de t. Vous pouvez vous le représenter avec un dessin, par exemple !

On peut ensuite montrer que dans le cas de contraintes d'égalités cet ensemble est relié aux gradients des fonctions qui constituent les contraintes.

Théorème IV.1.5 *Soient pour $i \in \{1, ..., p\}$, $g_i : \mathbb{R}^d \to \mathbb{R}$, p fonctions $\mathcal{C}^1$. On suppose que l'espace K est définit par :*

$$K := \{u \in \mathbb{R}^d, \quad \text{tels que } g_1(u) = ... = g_p(u) = 0\}.$$

Alors, pour tout $u \in K$, on a

$$T_K(u) \subset \left\{w \in \mathbb{R}^d, \langle \nabla g_j(u), w\rangle = 0, \, \forall j \in \{1, ..., p\}\right\}. \tag{IV.7}$$

Autrement dit $T_K(u) \subset (Vect((\nabla g_j(u))_{j\in\{1,...,p\}}))^\perp$.

Si de plus les vecteurs $(\nabla g_j(u))_{j\in\{1,...,p\}}$ forment une famille libre, alors on a même l'égalité

$$T_K(u) = \left\{w \in \mathbb{R}^d, \langle \nabla g_j(u), w\rangle = 0, \, \forall j \in \{1, ..., p\}\right\}. \tag{IV.8}$$

Autrement dit $T_K(u) = (Vect((\nabla g_j(u))_{j\in\{1,...,p\}}))^\perp$.

Preuve : Soient $u \in K$ et $w \in T_K(u)$. Vu la définition de $T_K(u)$, on sait qu'on peut trouver un intervalle ouvert I contenant 0 et $\varphi : I \to K$, $\mathcal{C}^1$, telle que $\varphi(0) = u$ et $\varphi'(0) = w$. Alors comme φ est à valeurs dans K, on a $\forall j \in \{1, ..., p\}$, $\forall t \in I$, $g_j(\varphi(t)) = 0$, i.e. $g_j \circ \varphi(t) = 0$. Comme I est un intervalle ouvert contenant 0 et que φ et g_j sont deux fonctions $\mathcal{C}^1$, on peut dériver en 0. Ainsi, en dérivant la fonction $g_j \circ \varphi$ en 0, on obtient [2] pour tout $j \in \{1, ..., p\}$:

$$(g_j \circ \varphi)'(0) = \langle \nabla g_j(\varphi(0)), \varphi'(0)\rangle. \tag{IV.9}$$

Ce qui donne pour tout $j \in \{1, ..., p\}$:

$$(g_j \circ \varphi)'(0) = \langle \nabla g_j(u), w\rangle.$$

Et donc comme de plus $g_j \circ \varphi \equiv 0$, sur I ouvert contenant 0, on sait aussi en particulier que $(g_j \circ \varphi)'(0) = 0$. Et donc pour tout $j \in \{1, ..., p\}$, $\langle \nabla g_j(u), w\rangle$. Ce qui donne la première inclusion.

Montrons maintenant que $\left\{w \in \mathbb{R}^d, \langle \nabla g_j(u), w\rangle = 0, \, \forall j \in \{1, ..., p\}\right\} = (Vect((\nabla g_j(u))_{j\in\{1,...,p\}}))^\perp$. Soit $w \in \left\{w \in \mathbb{R}^d, \langle \nabla g_j(u), w\rangle = 0, \, \forall j \in \{1, ..., p\}\right\}$. Si $v \in Vect((\nabla g_j(u))_{j\in\{1,...,p\}})$, alors il existe $(\lambda_i)_{i\in\{1,...,p\}} \in \mathbb{R}^p$ tel que $v = \sum_{i=1}^p \lambda_i \nabla g_i(u)$. Donc $\langle v, w\rangle = \sum_{i=1}^p \lambda_i \langle \nabla g_j(u), w\rangle = 0$ et donc $w \in (Vect((\nabla g_j(u))_{j\in\{1,...,p\}}))^\perp$. Réciproquement, si $w \in (Vect((\nabla g_j(u))_{j\in\{1,...,p\}}))^\perp$, alors en particulier pour tout $j \in \{1, ..., p\}$, $\langle w, \nabla g_j(u)\rangle = 0$ puisque $\nabla g_j(u) \in Vect((\nabla g_j(u))_{j\in\{1,...,p\}})$. Donc $w \in \left\{w \in \mathbb{R}^d, \langle \nabla g_j(u), w\rangle = 0, \, \forall j \in \{1, ..., p\}\right\}$.

La fin de cette démonstration ne s'adresse qu'aux personnes ayant déjà vu le théorème des Fonctions implicites. Pour les autres, passez et admettez la deuxième partie du

2. dérivation de fonctions composées

théorème.

Montrons l'inclusion réciproque dans le cas où les contraintes sont linéairement indépendantes (au sens de leurs vecteurs gradient). Soit donc $w \in \mathbb{R}^d$ tel que $\forall j \in \{1, ..., p\}$,

$$\langle \nabla g_j(u), w \rangle = 0. \tag{IV.10}$$

Choisissons $(v_{p+1}, ..., v_d) \in \mathbb{R}^d$ tels que $(\nabla g_1(u), ..., \nabla g_p(u), v_{p+1}, ..., v_d)$ soit une base de $\mathbb{R}^d$ (on sait le faire, on est en dimension finie et on complète une famille de vecteurs libres en une base). Définissons $F : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ telle que

$$F(x,t) = (g_1(x), ..., g_p(x), \langle v_{p+1}, x - u - tw \rangle, ..., \langle v_d, x - u - tw) \rangle.$$

L'application F est de classe $\mathcal{C}^1$. On a de plus $F(u, 0) = 0$. On va chercher à appliquer le théorème des fonctions implicites pour pouvoir, au voisinage de $(u, 0)$, "exprimer x en fonction de t". Pour cela on étudie la matrice des différentielles partielles "par rapport aux variables x", i.e. on étudie la matrice $\mathbb{J}_x(u, 0)$ donnée par

$$\mathbb{J}_x(u,0) = \left(\left(\frac{\partial F_i}{\partial x_j}(u,0) \right)_{i \in \{1,...,d\},\, j \in \{1,...,d\}} \right). \tag{IV.11}$$

Si cette matrice est inversible, on pourra appliquer le théorème des fonctions implicites. Or on trouve

$$\mathbb{J}_x(u,0) = \begin{pmatrix} {}^t\nabla g_1(u) \\ \vdots \\ {}^t\nabla g_p(u) \\ {}^t v_{p+1} \\ \vdots \\ {}^t v_d \end{pmatrix} \tag{IV.12}$$

Cette matrice est inversible par construction des vecteurs qui constituent les lignes de la matrice (ils forment une base de $\mathbb{R}^d$). On sait donc, par le théorème des fonctions implicites, qu'il existe un intervalle ouvert I contenant 0 et $\varphi : I \to \mathbb{R}^d$, telle que $\varphi(0) = u$, $\mathcal{C}^1$ telle que $F(t, \varphi(t)) = 0$ pour tout $t \in I$. Et donc en particulier, pour tout $t \in I$,

$$g_1(\varphi(t)) = ... = g_p(\varphi(t)) = 0.$$

Et donc $\varphi(I) \subset K$.
En dérivant les relations en t, on obtient

$$\langle \nabla g_j(\varphi(t)), \varphi'(t) \rangle = 0, \forall j \in \{1, ..., p\}.$$

En particulier

$$\langle \nabla g_i(u), \varphi'(0) \rangle = 0.$$

De plus on a (IV.10), donc pour tout $j \in \{1, ..., p\}$, $\langle \nabla g_j(u), w \rangle = 0$.
En combinant ces deux relations, on obtient pour tout $j \in \{1, ..., p\}$,

$$\langle \nabla g_i(u), \varphi'(0) - w \rangle = 0. \tag{IV.13}$$

De plus en exploitant encore que $F(\varphi(t),t)=0$ pour tout $t\in I$, on a

$$\langle v_j, \varphi(t)-u-tw\rangle = 0, \forall j\in\{p+1,...,d\},$$

pour tout $t\in I$. On en déduit donc en dérivant, que $\langle v_j,\varphi'(t)-w\rangle=0$, $\forall j\in\{p+1,...,d\}$. En écrivant ces égalités en $t=0$, on obtient, que pour tout $j\in\{p+1,...,d\}$,

$$\langle v_j,\varphi'(0)-w\rangle = 0. \tag{IV.14}$$

et donc le vecteur $\varphi'(0)-w\in\mathbb{R}^d$ est orthogonal à tous les vecteurs de la base de $\mathbb{R}^d$ choisie, on en déduit que :

$$\varphi'(0)=w.$$

Ce qui nous donne finalement le résultat voulu : $w\in T_K(u)$. □

On va maintenant pouvoir établir un résultat donnant une condition nécessaire d'optimalité.

Théorème IV.1.6 *Soit $\mathcal{J}:\mathbb{R}^d\to\mathbb{R}$ et $u^*\in K$. Supposons que $\mathcal{J}$ est différentiable en u^* et que les fonctions $(g_i)_{i\in\{1,...,p\}}$ sont toutes $\mathcal{C}^1$ dans un voisinage de u^*. On suppose que la famille de vecteurs $(\nabla g_j(u^*))_{j\in\{1,...,p\}}$ est libre. Si u^* est un point de minimum local de $\mathcal{J}$ sur K, alors il existe un unique $\lambda^*=(\lambda_1^*,\lambda_2^*,...,\lambda_p^*)\in\mathbb{R}^p$ tel que :*

$$\nabla\mathcal{J}(u^*)+\sum_{i=1}^{p}\lambda_i^*\nabla g_i(u^*)=0. \tag{IV.15}$$

Le vecteur $\lambda^*\in\mathbb{R}^p$ est alors appelé multiplicateur de Lagrange.

Preuve : Soit $w\in T_K(u^*)$, et soit $\varphi: I\to K$, tel que $\varphi(0)=u^*$ et $\varphi'(0)=w$. Par continuité de φ sur I, on sait qu'on peut trouver $\tilde{I}$ intervalle ouvert inclus dans I contenant 0 et tel que pour tout $t\in\tilde{I}$, $\mathcal{J}(\varphi(t))\geq\mathcal{J}(u^*)$ [3]. La fonction $\mathcal{J}\circ\varphi$ a donc un minimum local en 0, et donc $(\mathcal{J}\circ\varphi)'(0)=0$. Or $(\mathcal{J}\circ\varphi)'(0)=\langle\nabla\mathcal{J}(\varphi(0)),\varphi'(0)\rangle=\langle\nabla\mathcal{J}(u^*),w\rangle$. Donc $\langle\nabla\mathcal{J}(u^*),w\rangle=0$. On a donc

$$\nabla\mathcal{J}(u^*)\in(T_K(u^*))^{\perp}.$$

Mais $T_K(u^*)=(Vect((\nabla g_j(u^*))_{j\in\{1,...,p\}}))^{\perp}$, comme la famille de vecteurs $(\nabla g_j(u^*))_{j\in\{1,...,p\}}$ est libre (cf. théorème IV.1.5). Donc (voir rappels sur les orthogonaux au début de la section)

$$(T_K(u^*))^{\perp}=Vect((\nabla g_j(u^*))_{j\in\{1,...,p\}}).$$

et donc on a l'existence $\lambda^*\in\mathbb{R}^p$ vérifiant (IV.15). On obtient l'unicité de λ^* puisque les p gradients des contraintes en u^* sont indépendants [4].
□

Les réels $(\lambda_i)_{i\in\{1,...,n\}}$ sont appelés *multiplicateurs de Lagrange* associé au problème de minimisation sous contraintes.

3. Comme u^* est un minimum local, on sait qu'il existe $\eta>0$ tel que pour tout $w\in K$, si $\|u^*-w\|\leq\eta$, alors $\mathcal{J}(w)\geq\mathcal{J}(u^*)$. Comme $\varphi(0)=u^*$, et que φ est continue sur tout I contenant 0, il existe un voisinage $\tilde{I}$ de 0, tel que $\|\varphi(t)-u^*\|\leq\eta$. Et comme φ est à valeur dans K on a le résultat.

4. Supposez qu'il existe deux tels λ et aboutir à une combinaison linéaire nulle des vecteurs $(\nabla g_j(u^*))_{j\in\{1,...,p\}}$

Lorsque la famille de vecteurs $(\nabla g_j(u^*))_{j\in\{1,...,p\}}$ est libre, on dit qu'on est dans **un cas régulier** (ou u^* est un point régulier). Sinon, on dira qu'on est dans **un cas non régulier** (ou u^* est un point non régulier).

Interprétation en terme de Lagrangien. Si on introduit la fonction $\mathcal{L} : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$, $(u,\lambda) \mapsto \mathcal{J}(u)+\sum_{i=1}^{p} \lambda_i g_i(u) = \mathcal{J}(u)+\langle \lambda, g(u)\rangle$ appelée *Lagrangien* associé au problème de minimisation sous contraintes. Alors si u^* est un minimum local de $\mathcal{J}$ sur K et que les gradients des contraintes forment une famille libre, il existe $\lambda^* \in \mathbb{R}^p$ tel que :

$$\frac{\partial \mathcal{L}}{\partial u}(u^*,\lambda^*) = 0 \,\text{et}\, \frac{\partial \mathcal{L}}{\partial \lambda}(u^*,\lambda^*) = 0. \tag{IV.16}$$

Ce qui est simplement une réécriture du résultat du théorème précédent.

IV.1.4 Contraintes d'inégalités.

On suppose maintenant que l'espace des contraintes est donné par des contraintes de type inégalité. Autrement dit, soit $m \in \mathbb{N}^*$, $m \leq d$,

$$K := \left\{ u \in \mathbb{R}^d,\, h_i(u) \leq 0, \forall i \in \{1,...,m\} \right\}, \tag{IV.17}$$

avec pour $i \in \{1,...,m\}$, $h_i : \mathbb{R}^d \to \mathbb{R}$ des fonctions $\mathcal{C}^0$.

La situation se complique encore un peu plus. On a toujours envie de trouver un moyen de se balader autour d'un potentiel minimum local pour pouvoir tester s'il est optimal ou non. On n'a pas forcément de problème si $h_i(u) < 0$ pour tout $i \in \{1,...,p\}$, car dans ce cas là, grâce à la continuité des fonctions $(h_i)_{i\in\{1,...,p\}}$, on sait que dans un voisinage de ce point on est sûr qu'on a encore $h_i(v) \leq 0$, pour tout v dans ce voisinage. Cette contrainte ne pose donc en fait pas de problème. On donne le vocabulaire suivant qui suit cette remarque.

Definition IV.1.7 *Soit $u \in K$. On dit qu'une contrainte i (avec $i \in \{1,...,m\}$) est active en u si $h_i(u) = 0$. Elle est dite inactive en u, sinon. L'ensemble $I(u) = \{i \in \{1,...,m\},\, h_i(u) = 0\}$, est appelé ensemble des contraintes actives en u.*

On essaie de généraliser le résultat du théorème IV.1.5. On dispose du Lemme de Farkas que l'on admettra dans ce cours.

Lemma IV.1.8 (admis) *Soient $M \in \mathbb{N}^*$ et $a_1,...,a_M$, M éléments de $\mathbb{R}^d$. On considère les ensembles*

$$\mathcal{Q} = \left\{ v \in \mathbb{R}^d,\, \langle a_i, v\rangle \leq 0,\, \textit{pour } i \in \{1,...,M\} \right\},$$

et

$$\hat{\mathcal{Q}} = \left\{ v \in \mathbb{R}^d, \exists \lambda_1,...,\lambda_M \geq 0,\, \textit{tels que } v = -\sum_{i=1}^{M} \lambda_i a_i \right\}.$$

Alors pour tout $p \in \mathbb{R}^d$, on a : si $\langle p, w\rangle \geq 0$, pour tout $w \in \mathcal{Q}$, alors $p \in \hat{\mathcal{Q}}$. La réciproque est également vraie.

On voit que ce Lemme est une sorte de généralisation du théorème IV.1.5 au cas d'inégalités (en remplaçant les a_i par les gradient des contraintes).

À l'aide de ces définitions et du Lemme précédent, on peut là aussi établir une condition nécessaire d'optimalité.

Théorème IV.1.9 *On suppose que K est donné par des contraintes d'inégalité comme ci-dessus. Soit $u^* \in K$. On suppose que les fonctions $\mathcal{J}$ et $(h_i)_{i\in\{1,...,m\}}$ sont $\mathcal{C}^1$ dans un voisinage de $u^* \in K$ et que la famille $(\nabla h_i(u^*))_{i\in I(u^*)}$ est libre. Alors, si u^* est minimum local de $\mathcal{J}$ sur K, il existe λ_1^*, λ_2^*, ..., $\lambda_m^* \geq 0$, appelés multiplicateurs de Lagrange, tels que :*

$$\nabla \mathcal{J}(u^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(u^*) = 0,$$

$$\text{avec } \lambda_i^* \geq 0, \text{ et } \lambda_i^* = 0, \text{ si } h_i(u^*) < 0, \forall i \in \{1, ..., m\}.$$

On peut en fait établir un théorème plus général qui est valable sous des hypothèses plus générales. C'est dans ce cas l'hypothèse que la famille $(\nabla h_i(u^*))_{i\in I(u^*)}$ est libre qui va être impactée. On va remplacer cette hypothèse par une hypothèse de contraintes dites qualifiées. C'est toujours une modification "légère" de l'approche du cas de contraintes d'égalités pour trouver une façon de se balader et on définit la notion de **contraintes qualifiées** pour les contraintes actives.

Definition IV.1.10 *On dit que les contraintes sont qualifiées en $u \in K$ si chaque h_i pour $i \in \{1, ..., m\}$ est dérivable en u et qu'il existe une direction $\tilde{w} \in K$ telle que l'on ait pour tout $i \in I(u)$,*

(a) ou bien $\langle \nabla h_i(u), \tilde{w} \rangle < 0$,

ou

(b) ou bien $\langle \nabla h_i(u), \tilde{w} \rangle = 0$ et h_i est affine.

On peut faire quelques commentaires sur cette définition. Pour (a), ce que l'on voit c'est que la direction $\tilde{w}$ est en quelque sorte une direction de descente pour u relativement à chaque h_i, et on reste donc dans K. En effet, formellement, on peut écrire au voisinage de $t = 0$, $h_i(u + t\tilde{w}) = h_i(u) + t\langle \nabla h_i(u), \tilde{w} \rangle + o(t)$. Donc si (a) est vérifiée, on voit qu'on peut trouver $t > 0$ suffisamment petit pour que si $h_i(u) = 0$, alors $h_i(u + t\tilde{w}) < 0$. Et donc $u + t\tilde{w} \in K$. Pour (b), si toutes les contraintes sont affines, on voit qu'on peut tout simplement prendre $\tilde{w} = 0$ et les contraintes sont automatiquement qualifiées. En effet, supposons que h_i est affine, alors on peut l'écrire $h_i(u) = \langle a_i, u \rangle + b_i$, avec $a_i \in \mathbb{R}^d$ et $b_i \in \mathbb{R}$. Donc si $h_i(u) = 0$ pour toute direction $w \in \mathbb{R}^d$ et tout $t \in \mathbb{R}$, on a $h_i(u + tw) = \underbrace{\langle a_i, u \rangle + b_i}_{h_i(u)=0} + t\langle a_i, w \rangle$. Donc pour garantir que $h_i(u + tw) \leq 0$ sachant que $h_i(u) = 0$, il suffit de prendre $w = 0$. On distingue ces deux cas dans la définition car les contraintes affines sont qualifiées sous des conditions moins strictes et la définition mérite d'être adaptée à ce cas vu l'importance des contraintes affines dans les applications.

Le théorème devient alors :

Théorème IV.1.11 *On suppose que K est donné par des contraintes d'inégalité comme ci-dessus. Soit $u^* \in K$. On suppose que les fonctions $\mathcal{J}$ et $(h_i)_{i\in\{1,...,m\}}$ sont dérivables en $u^* \in K$ et que les*

contraintes sont qualifiées en $u^* \in K$. Alors, si u^* est minimum local de $\mathcal{J}$ sur K, il existe λ_1^*, λ_2^*, ..., $\lambda_m^* \geq 0$, appelés multiplicateurs de Lagrange, tels que :

$$\nabla \mathcal{J}(u^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(u^*) = 0,$$

$$\textit{avec } \lambda_i^* \geq 0, \textit{ et } \lambda_i^* = 0, \textit{ si } h_i(u^*) < 0, \forall i \in \{1, ..., m\}.$$

Preuve : Notons $\tilde{K}(u^*) := \left\{ w \in \mathbb{R}^d, \langle \nabla h_i(u^*), w \rangle \leq 0, \forall i \in I(u^*) \right\}$. On se donne alors $\tilde{w}$ permettant de vérifier la qualification des contraintes. On se donne également $w \in \tilde{K}(u^*)$ et $\delta > 0$. Montrons que $u^* + \varepsilon(w + \delta\tilde{w}) \in K$ pour tout $\varepsilon > 0$ assez petit. On examine trois cas de figure.

1. Si $i \notin I(u^*)$, on a $h_i(u^*) < 0$ et $h_i(u^* + \varepsilon(w + \delta\tilde{w})) < 0$, par continuité de h_i et si ε est assez petit (puisque dans ce cas $u^* + \varepsilon(w + \delta\tilde{w})$ est dans un voisinage de u^*).
2. Si $i \in I(u^*)$ et $\langle \nabla h_i(u^*), \tilde{w} \rangle < 0$, alors

$$\begin{aligned} h_i(u^* + \varepsilon(w + \delta\tilde{w})) &= h_i(u^*) + \varepsilon \langle \nabla h_i(u^*), w + \delta\tilde{w} \rangle + o(\varepsilon), & \text{(IV.18)} \\ &\leq \varepsilon\delta \langle \nabla h_i(u^*), \tilde{w} \rangle + o(\varepsilon) & \text{(IV.19)} \end{aligned}$$

   Et donc $h_i(u^* + \varepsilon(w + \delta\tilde{w})) < 0$, pour $\varepsilon > 0$ assez petit.
3. Si $i \in I(u^*)$ et $\langle \nabla h_i(u^*), \tilde{w} \rangle = 0$, alors h_i est affine et

$$\begin{aligned} h_i(u^* + \varepsilon(w + \delta\tilde{w})) &= h_i(u^*) + \varepsilon \langle \nabla h_i(u^*), w + \delta\tilde{w} \rangle, & \text{(IV.20)} \\ &\leq \varepsilon \langle \nabla h_i(u^*), w \rangle \leq 0. & \text{(IV.21)} \end{aligned}$$

Donc finalement, si u^* est un minimum local de $\mathcal{J}$ sur K, on déduit de ce qui précède que pour tout $\delta > 0$ si $\varepsilon > 0$ est suffisamment petit $u^* + \varepsilon(w + \delta\tilde{w}) \in K$. Et donc on peut écrire

$$\mathcal{J}(u^* + \varepsilon(w + \delta\tilde{w})) - \mathcal{J}(u^*) \geq 0. \tag{IV.22}$$

En divisant par $\varepsilon > 0$ et en faisant tendre $\varepsilon \to 0^+$, on trouve pour tout $w \in \tilde{K}(u^*)$, $\forall \delta > 0$

$$\langle \nabla \mathcal{J}(u^*), w + \delta\tilde{w} \rangle \geq 0. \tag{IV.23}$$

On peut alors faire tendre $\delta \to 0^+$ et on trouve pour tout $w \in \tilde{K}(u^*)$,

$$\langle \nabla \mathcal{J}(u^*), w \rangle \geq 0. \tag{IV.24}$$

Le lemme de Farkas appliqué à $p = \nabla \mathcal{J}(u^*)$ et $a_i = \nabla h_i(u^*)$ pour $i \in I(u^*)$ (dans ce cas $\mathcal{Q} = \tilde{K}(u^*)$) permet de conclure. □

Une remarque pratique sur les contraintes qualifiées : si la famille $(\nabla h_i)_{i\in I(u^*)}$ est libre alors les contraintes sont qualifiées !

En effet, les conditions de qualification sont des conditions qui sont suffisantes pour pouvoir se balader dans K à partir d'un point u de K. Les conditions de qualification peuvent être parfois difficiles à vérifier en pratique. Donnons quelques cas particuliers où ces conditions sont vérifiées.

Ce qu'on voit tout d'abord, c'est qu'en fait ce ne sont que les contraintes actives qui jouent un rôle dans la vérification des contraintes de qualification et la condition nécessaire d'optimalité puisque pour les contraintes inactives $i \notin I(u^*)$, on a $\lambda_i^* = 0$. Et les contraintes actives sont en fait des contraintes d'égalité en ce point. On semble donc ramené au cas de contraintes d'égalité et il paraît alors naturel de se demander si quand la famille $(\nabla h_i(u^*))_{i\in I(u^*)}$ est une famille libre, les contraintes sont qualifiées. Et c'est le cas ! En effet, supposons que la famille $(\nabla h_i(u^*))_{i\in I(u^*)}$ est libre. Posons alors

$$\tilde{w} = \sum_{j\in I(u^*)} \alpha_j \nabla h_j(u^*).$$

On cherche à déterminer les α_j pour $i \in I(u^*)$ pour que $\langle \nabla h_i(u^*), \tilde{w}\rangle = -1$ (car dans ce cas on aura trouvé une direction qui satisfait la condition de qualification). Cela est possible car écrire pour tout $i \in I(u^*)$, $\langle \nabla h_i(u^*), \tilde{w}\rangle = -1$ revient à écrire pour tout $i \in I(u)$,

$$\sum_{j\in I(u^*)} \alpha_j \langle \nabla h_i(u^*), \nabla h_j(u^*)\rangle = -1. \tag{IV.25}$$

Si on note maintenant $A = (a_{ij})_{(i,j)\in I(u^*)\times I(u^*)}$ la matrice de taille $card(I(u^*)) \times card(I(u^*))$, avec pour tout $(i,j) \in I(u^*) \times I(u^*)$,

$$a_{ij} = (\langle \nabla h_i(u^*), \nabla h_j(u^*)\rangle),$$

alors (IV.25) se réécrit

$$A\boldsymbol{\alpha} = V, \tag{IV.26}$$

avec $\boldsymbol{\alpha} = (\alpha_j)_{j\in I(u^*)}$ et $V = \begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix}$. Comme la famille $(\nabla h_i(u^*))_{i\in I(u^*)}$ est libre, on sait que la matrice A est inversible [5] et donc il existe un unique $\boldsymbol{\alpha}$ vérifiant cette égalité.

Avec ce choix de $\boldsymbol{\alpha}$, on a bien $\tilde{w}$ qui vérifie la condition de qualification.

Cas convexe

Lorsque les fonctions $\mathcal{J}$, h_1, ..., h_m sont convexes, la condition nécessaire devient une condition suffisante.

5. Supposons que $(v_i)_{i\in\{1,...,l\}}$ ($l \in \mathbb{N}^*$) est une famille libre et notons $B = (\langle v_i, v_j\rangle)_{(i,j)\in\{1,...,l\}^2}$ est la matrice de taille $l \times l$. Soit $\boldsymbol{\beta} = (\beta_k)_{k\in\{1,...,l\}} \in \mathbb{R}^l$, on peut montrer que $B\boldsymbol{\beta} = 0$ implique $\boldsymbol{\beta} = 0$. En effet pour tout $i \in \{1, ..., l\}$, $(B\boldsymbol{\beta})_i = \sum_{k=1}^l \beta_k \langle v_i, v_k\rangle = \langle v_i, \sum_{k=1}^l \beta_k v_k\rangle$. Puis en multipliant l'égalité i par β_i et en sommant pour $i \in \{1, ..., l\}$, on trouve $\| \sum_{i=1}^l \beta_i v_i\|^2 = 0$ et donc comme la famille $(v_i)_{i\in\{1,...,l\}}$ est libre, on déduit que $\boldsymbol{\beta} = 0$. La matrice carrée B est donc inversible.

Théorème IV.1.12 *On suppose que K est donné par des contraintes d'inégalité comme ci-dessus, et que les fonctions $\mathcal{J}$ et $(h_i)_{i\in\{1,...,m\}}$ sont convexes et $\mathcal{C}^1$. Si il existe λ_1^*, λ_2^*, ..., $\lambda_m^* \geq 0$, appelés multiplicateurs de Lagrange, tels que :*

$$\nabla\mathcal{J}(u^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(u^*) = 0,$$

$$\lambda_i^* \geq 0,\ \lambda_i^* = 0,\ si\ h_i(u^*) < 0,\ \forall i \in \{1,...,m\},$$

alors, u^* est un minimum global de $\mathcal{J}$ sur K.

Preuve : Par une des caractérisation des fonctions convexes, on a pour tout $u \in \mathbb{R}^d$,

$$\begin{aligned} \mathcal{J}(u) + \sum_{i=1}^{m} \lambda_i^* h_i(u) &\geq \mathcal{J}(u^*) + \sum_{i=1}^{m} \lambda_i^* h_i(u^*) + (\nabla\mathcal{J}(u^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(u^*), u - u^*)), \quad \text{(IV.27)} \\ &\geq \mathcal{J}(u^*), \quad \text{(IV.28)} \end{aligned}$$

puisqu'on utilise l'équation vérifiée par u^* et le fait que $\lambda_i^* = 0$, si $h_i(u^*) < 0$ (sinon $h_i(u^*) = 0$ et ne contribue de toutes façons pas non plus dans la somme). □

Remarque IV.1.13 *On utilisera souvent dans la suite h définie pour tout $u \in \mathbb{R}^d$, par $h(u) = (h_i(u))_{i\in\{1,...,m\}}$, i.e. $h : \mathbb{R}^d \to \mathbb{R}^m, u \mapsto (h_i(u))_{i\in\{1,...,m\}}$.*

IV.1.5 Cas de contraintes d'égalités et d'inégalités

De façon très naturelle, on peut envisager le cas où les deux types de contraintes sont mélangées. On se fixe $p \in \mathbb{N}$ et $m \in \mathbb{N}$ et dans ce cas K donné par

$$K := \left\{ u \in \mathbb{R}^d,\ g_i(u) = 0,\ \forall i \in \{1,...,p\},\ h_i(u) \leq 0,\ \forall i \in \{1,...,m\} \right\}. \quad \text{(IV.29)}$$

On adapte les définitions à ce contexte.

On note toujours $I(u) = \{i \in \{1,...,m\},\ h_i(u) = 0\}$ l'ensemble des contraintes actives en u.
On dispose alors du théorème suivant.

Théorème IV.1.14 (admis) *On suppose que K est donné par des contraintes d'inégalité comme ci-dessus. Soit $u^* \in K$. On suppose que les fonctions $\mathcal{J}$, $(g_i)_{i\in\{1,...,p\}}$ et $(h_i)_{i\in\{1,...,m\}}$ sont dérivables en $u^* \in K$ et que $(\nabla g_i(u^*))_{i\in\{1,...,p\}} \cup (\nabla h_i(u^*))_{i\in I(u^*)}$ est une famille libre. Alors, si u^* est minimum local de $\mathcal{J}$ sur K, il existe μ_1, ..., μ_p et λ_1^*, λ_2^*, ..., $\lambda_m^* \geq 0$, appelés multiplicateurs de Lagrange, tels que :*

$$\nabla\mathcal{J}(u^*) + \sum_{i=1}^{p} \mu_i^* \nabla g_i(u^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(u^*) = 0,$$

$$avec\ \lambda_i^* \geq 0,\ et\ \lambda_i^* = 0,\ si\ h_i(u^*) < 0,\ \forall i \in \{1,...,m\}.$$

Comme dans le cas de contraintes d'inégalité seules, on peut donner un énoncé de ce théorème avec des hypothèses moins restrictives, en donnant la notion de contraintes qualifiées qui suit.

Definition IV.1.15 *On dit que les contraintes définissant K sont qualifiées en $u \in K$ si les vecteurs $(\nabla g_i(u))_{i\in\{1,...,p\}}$ sont linéairement indépendants (i.e. c'est une famille libre) et s'il existe une direction $\tilde{w} \in \cap_{i=1}^{p}(\nabla g_i(u))^{\perp}$ telle que l'on ait pour tout $i \in I(u)$,*

$$\langle \nabla h_i(u), \tilde{w}\rangle < 0. \tag{IV.30}$$

Théorème IV.1.16 (admis) *On suppose que K est donné par des contraintes d'inégalité comme ci-dessus. Soit $u^* \in K$. On suppose que les fonctions $\mathcal{J}$, $(g_i)_{i\in\{1,...,p\}}$ et $(h_i)_{i\in\{1,...,m\}}$ sont dérivables en $u^* \in K$ et que les contraintes sont qualifiées en $u^* \in K$. Alors, si u^* est minimum local de $\mathcal{J}$ sur K, il existe μ_1, ..., μ_p et λ_1^*, λ_2^*, ..., $\lambda_m^* \geq 0$, appelés multiplicateurs de Lagrange, tels que :*

$$\nabla \mathcal{J}(u^*) + \sum_{i=1}^{p} \mu_i^* \nabla g_i(u^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(u^*) = 0,$$

$$\text{avec } \lambda_i^* \geq 0, \text{ et } \lambda_i^* = 0, \text{ si } h_i(u^*) < 0, \ \forall i \in \{1, ..., m\}.$$

IV.1.6 Interprétation en terme de Lagrangien dans tous les cas.

On peut préciser un peu plus notre interprétation en terme de Lagrangien. Si on est dans le cas de contraintes d'égalité, on introduit la fonction $\mathcal{L} : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$, $(u, \lambda) \mapsto \mathcal{J}(u) + \sum_{i=1}^{p} \lambda_i g_i(u) = \mathcal{J}(u) + \langle \lambda, g(u)\rangle$ appelée *Lagrangien*[6] associé au problème de minimisation sous contraintes dans le cas des contraintes d'égalités où K est donné par (IV.3). Si on est dans le cas de contraintes d'inégalités, on introduit la fonction $\mathcal{L} : \mathbb{R}^d \times \mathbb{R}_+^m \to \mathbb{R}$, $(u, \lambda) \mapsto \mathcal{J}(u) + \sum_{i=1}^{m} \lambda_i h_i(u) = \mathcal{J}(u) + \langle \lambda, h(u)\rangle$ appelée *Lagrangien* associé au problème de minimisation sous contraintes dans le cas des contraintes d'inégalités où K est donné par (IV.17). On note $\Lambda = \mathbb{R}^p$ dans le cas de contraintes dégalités et $\Lambda = \mathbb{R}_+^m$ dans le cadre de contraintes d'inégalités.

On définit la notion de point selle de Lagrangien.

Definition IV.1.17 *Le point $(u^*, \lambda^*) \in \mathbb{R}^d \times \Lambda$ est un point selle du Lagrangien $\mathcal{L}$ sur $\mathbb{R}^d \times \Lambda$ si pour tout $u \in \mathbb{R}^d$, pour tout $\lambda \in \Lambda$,*

$$\mathcal{L}(u^*, \lambda) \leq \mathcal{L}(u^*, \lambda^*) \leq \mathcal{L}(u, \lambda^*) \tag{IV.31}$$

On peut montrer que si $(u^*, \lambda^*) \in \mathbb{R}^d \times \Lambda$ est un point selle de $\mathcal{L}$ sur $\mathbb{R}^d \times \Lambda$, alors u^* est un point de minimum de $\mathcal{J}$ sur K.

Pour le théorème suivant (cf. théorème ci-dessous), on adopte la notation suivante :

— Si on est dans le cas de contraintes d'égalités, on pose $k_i = g_i$ pour $i \in \{1, ..., p\}$ et $l = p$.

— Si on est dans le cas de contraintes d'inégalités, on pose $k_i = h_i$ pour $i \in \{1, ..., m\}$ et $l = m$.

6. Ici $g : \mathbb{R}^d \to \mathbb{R}^p, u \mapsto (g_i(u))_{i\in\{1,...,p\}}$.

Théorème IV.1.18 *On suppose que les fonctions $k = (k_i)_{i\in\{1,...,l\}}$, définissant l'espace des contraintes K (cas d'égalité ou d'inégalité) sont toutes continues. On définit $\mathcal{L}$ comme ci-dessus (suivant les types de contraintes). Soit alors $\mathcal{U}$ un ouvert de $\mathbb{R}^d$ contenant K. Soit $(u^*, \lambda^*) \in \mathcal{U} \times \Lambda$ un point selle de $\mathcal{L}$ sur $\mathcal{U} \times \Lambda$, alors u^* est un point de minimum global de $\mathcal{J}$ sur K. De plus, si $\mathcal{J}$ et toutes les fonctions définissant les contraintes $k = (k_i)_{i\in\{1,...,l\}}$ sont dérivables en u^*, alors on a*

$$\nabla\mathcal{J}(u^*) + \sum_{i=1}^{l} \lambda_i^* \nabla k_i(u^*) = 0. \tag{IV.32}$$

Preuve : Pour simplifier l'écriture on note pour tout $(v, \beta) \in \mathcal{U} \times \Lambda$, $\sum_{i=1}^{l} \beta_i \nabla k_i(v)$ par $\beta \cdot \nabla k(v)$.

En écrivant la condition de point selle, on trouve $\forall v \in \mathcal{U}$, $\forall \beta \in \Lambda$,

$$\mathcal{J}(u^*) + \langle \beta, k(u^*)\rangle \leq \mathcal{J}(u^*) + \langle \lambda^*, k(u^*)\rangle \leq \mathcal{J}(v) + \langle \lambda^*, k(v)\rangle. \tag{IV.33}$$

— *Pour le cas des contraintes d'égalité.* On a directement, $\langle(\beta - \lambda^*), k(u^*)\rangle \leq 0$ pour tout $\beta \in \mathbb{R}^p$, en utilisant l'inégalité de gauche. En choisissant $\beta = \lambda^* + \zeta$ puis $\beta = \lambda^* - \zeta$, avec $\zeta \in \mathbb{R}^p$, on trouve $\langle \zeta, k(u^*)\rangle = 0$ pour tout $\zeta \in \mathbb{R}^p$. Et donc $k_i(u^*) = 0$ pour tout $i \in \{1, ..., p\}$ (il suffit de choisir pour $i \in \{1, ..., l\}$, $\zeta = (0, ..., 0 \underbrace{, 1,}_{i\text{ème position}} 0, ..., 0)$. Donc $u^* \in K$. En prenant ensuite la deuxième inégalité et $v \in K$, on déduit que $\mathcal{J}(u^*) \leq \mathcal{J}(v)$, $\forall v \in K$. On a donc le résultat voulu.

— *Pour le cas des contraintes d'inégalité.* On a $\lambda^* \in \mathbb{R}^m_+$. La première inégalité donne $\langle(\beta - \lambda^*), k(u^*)\rangle \leq 0$, pour tout $\beta \in \mathbb{R}^m_+$. En choisissant $\beta = \lambda^* + \zeta$, avec $\zeta \in \mathbb{R}^m_+$, on trouve $\langle \zeta, k(u^*)\rangle \leq 0$, pour tout $\zeta \in \mathbb{R}^m_+$. Donc en particulier, $k_i(u^*) \leq 0$, pour tout $i \in \{1, ..., m\}$ (il suffit de choisir pour $i \in \{1, ..., l\}$, $\zeta = (0, ..., 0 \underbrace{, 1,}_{i\text{ème position}} 0, ..., 0)$.
De plus $\langle \zeta, k(u^*)\rangle \leq 0$ avec $\zeta = \lambda^*$ donne $\langle \lambda^*, k(u^*)\rangle \leq 0$. Mais l'inégalité de gauche de l'encadrement avec $\beta = 0$ donne aussi $\langle \lambda^*, k(u^*)\rangle \geq 0$. Et donc $\langle \lambda^*, k(u^*)\rangle = 0$. Cette condition nous dit que si $i \in \{1, ..., m\}$ est tel que $k_i(u^*) < 0$, alors $\lambda_i^* = 0$.
De plus, en prenant ensuite la deuxième inégalité et $v \in K$, on déduit que $\mathcal{J}(u^*) \leq \mathcal{J}(v)$, $\forall v \in K$ (en utilisant $\langle \lambda^*, k(u^*)\rangle = 0$, $\lambda^* \in \mathbb{R}^m_+$, $k_i(v) \leq 0$, $\forall i \in \{1, ..., l\}$). On a donc le résultat voulu.

On voit que la deuxième inégalité de l'encadrement nous dit que u^* est un point de minimum de $v \mapsto \mathcal{J}(v) + \langle \lambda^*, k(v)\rangle$ sur $\mathcal{U}$ (et donc sans contraintes). On sait donc (comme $\mathcal{U}$ est ouvert) que si toutes les fonctions sont dérivables $\nabla\mathcal{J}(u^*) + \lambda^* \cdot \nabla k(u^*) = 0$. Ce qui termine la preuve du théorème.

□

On est donc en fait passé d'un problème avec contraintes à un problème sans contraintes.

IV.2 Algorithmes numériques pour les problèmes avec contraintes.

Comme dans le cas sans contraintes, on voudrait pouvoir approcher le point de minimum sous contraintes (s'il existe et est unique). Il nous faut donc adapter nos algorithmes. Car si on reprend les mêmes, on n'est pas sûr a priori de rester dans K à chaque itération (toujours cette même problématique...). Pour plus de détails, vous pouvez regarder par exemple [1].

IV.2.1 Théorème de projection sur un convexe en dimension finie

On commence par énoncer le théorème de projection sur un convexe fermé en dimension finie qui joue un rôle important ici. Pour les MPA, vous avez vu ce théorème dans un cadre bien plus général, pas forcément en dimension finie (cf. UE Analyse fonctionnelle et espaces de Hilbert).

Théorème IV.2.1 (Théorème de projection sur un convexe fermé en dimension finie.) *Soit K une partie convexe fermée et non vide de $\mathbb{R}^d$ et $x \in \mathbb{R}^d$. Alors, il existe un unique $x_K \in K$, tel que*

$$\|x - x_K\| = \min_{y \in K} \|x - y\|. \tag{IV.34}$$

De plus, x_K est caractérisé par

$$\langle x - x_K, y - x_K \rangle \leq 0, \, \forall y \in K. \tag{IV.35}$$

On dira alors que x_K est la projection de x sur K et on notera $x_K = p_K(x)$. L'application $p_K : \mathbb{R}^d \to K$ ainsi définie est 1-lipschitzienne.

Preuve : On fera cette preuve comme exercice de TD. □

Remarque IV.2.2 *Si $u \in K$, alors $p_K(u) = u$.*
De plus, si K est convexe fermé et non vide et $\mathcal{J}$ différentiable en u^* et que $u^* \in K$ est un point de minimum local de $\mathcal{J}$ sur K, alors on a $\langle \nabla \mathcal{J}(u^*), v - u^* \rangle \geq 0$, $\forall v \in K$ (cf. théorème IV.1.1). Donc pour tout $v \in K$, pour tout $\rho > 0$,

$$\langle \rho \nabla \mathcal{J}(u^*), v - u^* \rangle \geq 0. \tag{IV.36}$$

D'où pour tout $v \in K$, pour tout $\rho > 0$,

$$\langle u^* - \rho \nabla \mathcal{J}(u^*) - u^*, v - u^* \rangle \leq 0. \tag{IV.37}$$

Par la caractérisation du théorème de projection sur un convexe fermé, on trouve que $p_K(u^* - \rho \nabla \mathcal{J}(u^*)) = u^*$.

IV.2.2 Algorithme de gradient à pas constant avec projection.

On se place ici dans le cas où K, l'espace des contraintes est un ensemble convexe fermé et non vide.

L'algorithme de gradient à pas fixe ($\rho > 0$) avec projection s'écrit :

Algorithme du gradient à pas constant avec projection
— *Initialisation* : $u^0 \in K$ donné et $\rho > 0$ donné.
— *Itération* : $k \in \mathbb{N}$,

$$u_{k+1} = p_K(u_k - \rho \nabla \mathcal{J}(u_k)).$$

On voit qu'en comparaison avec un algorithme de gradient à pas fixe "classique" dans le cas sans contraintes, on a rajouté à chaque itération une étape de projection sur l'espace K.

Théorème IV.2.3 *Soit $\alpha > 0$. On suppose que $\mathcal{J}$ est α-convexe, $\mathcal{C}^1$ et à gradient Lipschitzien sur $\mathbb{R}^d$ de constante de Lipschitz $M > 0$. Alors, si $0 < \rho < \frac{2\alpha}{M^2}$, l'algorithme de gradient à pas fixe avec projection converge, i.e. pour tout $u^0 \in K$, la suite $(u_k)_{k\in\mathbb{N}}$ définie ci-dessus par l'algorithme converge vers la solution u^* du problème de minimisation.*

Preuve : On a tout d'abord existence et unicité du point de minimum $u^* \in K$ car $\mathcal{J}$ est α-convexe (avec $\alpha > 0$) et $K \subset \mathbb{R}^d$ est un fermé non vide. On sait déjà par la remarque IV.2.2 que

$$u^* = p_K(u^* - \rho \nabla \mathcal{J}(u^*)). \tag{IV.38}$$

La preuve ressemble beaucoup à la preuve de la convergence dans le cas sans contraintes. Soit $k \in \mathbb{N}$ fixé. Par la définition de la méthode de gradient à pas constant avec projection, on a $u_{k+1} = p_K(u_k - \rho \nabla \mathcal{J}(u_k))$. Donc

$$\begin{aligned} \|u_{k+1} - u^*\|^2 &= \|p_K(u_k - \rho \nabla \mathcal{J}(u_k)) - \underbrace{u^*}_{=p_K(u^* - \rho \nabla \mathcal{J}(u^*))}\|^2, & \text{(IV.39)} \\ &\leq \|u_k - \rho \nabla \mathcal{J}(u_k) - (u^* - \rho \nabla \mathcal{J}(u^*))\|^2, & \text{(IV.40)} \end{aligned}$$

comme par le théorème IV.2.1, on sait que p_K est 1-Lipschitzienne.

À partir de ce moment là, on est dans la même situation que dans la preuve du gradient à pas constant et sans contraintes.

On obtient en développant l'expression

$$\|u_{k+1} - u^*\|^2 \leq \|u_k - u^*\|^2 - 2\rho \langle \nabla \mathcal{J}(u_k) - \nabla \mathcal{J}(u^*), u_k - u^* \rangle + \rho^2 \|\nabla \mathcal{J}(u_k) - \nabla \mathcal{J}(u^*)\|^2$$

Les hypothèses (α-convexité et III.1) permettent d'écrire (en utilisant une caractérisation de l'α-convexité de la proposition II.4.2) alors

$$\|u_{k+1} - u^*\|^2 \leq (1 - 2\rho\alpha + \rho^2 M^2) \|u_k - u^*\|^2.$$

La fin de la preuve est alors la même que le théorème III.2.3, on ne la re-détaille pas ici. Le résultat en découle. □

Cet algorithme a l'air assez simple à mettre en œuvre, mais il ne faut pas oublier qu'à chaque itération il y a une projection à faire. Et en fait ce n'est pas si simple d'avoir accès à l'expression de la projection : il y a une différence entre savoir que la projection existe et en donner une expression explicite ! Ce n'est pas si facile. Un cas où on sait le faire de façon élémentaire est le cas où K est un pavé (un produit cartésien de segments), on verra cela en TD.

Le problème des tests d'arrêts se pose là aussi (cf. TP).

IV.2.3 Algorithme d'Uzawa

On se place dans le cadre de contraintes d'inégalité. On cherche ici à trouver directement un point selle du Lagrangien. On définit donc

$$\mathcal{L} : \mathbb{R}^d \times \mathbb{R}^m_+ \to \mathbb{R},\ (u, \lambda) \mapsto \mathcal{J}(u) + \sum_{i=1}^{m} \lambda_i h_i(u) = \mathcal{J}(u) + \langle \lambda, h(u) \rangle.$$

En exploitant la définition d'un point selle IV.1.17, on obtient alors avec la première inégalité, si $(u^*, \lambda^*) \in \mathbb{R}^d \times \Lambda$ est un point selle de $\mathcal{L}$, alors pour tout $\beta \in \mathbb{R}^m_+$,

$$\mathcal{L}(u^*, \beta) \leq \mathcal{L}(u^*, \lambda^*) \tag{IV.41}$$

Donc pour tout $\beta = (\beta_1, ..., \beta_m) \in \mathbb{R}^m_+$,

$$\sum_{i=1}^{m} (\beta_i - \lambda_i^*) h_i(u^*) \leq 0. \tag{IV.42}$$

Cela donne pour tout $\mu > 0$, $\sum_{i=1}^m (\beta_i - \lambda_i^*)(\mu h_i(u^*) - \lambda_i^* + \lambda_i^*) \leq 0$.
Autrement dit pour tout $\beta \in \mathbb{R}^m_+$ et pour tout $\mu > 0$

$$\langle \beta - \lambda^*, (\mu h(u^*) + \lambda^*) - \lambda^* \rangle \leq 0. \tag{IV.43}$$

On utilise alors de nouveau la caractérisation de la projection sur un convexe fermé IV.2.1 sur $\mathbb{R}^m_+$ (qui est bien un convexe fermé non vide), et on déduit que pour tout $\mu > 0$,

$$\lambda^* = p_{\mathbb{R}^m_+}(\mu h(u^*) + \lambda^*). \tag{IV.44}$$

En s'inspirant de cela, l'algorithme est alors donné par

Algorithme d'Uzawa

Initialisation : $\mu > 0$ et $\lambda_0 \in \mathbb{R}^m_+$ donnés.
Itération : $k \in \mathbb{N}$.

— u_k est calculé comme solution de

$$\mathcal{L}(u_k, \lambda_k) = \min_{u \in \mathbb{R}^d} \mathcal{L}(u, \lambda_k)$$

(c'est un problème d'optimisation sans contraintes).

— $\lambda_{k+1} = p_{\mathbb{R}^m_+}(\lambda_k + \mu h(u_k))$.

On a là aussi un théorème de convergence de l'algorithme.

Théorème IV.2.4 (admis) *On suppose que $\mathcal{J} : \mathbb{R}^d \to \mathbb{R}$ est α-convexe et $\mathcal{C}^1$, que $h = (h_i)_{i \in \{1,...m\}}$ est une fonction convexe (au sens d'une fonction de $\mathbb{R}^d$ à valeurs dans $\mathbb{R}^m$). On suppose que h est lipschitzienne*[7]*, i.e. qu'il existe $C > 0$ tel que $\|h(w) - h(w)\| \leq C\|v - w\|$, $\forall (v, w) \in \mathbb{R}^d \times \mathbb{R}^d$. On suppose également qu'il existe un point selle (u^*, λ^*) du Lagrangien sur $\mathbb{R}^d \times \Lambda$. Alors, si $0 < \mu < \frac{2\alpha}{C^2}$, l'algorithme d'Uzawa converge, i.e. quelque soit l'élément initial λ^0, la suite $(u_k)_{k \in \mathbb{N}}$ définie par l'algorithme d'Uzawa correspondant converge vers la solution u^* du problème de minimisation sous contraintes.*

7. Attention ici à bien adapter les normes, suivant les vecteurs auxquels elle s'applique, $\|\cdot\|$ est la norme euclidienne sur $\mathbb{R}^d$ ou sur $\mathbb{R}^m$.

IV.2.4 Méthode de pénalisation

On donne ici encore une autre méthode qui permet de se ramener à un problème de minimisation sans contraintes.

Considérons un problème de minimisation avec contraintes d'inégalités. On introduit une fonction $\varphi : \mathbb{R}^d \to \mathbb{R}$, continue et telle que, pour tout $v \in \mathbb{R}^d$,

$$\varphi(v) \geq 0, \tag{IV.45}$$

et

$$\varphi(v) = 0 \text{ si et seulement si } v \in K. \tag{IV.46}$$

La méthode de pénalisation consiste à minimiser la fonctionnelle :

$$\mathcal{J}_\eta : v \mapsto \mathcal{J}(v) + \eta\varphi(v),$$

on cherche donc une solution à

$$\min_{u\in\mathbb{R}^d} \mathcal{J}_\eta(u)$$

Pour φ on peut par exemple penser à prendre :

$$\varphi : v \mapsto \sum_{i=1}^{m} (\max(h_i(v), 0))^2.$$

On notera alors $\tilde{\mathcal{J}}_\eta$ la fonction $\mathcal{J}_\eta$ avec cette fonction φ particulière. On a là aussi un théorème de convergence, mais dans le cas strictement convexe.

Théorème IV.2.5 *On suppose que $\mathcal{J}$ est continue, strictement convexe et infinie à l'infini, que les fonction h_i sont convexes et continues pour tout $i \in \{1, ..., m\}$ et que l'ensemble*

$$K = \Big\{ v \in \mathbb{R}^d,\ h_i(v) \leq 0,\ \forall i \in \{1, ..., m\} \Big\}$$

est non vide.

Si u^* est solution du problème de minimisation de $\mathcal{J}$ sous contraintes données par l'ensemble K, alors u_ε solution de

$$\mathcal{J}(u_\varepsilon) = \min_{v\in\mathbb{R}^d} \tilde{\mathcal{J}}_{\frac{1}{\varepsilon}}(v)$$

est telle que

$$\lim_{\varepsilon\to 0} u_\varepsilon = u^*$$

Deuxième partie

Approximation par éléments finis.

Chapitre V

Forme variationnelle et principe général éléments finis

La méthode des éléments finis regroupe une classe de méthodes numériques permettant l'approximation d'équations aux dérivées partielles en se basant sur une formulation particulière de ces dernières, appelée formulation variationnelle. La notion de passage d'un problème en dimension infinie à un problème en dimension finie est aussi essentielle. Commençons par mettre en avant un lien possible entre problème d'optimisation et formulation variationnelle.

V.1 Un problème de minimisation en dimension infinie et lien avec la résolution d'équations différentielles

V.1.1 Problème physique et modélisation

On considère une corde (de longueur 1) tendue attachée à ses deux extrémités et initialement en position horizontale à laquelle on suspend une charge. Les abscisses sont décrites par le point $x \in [0,1]$ et le déplacement vertical de la corde par une fonction $u : x \mapsto u(x)$ par rapport au repos lorsque la charge y est suspendue. On suppose que la corde est fixée aux deux extrémités, ce qui se traduit par les conditions $u(0) = u(1) = 0$. Le champ de forces exercé par le poids sur la corde est modélisé par une densité linéique $f : [0,1] \to \mathbb{R}$.
Une illustration se trouve en Figure V.1 (haut).

Pour décrire le modèle, on fait un bilan d'énergie sur un élément de longueur dx, en faisant l'hypothèse de petits déplacements. Une illustration se trouve en Figure V.1 (bas).

Pour cela on procède en deux étapes :

(a) *L'énergie infinitésimale due à l'étirement de l'élément de corde* est proportionnelle à l'augmentation de la longueur entre la situation au repos et la corde en tension :

$$\xi_1 = k(x)(\sqrt{(u(x+dx)-u(x))^2 + dx^2} - dx) \approx k(x)\left(\sqrt{\left(\frac{\partial u}{\partial x}(x)\right)^2 + 1} - 1\right) dx, \qquad \text{(V.1)}$$

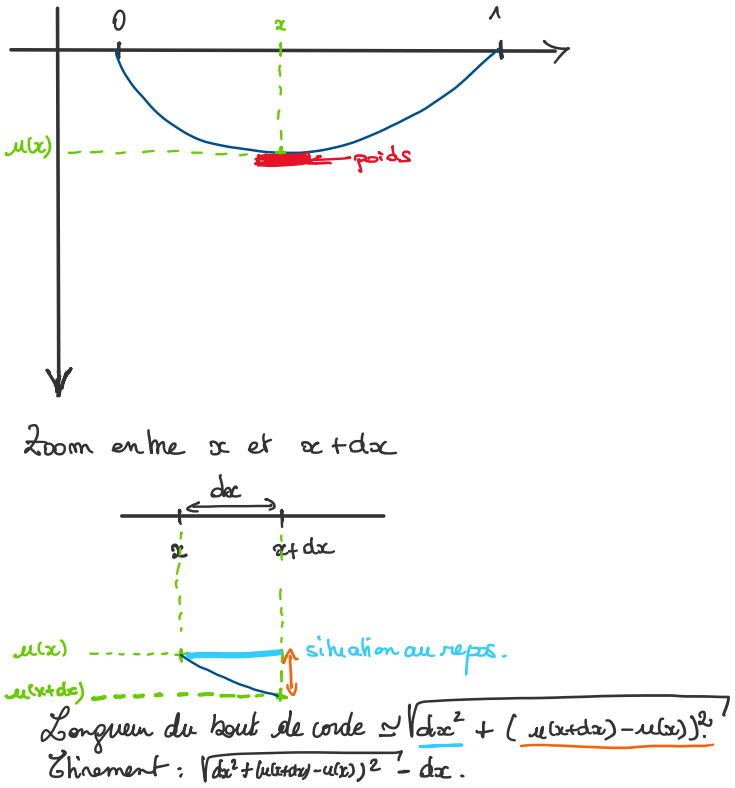

FIGURE V.1 – Illustration du problème de la corde (haut), bilan sur un élément de longueur dx (bas)

où le coefficient de proportionnalité $k(x)$ est donné et s'appelle la raideur de la corde au point x.

(b) *L'énergie potentielle infinitésimale due au poids suspendu sur la corde* est l'opposé du travail de ce poids au cours du déplacement décrit par u et vaut :

$$\xi_2 = -f(x)u(x)dx. \tag{V.2}$$

L'énergie globale $\mathcal{E}(u)$ est donnée par

$$\int_0^1 k(x)\left(\sqrt{1+\left(\frac{\partial u}{\partial x}(x)\right)^2}-1\right)dx - \int_0^1 f(x)u(x)dx. \tag{V.3}$$

En supposant que le déplacement u et sa dérivée sont petits, on peut supposer que

$$\sqrt{1+\left(\frac{\partial u}{\partial x}(x)\right)^2} \approx \frac{1}{2}\left(\frac{\partial u}{\partial x}(x)\right)^2. \tag{V.4}$$

On a utilisé ici le développement limité en 0 : $\sqrt{1+y} = 1 + \frac{1}{2}y + o(y)$ avec $y = \left(\frac{\partial u}{\partial x}(x)\right)^2$ et on a négligé le terme en o.

On obtient alors une expression simplifiée pour $\mathcal{E}(u)$:

$$\mathcal{E}(u) = \frac{1}{2}\int_0^1 k(x)\left(\frac{\partial u}{\partial x}(x)\right)^2 dx - \int_0^1 f(x)u(x)dx. \tag{V.5}$$

V.1.2 Un problème d'optimisation

Physiquement le système cherche à minimiser son énergie. *La position de la corde à l'équilibre est l'unique u qui minimise l'énergie $\mathcal{E}(u)$ et qui vérifie les conditions aux limites $u(0) = u(1) = 0$.*

Essayons de formaliser un peu le problème d'optimisation. L'espace V (où l'on cherche la variable d'optimisation) que l'on considèrerait pourrait être a priori $W := \{v \in \mathcal{C}^1([0,1]),\ \text{tel que } v(0) = v(1) = 0\}$, la fonction coût serait $\mathcal{E} : W \to \mathbb{R}, v \mapsto \mathcal{E}(v)$, et on peut donc écrire le problème d'optimisation comme : *Trouver $u^* \in W$ tel que*

$$\mathcal{E}(u^*) = \min_{u \in W} \mathcal{E}(v). \tag{V.6}$$

On voit que c'est un problème d'optimisation non linéaire en dimension **infinie** et sans contraintes.

La théorie du cours ne permet pas de traiter ce problème d'optimisation, puisque nous n'avons vu que des résultats en dimension finie et ici l'ensemble auquel appartient u n'est pas de dimension finie.

La façon de généraliser les résultats vus en cours repose sur la notion d'espaces de Hilbert [1] qui permettent de retrouver un cadre adapté.

On peut ainsi montrer la proposition suivante que l'on admettra :

Proposition. Soient $f \in \mathcal{C}^0([0,1],\mathbb{R})$, $k \in \mathcal{C}^1([0,1],\mathbb{R})$ telle que $\inf_{[0,1]} k > 0$, alors il existe un unique $u \in V$ tel que pour tout $v \in V, \mathcal{E}(u) \leq \mathcal{E}(v)$, avec $V := \{v \in \mathcal{C}^2([0,1]),\ v(0) = v(1) = 0\}$.

Dans la suite, on se place dans le cadre simplifié où $k \equiv 1$.

Essayons dans ce cas d'écrire au moins formellement (on ne justifie rien !) une condition nécessaire d'optimalité.

Si u est un point de minimum local de $\mathcal{E}$ sur V, alors pour $v \in V$ et $t \in \mathbb{R}$ suffisamment petit, $u + tv \in V$ (V est un espace vectoriel) et

$$\mathcal{E}(u + tv) - \mathcal{E}(u) \geq 0 \tag{V.7}$$

Ré-écrivons $\mathcal{E}$ (on enlève ici la dérivée partielle puisque u est une fonction définie sur un intervalle de $\mathbb{R}$) :

$$\mathcal{E}(u) = \frac{1}{2}\int_0^1 (u'(s))^2 ds - \int_0^1 f(s)u(s)ds. \tag{V.8}$$

1. Pour les MPA...

avec $u \in V := \left\{ v \in \mathcal{C}^2([0,1]),\ \text{telle que } v(0) = v(1) = 0 \right\}$.
On écrit pour tout $v \in V$ et $t \in \mathbb{R}$:

$$\begin{aligned} \mathcal{E}(u+tv) &= \frac{1}{2}\int_0^1 ((u+tv)'(s))^2 ds - \int_0^1 f(s)(u(s)+tv(s))ds, & \text{(V.9)} \\ &= \mathcal{E}(u) + t(\int_0^1 u'(s)v'(s)ds - \int_0^1 f(s)v(s)) + \frac{1}{2}t^2\int_0^1 (v'(s))^2 ds. & \text{(V.10)} \end{aligned}$$

Ce qui donne pour $t \neq 0$,

$$\frac{\mathcal{E}(u+tv)-\mathcal{E}(u)}{t} = \int_0^1 u'(s)v'(s)ds - \int_0^1 f(s)v(s) + \frac{1}{2}t\int_0^1 (v'(s))^2 ds. \quad \text{(V.11)}$$

Comme dans le cas de la dimension finie, on divise alors (V.22) par $t > 0$ et on fait tendre t vers 0^+. On déduit que

$$\int_0^1 u'(s)v'(s)ds - \int_0^1 f(s)v(s) \geq 0.$$

Ensuite on divise alors (V.22) par $t < 0$ et on fait tendre t vers 0^-. On déduit que

$$\int_0^1 u'(s)v'(s)ds - \int_0^1 f(s)v(s) \leq 0.$$

Finalement

$$\int_0^1 u'(s)v'(s)ds - \int_0^1 f(s)v(s) = 0. \quad \text{(V.12)}$$

On obtient donc, pour tout $v \in V$,

$$\int_0^1 u'(s)v'(s)ds = \int_0^1 f(s)v(s)ds. \quad \text{(V.13)}$$

C'est la condition nécessaire d'optimalité.

Vu la régularité des fonctions de V, on peut se permettre de faire une intégration par partie dans le membre de gauche. Cela donne pour tout $v \in V$,

$$-\int_0^1 u''(s)v(s) = \int_0^1 f(s)v(s)ds.$$

On peut également montrer [2] que sous les hypothèses de la proposition précédente, cette égalité donne que $u \in V$ est solution de

$$\begin{aligned} -u''(x) &= f(x),\ \forall x \in]0,1[, & \text{(V.14)} \\ u(0) &= u(1) = 0. & \text{(V.15)} \end{aligned}$$

2. Pour les IM, vous pouvez admettre ce résultat. Pour les MPA : il faut appliquer cette égalité à des v qui sont $\mathcal{C}_c^\infty$ puis passer par densité à une égalité valable pour des v dans $L^2([0,1])$. Puis on choisit $v = -u'' - f$ et on obtient $\int_0^1 (-u''(s) - f(s))^2 ds = 0$ et donc le résultat ci-dessous

Cette équation est une équation différentielle avec **conditions aux bords** (ou **conditions aux limites**). Vous avez vu cette équation dans le cours d'EDP différences finies au premier semestre. On l'appelle communément **l'équation de Poisson** (ici en dimension 1)

On dit que (V.28) est la *formulation variationnelle* de (V.14)-(V.15).

V.1.3 Formalisation générale

On peut reécrire la formulation variationnelle sous une forme standard classique : Trouver $v \in \mathcal{V}$ tel que

$$a(u,v) = l(v),\ \forall v \in \mathcal{V}, \tag{V.16}$$

avec $\mathcal{V}$ l'espace dans lequel on cherche une solution, $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}, (u,v) \mapsto \int_0^1 u'(s)v'(s)ds$ et $l : \mathcal{V}, v \mapsto \int_0^1 f(s)v(s)ds$. On verra que c'est le formalisme usuel.

On a montré que sous de bonnes hypothèses, on peut obtenir (V.14)-(V.15) à partir de (V.28).

Mais on peut aussi montrer que l'on peut obtenir (V.28) à partir de (V.14)-(V.15). Il suffit de multiplier par une fonction $v \in \mathcal{V}$, d'intégrer sur $[0,1]$ et de faire une intégration par parties pour obtenir la formulation variationnelle. En effet, prenons (V.14)-(V.15), et multiplions par $v \in V$, on obtient :

$$-u''(x)v(x) = f(x)v(x),\ \forall x \in]0,1[,. \tag{V.17}$$

(V.18)

Puis intégrons sur $[0,1]$, puis intégrons par partie le membre de gauche, on obtient :

$$-\int_0^1 u''(s)v(s)ds = \int_0^1 f(s)v(s)ds, \tag{V.19}$$

$$\int_0^1 u'(s)v'(s)ds + \big[v(s)u'(s)\big]_0^1 = \int_0^1 f(s)v(s)ds, \tag{V.20}$$

Mais comme $v \in V$, on a $v(0) = v(1) = 0$, donc

$$\int_0^1 u'(s)v'(s)ds = \int_0^1 f(s)v(s)ds, \tag{V.21}$$

On peut aussi donner un sens à la formulation variationnelle même sans avoir $v \in \mathcal{C}^2([0,1])$ (et c'est un aspect essentiel de la formulation variationnelle). Par exemple, si v est $\mathcal{C}^1([0,1])$ (ou $\mathcal{C}^1_m([0,1])$), alors la formulation a un sens. En réalité, on aura plutôt recours à un *espace de Sobolev* $H^1_0([0,1])$ (admis pour les IM).

En fait, le cadre approprié pour travailler avec les formulations variationnelles est le cadre des espaces de Hilbert qui sont des espaces vectoriels munis d'un produit scalaire et dont la norme associée

rend l'espace complet (voir paragraphe de rappels et notations ci-dessous).

Il a été vu en cours d'EDP Différences finies au premier semestre que l'on peut construire une méthode numérique pour approcher directement la solution de l'équation de Poisson (par différences finies). L'idée de *la méthode des éléments finis* est de plutôt travailler directement avec la formulation variationnelle des équations et de se ramener à un problème en dimension finie. Cette stratégie va permettre d'envisager ensuite des contextes associés à des équations aux dérivées partielles bien plus généraux.

V.2 Formalisme général

Pour nous permettre de ne pas nous réduire à ne considérer que l'équation de Poisson, on va développer un cadre mathématique plus général sur lequel on va développer la stratégie de la méthode des éléments finis. Ce cadre s'appliquera en particulier à l'équation de Poisson, mais pas que !

On commence par quelques rappels et notations. Si ce n'est pas des rappels, vous pouvez admettre les résultats qui suivent.

V.2.1 Quelques notations et rappels

Soit $(V, \|\cdot\|)$ un $\mathbb{R}$-espace vectoriel normé.

Formes linéaires

On donne la définition de ce qu'est une *forme linéaire*.

Definition V.2.1 *On dit que $l : V \to \mathbb{R}$ est **une forme linéaire** si $\forall (u, \tilde{u}) \in V \times V$, $\forall \lambda \in \mathbb{R}$,*
— $l(\lambda u + \tilde{u}) = \lambda l(u) + l(\tilde{u})$.

Formes bilinéaires

On donne la définition de ce qu'est une *forme bilinéaire*.

Definition V.2.2 *On dit que $a : V \times V \to \mathbb{R}$ est **une forme bilinéaire** (à valeurs réelles) si $\forall (u, v) \in V \times V$, $\forall (\tilde{u}, \tilde{v}) \in V \times V$, $\forall \lambda \in \mathbb{R}$,*
— $a(\lambda u + \tilde{u}, v) = \lambda a(u, v) + a(\tilde{u}, v)$,
— $a(u, \lambda v + \tilde{v}) = \lambda a(u, v) + a(u, \tilde{v})$.
*On dit que la forme bilinéaire est **symétrique**, si :*

$$a(u, v) = a(v, u),\ \forall (u, v) \in V \times V.$$

*On dit que la forme bilinéaire est **définie positive** si $\forall v \in V$,*
— $a(v, v) \geq 0$,
et
— $a(v, v) = 0 \Rightarrow v = 0$.

Produit scalaire

Dans la partie Optimisation de ce cours, on a très souvent utilisé le produit scalaire euclidien sur $\mathbb{R}^d$. On peut plus généralement définir ce qu'est un produit scalaire sur un espace vectoriel.

Definition V.2.3 *Un produit scalaire sur un espace vectoriel V est une forme bilinéaire symétrique définie positive sur $V \times V$.*

Exemples :

- $\langle \cdot, \cdot \rangle : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $(x, y) \mapsto \sum_{i=1}^{d} x_i y_i$ est un produit scalaire sur le $\mathbb{R}$-espace vectoriel $\mathbb{R}^d$. C'est le produit scalaire euclidien sur $\mathbb{R}^d$ qu'on a très souvent rencontré dans ce cours...
- Soit Ω un ouvert de $\mathbb{R}^d$, la forme bilinéaire $\langle \cdot, \cdot \rangle : L^2(\Omega) \times L^2(\Omega), (f, g) \mapsto \int_\Omega f(x)g(x)dx$, est un produit scalaire sur l'espace $L^2(\Omega)$.

L'inégalité de Cauchy-Schwarz est une inégalité centrale.

Proposition V.2.4 (Inégalité de Cauchy-Schwarz) *Si a est un produit scalaire, on a pour tout $(u, v) \in V \times V$, $|a(u, v)| \leq a(u, u)^{\frac{1}{2}} a(v, v)^{\frac{1}{2}}$.*

Proposition V.2.5 *Si a est un produit scalaire, alors l'application $u \mapsto a(u, u)^{\frac{1}{2}}$ est une norme de V. On dit que c'est la norme canoniquement associée au produit scalaire.*

Espaces complets

Definition V.2.6 *On dit que $(u_n)_{n\in\mathbb{N}}$ est une suite de Cauchy de V si $\forall \varepsilon > 0$, $\exists N \in \mathbb{N}$ tel que $\forall (p, q) \in \mathbb{N} \times \mathbb{N}$, tels que $p \geq N$ et $q \geq N$, on ait $\|u_p - u_q\| \leq \varepsilon$.*

Definition V.2.7 *Un espace vectoriel normé est complet si toute suite de Cauchy converge.*

Definition V.2.8 *Soit $(H, \langle \cdot, \cdot \rangle)$ un espace vectoriel muni d'un produit scalaire. C'est un Hilbert s'il est complet pour la norme canoniquement associée au produit scalaire.*

V.2.2 Formulation variationnelle

On se donne $\mathcal{V}$ un $\mathbb{R}$-espace vectoriel muni d'un produit scalaire le rendant complet : c'est un espace de Hilbert. On notera $(\mathcal{V}, \langle \cdot, \cdot \rangle)$ cet espace de Hilbert ($\langle \cdot, \cdot \rangle$ le produit scalaire et $\|\cdot\|$ la norme canoniquement associée). On se donne *une forme bilinéaire* $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}, (u, v) \mapsto a(u, v)$ et $l : \mathcal{V} \to \mathbb{R}, v \mapsto l(v)$ *une forme linéaire* sur $\mathcal{V}$.

On s'intéresse à la résolution du problème $(\mathcal{F})$: *Trouver $u \in \mathcal{V}$ tel que*

$$a(u, v) = l(v), \ \forall v \in \mathcal{V}.$$

Cette forme est appelée *Formulation variationnelle.*

Remarque V.2.9 *Le problème présenté en section précédente rentre dans ce cadre en utilisant* $H^1([0,1]) = \left\{v \in L^2([0,1]),\, v' \in L^2([0,1])\right\}$ *(espace connu pour les MPA, admis pour les IM), et*

$$\mathcal{V} := \left\{v \in H^1([0,1]),\, v(0) = v(1) = 0\right\}.$$

On a les résultats et définition suivants :

Proposition V.2.10 *La forme bilinéaire* $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ *est continue si et seulement si il existe* $M > 0$ *tel que pour tout* $(u, v) \in \mathcal{V} \times \mathcal{V}$, $|a(u, v)| \leq M\|u\|\|v\|$.

Definition V.2.11 *On dit que* $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ *est coercive sur* $\mathcal{V} \times \mathcal{V}$ *si il existe* $\alpha > 0$ *tel que pour tout* $v \in \mathcal{V}$, $a(v, v) \geq \alpha\|v\|^2$.

On peut relier la résolution du problème donné par la formulation variationnelle à la recherche d'un minimum d'une fonctionnelle **dans le cas où** a **est symétrique**.

Proposition V.2.12 *[admis] Soit* $(\mathcal{V}, \langle\cdot,\cdot\rangle)$ *un espace de Hilbert et* $a : \mathcal{V}\times\mathcal{V} \to \mathbb{R}$ *une forme bilinéaire symétrique continue et coercive et* $l : \mathcal{V} \to \mathbb{R}$ *une forme linéaire continue. Alors résoudre le problème variationnel* $(\mathcal{F})$ *revient à minimiser la fonction coût* $\mathcal{J} : \mathcal{V} \to \mathbb{R}, v \mapsto \frac{1}{2}a(v, v) - l(v)$ *qui admet un unique minimum sur* $\mathcal{V}$.

Idée de la preuve.
La preuve repose sur le fait que si a est coercive, alors $\mathcal{J}$ est α-convexe. Et le résultat d'existence du minimum, connu en dimension finie pour un fermé non vide, se généralise en dimension infinie à un convexe fermé non vide. Cela assure l'existence du minimum. L'unicité est assurée par l'α-convexité avec $\alpha > 0$, donc la stricte convexité. La caractérisation par $(\mathcal{F})$ se fait à l'aide de la stratégie déjà évoquée : on écrit et développe $\mathcal{J}(u + tv) - \mathcal{J}(u)$ puis on fait tendre t vers 0 le tout en utilisant la symétrie de a. Voyons l'idée formellement : Si u est un point de minimum local de $\mathcal{J}$ sur $\mathcal{V}$, alors pour $v \in \mathcal{V}$ et $t \in \mathbb{R}$ suffisamment petit, $u + tv \in \mathcal{V}$ ($\mathcal{V}$ est un espace vectoriel) et

$$\mathcal{J}(u + tv) - \mathcal{J}(u) \geq 0 \tag{V.22}$$

On écrit pour tout $v \in V$ et $t \in \mathbb{R}$:

$$\begin{aligned} \mathcal{J}(u + tv) &= \frac{1}{2}a(u + tv, u + tv) - l(u + tv), &\text{(V.23)}\\ &= \frac{1}{2}a(u, u) - l(u) + t\frac{1}{2}(a(u, v) + a(v, u)) - tl(v) + \frac{1}{2}t^2a(v, v). &\text{(V.24)}\end{aligned}$$

En utilisant la symétrie de a et l'expression de $\mathcal{J}$, on trouve

$$\mathcal{J}(u + tv) = \mathcal{J}(u) + t(a(u, v) - l(v)) + \frac{1}{2}t^2a(v, v). \tag{V.25}$$

Ce qui donne

$$\frac{\mathcal{J}(u + tv) - \mathcal{J}(u)}{t} = a(u, v) - l(v) + \frac{1}{2}ta(v, v). \tag{V.26}$$

On divise alors (V.22) par $t > 0$ et on fait tendre t vers 0^+. On déduit que

$$a(u, v) - l(v) \geq 0.$$

Ensuite on divise alors (V.22) par $t < 0$ et on fait tendre t vers 0^-. On déduit que

$$a(u, v) - l(v) \leq 0.$$

Finalement

$$a(u, v) - l(v) = 0. \tag{V.27}$$

On obtient donc, pour tout $v \in \mathcal{V}$,

$$a(u, v) = l(v). \tag{V.28}$$

□

Remarque V.2.13 *On remarque que cette écriture contient en particulier le cas de la dimension finie. De plus $\mathcal{J}$ est une fonctionnelle quadratique !*

En fait, en utilisant la théorie Hilbertienne, on peut montrer qu'il y a encore existence et unicité de la solution au problème, même si on enlève l'hypothèse de symétrie.

C'est **le théorème de Lax Milgram**.

Théorème V.2.14 (Lax Milgram, admis) *Soit $(\mathcal{V}, \langle \cdot, \cdot \rangle)$ un espace de Hilbert réel (on note $\| \cdot \|$) la norme associée, l une forme linéaire continue sur $\mathcal{V}$ et a une forme bilinéaire continue et coercive sur $\mathcal{V} \times \mathcal{V}$. Alors le problème variationnel $(\mathcal{F})$ admet une unique solution.*

Remarque V.2.15 *Par contre, dans le cas où a n'est pas symétrique, on ne peut plus caractériser la solution comme le minimum de la fonctionnelle $\mathcal{J}$.*

Pour l'exemple de la section précédente, on peut montrer que a est bilinéaire symétrique continue et coercive sur un espace $\tilde{\mathcal{V}}$, différent de $\mathcal{V}$ ($\mathcal{V} \subset \tilde{\mathcal{V}}$) et l continue sur ce même espace $\tilde{\mathcal{V}}$. Cet espace est $H_0^1([0, 1])$ (admis pour les IM).

V.3 Principe de la méthode des éléments finis, passage en dimension finie.

La méthode des éléments finis se base sur la formulation variationnelle associée à une équation aux dérivées partielles (dans le même esprit que ce que l'on a vu avec l'exemple de la corde). Le but est d'*approcher directement la solution du problème variationnel plutôt que d'approcher directement l'équation aux dérivées partielles*. Pour cela, on va chercher à approcher la solution $u \in \mathcal{V}$ dans un espace de **dimension finie**.

Le point de départ de la méthode est donc une formulation variationnelle de type $(\mathcal{F})$ que l'on va utiliser dans un contexte particulier.

V.3.1 Formulation variationnelle discrète

On se donne un paramètre $h > 0$ (dont l'interprétation sera précisée plus tard) et on introduit un sous ensemble $\mathcal{V}_h \subset \mathcal{V}$ de dimension finie $N_h \in \mathbb{N}^*$ et on restreint la formulation variationnelle à $\mathcal{V}_h$, i.e. on transforme le problème $(\mathcal{F})$ du paragraphe précédent (qui n'est pas en dimension finie a priori) en un problème $(\mathcal{F}_h)$ posé en dimension finie comme suit :

$(\mathcal{F}_h)$ *Trouver $u_h \in \mathcal{V}_h$, tel que pour tout $v_h \in \mathcal{V}_h$,*

$$a(u_h, v_h) = l(v_h).$$

On appelle $(\mathcal{F}_h)$ **la formulation variationnelle discrète** de $(\mathcal{F})$ associée à l'espace $\mathcal{V}_h$.

On peut montrer que sous les mêmes hypothèses que dans le paragraphe précédent, on a existence et unicité d'une solution u_h dans $\mathcal{V}_h$.

Proposition V.3.1 *Soit $\mathcal{V}$ un espace de Hilbert réel et $\mathcal{V}_h$ un sous-espace de dimension finie de $\mathcal{V}$. Soit $a : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ une forme bilinéaire continue et coercive sur $\mathcal{V} \times \mathcal{V}$ et $l : \mathcal{V} \to \mathbb{R}$ une forme linéaire continue sur $\mathcal{V}$.*
Alors le problème $(\mathcal{F}_h)$ a une solution unique et est équivalent à la résolution d'un système linéaire dont la matrice est inversible. Si de plus a est symétrique, alors la matrice du système linéaire à résoudre est symétrique définie positive.

Preuve : On se donne une base de $\mathcal{V}_h$ que l'on note $\mathcal{B}_h := (\varphi_i)_{i\in\{1,...,N_h\}}$. On sait qu'on peut alors décomposer tout élément de $\mathcal{V}_h$ sur cette base. Donc il existe $(u_h^i)_{i\in\{1,...,N_h\}}$, N_h valeurs réelles telles que

$$u_h = \sum_{i=1}^{N_h} u_h^i \varphi_i.$$

Notons $U_h = \begin{pmatrix} u_h^1 \\ u_h^2 \\ \vdots \\ u_h^{N_h} \end{pmatrix} \in \mathbb{R}^{N_h}$ le vecteur de coordonnées de u_h dans la base $(\varphi_i)_{i\in\{1,...,N_h\}}$.

Montrons que le problème $(\mathcal{F}_h)$ est équivalent au problème $(\mathcal{F}_h^{bis})$ suivant :

$(\mathcal{F}_h^{bis})$ *: Trouver $u_h \in \mathcal{V}_h$ tel que*

$$a(u_h, \varphi_i) = l(\varphi_i),\ \forall i \in \{1, ..., N_h\}. \tag{V.29}$$

En effet, $(\mathcal{F}_h) \Rightarrow (\mathcal{F}_h^{bis})$, en prenant $v_h = \varphi_i \in \mathcal{V}_h$ dans $(\mathcal{F}_h)$.

Et $(\mathcal{F}_h^{bis}) \Rightarrow (\mathcal{F}_h)$. Soit u_h solution de $(\mathcal{F}_h^{bis})$. Soit $v_h \in \mathcal{V}_h$ et $(v_h^i)_{i\in\{1,...,N_h\}}$ ses coordonnées dans la base $(\varphi_i)_{i\in\{1,...,N_h\}}$. On a donc comme u_h solution de $(\mathcal{F}_h^{bis})$, en multipliant l'équation correspondante (V.29) par v_h^i

$$v_h^i a(u_h, \varphi_i) = v_h^i l(\varphi_i),\ \forall i \in \{1, ..., N_h\}. \tag{V.30}$$

En utilisant la bilinéarité de a et la linéarité de l, on déduit que

$$a(u_h, v_h^i \varphi_i) = l(v_h^i \varphi_i),\ \forall i \in \{1, ..., N_h\}. \tag{V.31}$$

Puis en sommant sur $i \in \{1, ..., N_h\}$ et en utilisant là encore la bilinéarité de a et la linéarité de l, on a

$$a(u, \underbrace{\sum_{i=1}^{N_h} v_h^i \varphi_i}_{=v_h}) = l(\underbrace{\sum_{i=1}^{N_h} v_h^i \varphi_i}_{=v_h}),\ \forall i \in \{1, ..., N_h\}. \tag{V.32}$$

Ce qui donne que u_h est solution de $(\mathcal{F}_h)$.

En utilisant la décomposition de u_h sur la base $\mathcal{B}_h$, on a donc en particulier que le problème $(\mathcal{F}_h^{bis})$ est équivalent à trouver $(u_h^j)_{j \in \{1,...,N_h\}}$ tel que

$$a(\sum_{j=1}^{N_h} u_h^j \varphi_j, \varphi_i) = l(\varphi_i),\ \forall i \in \{1, ..., N_h\}, \tag{V.33}$$

$$\sum_{j=1}^{N_h} u_h^j a(\varphi_j, \varphi_i) = l(\varphi_i),\ \forall i \in \{1, ..., N_h\}, \tag{V.34}$$

$$\tag{V.35}$$

par bilinéarité de a.

On peut réecrire cette dernière inégalité sous la forme d'un système linéaire

$$\mathcal{A}_h U_h = L_h, \tag{V.36}$$

avec $\mathcal{A}_h = (a(\varphi_j, \varphi_i))_{(i,j) \in \{1,...,N_h\}^2}$ qui est une matrice carrée et $L_h = (l(\varphi_i))_{i \in \{1,...,N_h\}}$.

Pour montrer l'existence et l'unicité de solution, il suffit donc de montrer que $\mathcal{A}_h$ est inversible.

Commençons par remarquer que si $(u_h, v_h) \in \mathcal{V}_h \times \mathcal{V}_h$, alors si U_h (resp. V_h) désigne le vecteur de coordonnées de u_h (resp. v_h) dans la base $\mathcal{B}_h$:

$$a(u_h, v_h) = a(\sum_{i=1}^{N_h} u_h^i \varphi_i, \sum_{j=1}^{N_h} v_h^j \varphi_j) \tag{V.37}$$

$$= a(\sum_{i=1}^{N_h} u_h^i \varphi_i, \sum_{j=1}^{N_h} v_h^j \varphi_j) \tag{V.38}$$

$$= \sum_{i=1}^{N_h} \sum_{j=1}^{N_h} u_h^i v_h^j a(\varphi_i, \varphi_j) \tag{V.39}$$

$$= {}^t U_h \mathcal{A}_h V_h \tag{V.40}$$

Ensuite, soit $v_h \in \mathcal{V}_h$, de cordonnées $(v_h^i)_{i\in\{1,\dots,N_h\}}$ dans la base $\mathcal{B}_h$. On note V_h le vecteur de coordonnées $(v_h^i)_{i\in\{1,\dots,N_h\}}$. On a $\mathcal{A}_h V_h = 0 \Rightarrow {}^t V_h \mathcal{A}_h V_h = 0$. Mais par (V.40), on a ${}^t V_h \mathcal{A}_h V_h = a(v_h, v_h)$ et comme a est coercive, on a $a(v_h, v_h) \geq \alpha \|v_h\|^2$ et on en déduit que $\|v_h\| = 0$ et donc $v_h = 0$.

On a donc $\mathcal{A}_h$ inversible et existence et unicité d'une solution au système linéaire.

De plus, si a est symétrique, alors vu l'expression de $\mathcal{A}_h$, on voit que $\mathcal{A}_h$ est symétrique. En utilisant ensuite ${}^t V_h \mathcal{A}_h V_h = a(v_h, v_h)$ et la coercivité de a, on a que $\mathcal{A}_h$ est symétrique définie positive.

□

Remarque V.3.2 *On aurait aussi pu montrer l'existence et l'unicité de la solution par le théorème de Lax-Milgram.*

Remarque V.3.3 *Lorsque a est symétrique, on déduit aussi du paragraphe précédent que u_h réalise le minimum de la fonctionnelle associée sur $\mathcal{V}_h$, et donc on s'est ramené à un problème d'optimisation en dimension finie d'une fonctionnelle quadratique.*

V.3.2 Erreur.

On aimerait maintenant pouvoir quantifier l'erreur commise entre la solution exacte u et son approximation u_h. C'est l'objet du Lemme suivant qui relie la qualité de l'approximation à une erreur de meilleure approximation de l'espace $\mathcal{V}_h$.

Lemma V.3.4 ***Lemme de Céa.*** *On suppose que $(\mathcal{V}, \langle\cdot,\cdot\rangle)$ est un espace de Hilbert réel, $\mathcal{V}_h$ un sous espace de dimension finie de $\mathcal{V}$. Soit $a : \mathcal{V}\times\mathcal{V} \to \mathbb{R}$ une forme bilinéaire continue (avec constante de continuité $M > 0$) et coercive (avec constante de coercivité $\alpha > 0$) sur $\mathcal{V}\times\mathcal{V}$ et $l : \mathcal{V} \to \mathbb{R}$ une forme linéaire continue sur $\mathcal{V}$. On considère $u \in \mathcal{V}$ la solution de $(\mathcal{F})$ et u_h la solution de $(\mathcal{F}_h)$. On a alors l'estimation suivante*

$$\|u - u_h\| \leq \frac{M}{\alpha} \inf_{v_h\in\mathcal{V}_h} \|u - v_h\|. \tag{V.41}$$

Preuve : Tout d'abord, il y a existence et unicité d'une solution $u \in \mathcal{V}$ à $(\mathcal{F})$ et $u_h \in \mathcal{V}_h$ à $(\mathcal{F}_h)$ par le Lemme de Lax-Milgram dont les hypothèses sont vérifiées.
Comme $\mathcal{V}_h \subset \mathcal{V}$, on déduit de $(\mathcal{F})$ et $(\mathcal{F}_h)$ que

$$a(u - u_h, z_h) = 0,\ \forall z_h \in \mathcal{V}_h.$$

En utilisant la coercivité de a, ce qui précède et la continuité de a, on a pour tout $w_h \in \mathcal{V}_h$,

$$\alpha\|u-u_h\|^2 \leq a(u-u_h, u-u_h) = a(u-u_h, u-w_h+w_h-u_h) = a(u-u_h, u-w_h) \leq M\|u-u_h\|\|u-w_h\|. \tag{V.42}$$

Si $\|u - u_h\| \neq 0$, alors en divisant l'inégalité du dessus par $\|u - u_h\| \neq 0$, on a le résultat voulu en passant à l'inf sur les $w_h \in \mathcal{V}_h$.

Si $\|u - u_h\| = 0$, l'inégalité demandée est encore vraie.

Ce qui donne le résultat voulu. □

V.3.3 La méthode des éléments finis, principe et stratégie de résolution.

De ce qui précède, on envisage une stratégie de résolution. Il suffirait de créer une suite d'espaces $\mathcal{V}_h$ de paramètre h, tels que la quantité $\inf_{v_h \in \mathcal{V}_h} \|u - v_h\| \to 0$ lorsque $h \to 0$ (dans le Lemme de Céa). Cela assurerait que lorsque $h \to 0$, $u_h \to u$ dans $\mathcal{V}$.

Mais il se pose alors plusieurs questions :

(a) Comment créer ces espaces $\mathcal{V}_h$ pour que l'erreur de meilleure approximation de $\mathcal{V}_h$, $\inf_{v_h \in \mathcal{V}_h} \|u - v_h\|$, tende vers 0 lorsque h tend vers 0 ?

(b) Comment choisir $\mathcal{V}_h$ pour que l'on sache construire assez facilement une base de $\mathcal{V}_h$, nous permettant ainsi de résoudre $(\mathcal{F}_h)$ par résolution d'un système linéaire ? Et que celui-ci soit "facilement" résoluble.

La méthode des Éléments Finis permet de répondre à ces deux questions. Elle permet de construire de tels espaces avec $\lim_{h \to 0} N_h = +\infty$.

On commencera par bien distinguer le domaine ouvert sur lequel est posé le problème d'EDP de départ que l'on nommera Ω ($\Omega =]0, 1[$ dans la première section) et l'espace où l'on cherche la fonction solution u, noté $\mathcal{V}$ (un exemple d'espace $\mathcal{V}$ en première section). L'espace $\mathcal{V}$ dépend bien sûr du domaine Ω.

Pour construire l'approximation de la solution, on va se baser sur des approximations polynomiales par morceaux, i.e. les solutions discrètes u_h seront à rechercher dans un espace $\mathcal{V}_h$ constitué de polynômes par morceaux.

Pour la construction de tels espaces $\mathcal{V}_h$, les méthodes d'éléments finis s'appuient sur une discrétisation du domaine Ω (que l'on appellera un *maillage du domaine*). Le paramètre h s'interprétera alors comme le pas (la taille) du maillage et la limite $h \to 0$, comme le fait de considérer des maillages de plus en plus fins. Ce maillage rendra plus aisée la construction de la base $\mathcal{B}_h$ de l'espace $\mathcal{V}_h$ constituée des fonctions $(\varphi_i)_{i \in \{1, \dots, N_h\}}$. Ces fonctions auront des support localisés, ce qui signifie que leur support sera limité à quelques éléments du maillage.

On peut résumer les étapes à suivre pour approcher la solution d'une EDP par Éléments finis comme suit :

Etapes.

(1) Si on part d'une EDP, on identifie Ω, le domaine sur lequel doit être résolue l'EDP, les conditions aux limites, les opérateurs différentiels entrant en jeu (∇, div, Laplacien, un éventuel second membre,

(2) Identifier la formulation variationnelle associée au problème d'EDP : on identifie $\mathcal{V}$, a, l. La stratégie classique (comme en section 1.) pour l'obtenir consiste à

  (a) Multiplier l'équation par une fonction générale (que l'on appellera *fonction test*) de régularité bien choisie avec des conditions aux limites éventuelles (espace $\mathcal{V}$).

  (b) Intégrer l'équation obtenue sur le domaine Ω.

  (c) Faire une intégration par partie sur le terme portant sur l'opérateur différentiel qui s'ap-

plique à la solution recherchée (on cherche à faire baisser l'ordre de la dérivée qui apparaît dans la formulation initiale de l'EDP).

(d) Utiliser les conditions aux limites sur la solution à chercher ou sur la fonction test (au besoin, on ajuste l'espace $\mathcal{V}$)

(e) Identifier $\mathcal{V}$, a et l qui constitueront $(\mathcal{F})$.

(3) On discrétise le domaine Ω en utilisant un maillage $\mathcal{T}_h$ et on construit à partir de là un espace discret $\mathcal{V}_h$ (et les fonctions de base),

(4) On définit le problème variationnel discret $(\mathcal{F}_h)$,

(5) On résout le problème discret (écriture du système linéaire : on forme la matrice, le second membre à l'aide des fonctions de base et résolution du système linéaire),

(6) On post-traite les résultats (visualisation, calcul d'erreur par rapport à la solution exacte, précision).

Chapitre VI

Éléments finis $\mathbb{P}_1$ en 1D pour l'équation de Poisson.

Pour construire un premier exemple d'espace d'éléments finis, on choisit des approximations linéaires par morceaux pour construire l'espace de discrétisation $\mathcal{V}_h$. On se focalise sur une équation modèle : ici l'équation de Poisson en dimension 1. On suivra les "**Étapes**" à la fin du chapître précédent.

VI.1 Équation modèle et sa formulation variationnelle

On rappelle que l'espace $\mathcal{V}$ considéré pour la formulation variationnelle sera un espace de fonctions définies sur un domaine donné que l'on note Ω, le domaine où doit être résolue l'équation.
Ici, on choisit de considérer l'équation de Poisson en dimension 1 et on définit le domaine Ω comme étant l'intervalle $[\alpha, \beta]$ avec $(\alpha, \beta) \in \mathbb{R}^2$ et $\alpha < \beta$.

Donnons les étapes permettant d'arriver à la formulation variationnelle.

VI.1.1 Équation considérée

On écrit l'équation, on identifie le domaine sur lequel elle est posée, les conditions de bords et on identifie ses diverses caractéristiques.
L'équation considérée est :

$$\begin{aligned} -u''(x) &= f(x), \text{ pour } x \in]\alpha, \beta[, && \text{(VI.1)} \\ u(\alpha) &= 0, && \text{(VI.2)} \\ u(\beta) &= 0. && \text{(VI.3)} \end{aligned}$$

C'est une équation avec un opérateur dérivée seconde, un second membre (donné par f). Les conditions aux limites sont des *conditions de Dirichlet homogènes.* Elles sont données par $u(\alpha) = u(\beta) = 0$.

VI.1.2 On identifie la formulation variationnelle

On établit la formulation variationnelle et on identifie l'espace $\mathcal{V}$. Sa formulation variationnelle a été étudiée en première section sur l'intervalle $[0,1]$. On peut l'étendre à l'intervalle $]\alpha,\beta[$ (*à faire en exercice*). On notera $\mathcal{V}$ l'espace des fonctions sur lequel sera posé la formulation variationnelle. L'espace $\mathcal{V}$ correct est l'espace de Hilbert $H_0^1([\alpha,\beta])$ (pour ceux n'ayant jamais vu cet espace, gardez à l'idée que cet espace est une espèce de généralisation de l'espace $\mathcal{C}^1([\alpha,\beta])$ pour des fonctions n'étant que $L^2([\alpha,\beta])$ avec les conditions de bords). On trouve que la formulation variationnelle s'écrit : *Trouver $u \in \mathcal{V}$ tel que pour tout $v \in \mathcal{V}$,*

$$\int_\alpha^\beta u'(x)v'(x)dx = \int_\alpha^\beta f(x)v(x)dx. \tag{VI.4}$$

VI.1.3 Étude de la formulation variationnelle

Une fois la formulation variationnelle identifiée, on peut voir si on sait prouver existence et unicité d'une solution au problème variationnel. Pour cela on peut par exemple tenter d'appliquer le Lemme de Lax-Milgram. Ici on voit que $a : \mathcal{V}\times\mathcal{V} \to \mathbb{R}, (u,v) \mapsto \int_\alpha^\beta u'(x)v'(x)dx$ et $l : \mathcal{V} \to \mathbb{R}, v \mapsto \int_\alpha^\beta f(x)v(x)dx$.

VI.2 Discrétisation par éléments finis.

On va maintenant détailler les étapes pour développer l'approximation numérique de la solution par éléments finis.

VI.2.1 Maillage du domaine

On commence par effectuer ***un maillage* de ce domaine**. Ici cela correspond à "décrire" un domaine géométrique donné par un nombre fini d'entités (*éléments* ou *cellules*) géométriques (segments en 1D, triangles, quadrilatères, polygones en 2D, tétrahèdres, polyhèdres en 3D). Ces *éléments* ou *cellules* géométriques sont décrites par la donnée de *sommets*.

Pour un maillage donné, on dispose donc de plusieurs nombres : Le nombre de *sommets* du maillage (n_s) et le nombre de *cellules* (ou *éléments*) du maillage (n_c).

Ici, en 1D, on veut faire un maillage de l'intervalle $[\alpha,\beta]$. On utilisera donc une discrétisation de l'intervalle $[\alpha,\beta]$ obtenue à l'aide d'une subdivision de $[\alpha,\beta]$. Choisissons une subdivision uniforme par exemple (on aurait pu prendre plus général) de $N+2$ points ($N \in \mathbb{N}^*$) et de pas $h = \frac{\beta-\alpha}{N+1} > 0$. On a donc choisi $(x_i)_{i\in\{0,..,N+1\}}$ telle que $x_0 = \alpha$, $x_{N+1} = \beta$ et $x_{i+1} - x_i = h$ pour tout $i \in \{0,..,N\}$. On notera par la suite pour $i \in \{0,...,N\}$, $I_i = [x_i, x_{i+1}]$.

Les éléments (ou cellules) du maillage sont tous les segments I_i et les sommets sont tous les x_i.

Ici, le nombre de sommets du maillage est $N+2$ ($n_s = N+2$ points dans la subdivision) et le nombre de cellules est $N+1$ ($n_c = N+1$ intervalles dans le maillage). De plus, il y a 2 points (ou

sommets) de bord (x_0 et x_{N+1}) et N points intérieurs ($(x_i)_{i\in\{1,...,N\}}$). Enfin, il y a 2 sommets par cellules.

VI.2.2 Définition de l'espace $\mathcal{V}_h$ et premières remarques.

À partir de ce maillage, on définit l'espace $\mathcal{V}_h$ de dimension finie, tel que $\mathcal{V}_h \subset \mathcal{V}$. On définit ensuite l'espace [1]

$$\mathcal{V}_h := \left\{u \in \mathcal{C}^0([\alpha,\beta]),\ \text{ pour tout } i \in \{0,...,N\}, u_{/I_i} \in \mathbb{P}_1(I_i), u(\alpha) = u(\beta) = 0\right\},$$

où $\mathbb{P}_1(J)$ est l'espace des polynômes sur un intervalle J de degré inférieur ou égal à 1.

On admettra que $\mathcal{V}_h \subset \mathcal{V}$. On peut, à partir de là, montrer que $\mathcal{V}_h$ est un sous-espace vectoriel de $\mathcal{V}$. En effet si $(\lambda,\mu) \in \mathbb{R}^2$ et $(v,w) \in \mathcal{V}_h \times \mathcal{V}_h$, alors pour tout $i \in \{0,...,N\}$, $(\lambda v + \mu w)_{/I_i}$ est encore un polynôme de degré au plus 1 et donc appartient à $\mathbb{P}_1$ (c'est un espace vectoriel). De plus $\lambda v + \mu w$ est encore continue sur $[\alpha,\beta]$ comme combinaison linéaire de fonctions continues. Enfin $\lambda v(\alpha) + \mu w(\alpha) = 0$ et $\lambda v(\beta) + \mu w(\beta) = 0$, puisque $v(\alpha) = v(\beta) = 0$ et $w(\alpha) = w(\beta) = 0$. Donc $\lambda v + \mu w \in \mathcal{V}_h$. Et $\mathcal{V}_h$ est un sous espace vectoriel de $\mathcal{V}$.

Essayons de décortiquer un peu $\mathcal{V}_h$. Sur chaque sous-intervalle I_i, un élément de $\mathcal{V}_h$ est déterminé de façon unique dès que deux valeurs de cette fonction sont connues. En effet, si $u \in \mathcal{V}_h$, on sait que sur chaque I_i, u est un polynôme de degré au plus 1, donc s'écrit $ax+b$ avec $(a,b) \in \mathbb{R}$ à déterminer. Il y a donc deux inconnues à déterminer qui peuvent l'être en fixant deux valeurs de la fonction en deux points donnés. Il y a donc 2 inconnues à déterminer pour déterminer entièrement la fonction sur cet intervalle (on peut appeler ces inconnues des *degrés de liberté* ou *noeuds*). De plus, pour les intervalles I_0 (resp. I_N), une valeur est déjà connue : $u(\alpha) = 0$ (resp. $u(\beta) = 0$). Il faut également s'assurer de la continuité de l'approximation.

Un choix naturel pour déterminer ces deux degrés de liberté sur chaque intervalle I_i est de prendre les valeurs en les noeuds x_i et x_{i+1} et on détermine donc la fonction en utilisant la valeur de la fonction aux extrémités de chaque intervalle I_i. On pourrait donc penser qu'il y a $2*(N+1)-2$ degrés de libertés pour déterminer la fonction (2 par intervalle moins les deux du bord). Mais les fonctions de $\mathcal{V}_h$ doivent être continues, donc pour assurer la continuité, il suffit d'attribuer en chaque x_i la même valeur à droite et à gauche (i.e. si $v \in \mathcal{V}_h$, $v(x_i^-) = v(x_i^+)$ pour tout $i \in \{1,...,N\}$, ou encore $v_{/I_{i-1}}(x_i) = v_{/I_i}(x_i)$), pour tout $i \in \{1,...,N\}$. Il n'y a donc en fait que N degrés de liberté à fixer pour déterminer l'expression de la fonction de $\mathcal{V}_h$ recherchée. **Une fonctions de $\mathcal{V}_h$ est en fait entièrement déterminée par ses valeurs aux noeuds $(x_i)_{i\in\{1,...,N\}}$ (qui sont les sommets internes).** On va préciser un peu ce raisonnement et montrer que la dimension de $\mathcal{V}_h$ est N en passant par la détermination d'une base de $\mathcal{V}_h$.

Remarque VI.2.1 *Il faut bien remarquer que N est lié à h (le pas de la subdivision) par $h = \frac{\beta-\alpha}{N+1}$.*

1. Ici, avec notre choix, on peux montrer que $\mathcal{V}_h \subset \mathcal{V} = H_0^1([\alpha,\beta])$ **(admis)**.

VI.2.3 Détermination d'une base de $\mathcal{V}_h$

Un aspect important est maintenant de déterminer une base de l'espace $\mathcal{V}_h$. On peut voir qu'une base de l'espace est donnée par les fonctions dites *fonctions de Lagrange*. Ces fonctions sont définies comme suit. Pour $i \in \{1, ..., N\}$, on définit $\varphi_i : [\alpha, \beta] \to \mathbb{R}$ telle que $\varphi_i \in \mathcal{V}_h$ et pour tout $j \in \{1, ..., N\}$, $\varphi_i(x_j) = \delta_{i,j}$ (où $\delta_{i,j}$ est le symbole de Kronecher, i.e. $\delta_{i,j} = 0$, si $i \neq j$ et $\delta_{i,i} = 1$).
Déterminons les $(\varphi_i)_{i\in\{1,...,N\}}$. On peut avoir une expression explicite de ces fonctions. Pour cela, on écrit que sur chaque I_i (avec $i \in \{0, .., N\}$), et $x \in I_i$, $\varphi_i(x)$ s'écrit $a_i x + b_i$ avec $(a_i, b_i) \in \mathbb{R}^2$ à déterminer. En écrivant les égalités imposées par la définition de φ_i, on obtient [2]

$$\varphi_i(x) = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}}, & \text{pour } x_{i-1} \leq x \leq x_i, \\ \frac{x_{i+1} - x}{x_{i+1} - x_i}, & \text{pour } x_i \leq x \leq x_{i+1}, \\ 0 \text{ sinon.} \end{cases} \tag{VI.5}$$

Une représentation graphique est donnée en figure VI.1.

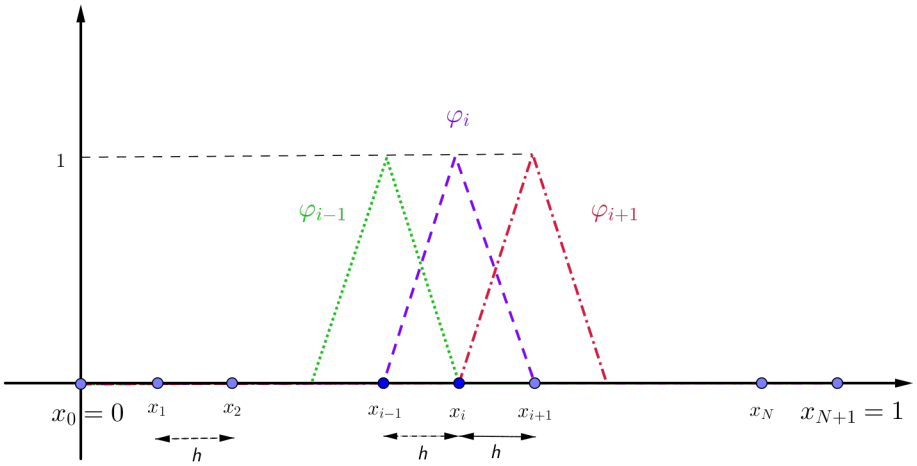

FIGURE VI.1 – Représentation graphique des fonctions de base $\mathbb{P}_1$. La représentation est faite sur $[0, 1]$, la passer sur $[\alpha, \beta]$

Remarque VI.2.2 *On peut faire plusieurs remarques.*

- *On voit que φ_i pour $i \in \{1, ..., N\}$ est nulle en dehors de $[x_{i-1}, x_{i+1}]$. On dira que son support est donc $[x_{i-1}, x_{i+1}]$.*
- *Ces fonctions sont souvent appelées les fonctions chapeau, du fait de leur forme graphique.*

Proposition VI.2.3 *La famille $(\varphi_i)_{i\in\{1,...,N\}}$ constitue une base de $\mathcal{V}_h$. La dimension de $\mathcal{V}_h$ est donc $N = \frac{\beta-\alpha}{h} - 1$.*

2. Faire le calcul en exercice.

Preuve : On peut montrer que la famille est libre. En effet considérons une combinaison linéaire nulle, i.e. $(\lambda_1, ..., \lambda_N) \in \mathbb{R}^N$ telle que

$$\sum_{j=1}^{N} \lambda_j \varphi_j = 0, \text{ sur } [\alpha, \beta]. \tag{VI.6}$$

Soit $i \in \{1, .., N\}$, on a donc en appliquant l'égalité ci-dessus à x_i,

$$\sum_{j=1}^{N} \lambda_j \varphi_j(x_i) = 0, \text{ sur } [\alpha, \beta]. \tag{VI.7}$$

En utilisant la définition de φ_i, cette égalité nous donne :

$$\lambda_i \underbrace{\varphi_i(x_i)}_{=1} = 0, \text{ sur } [\alpha, \beta]. \tag{VI.8}$$

Et donc $\lambda_i = 0$. On a donc ce résultat pour tout $i \in \{1, ..., N\}$, et la famille est donc libre. C'est une famille de N vecteurs libres de $\mathcal{V}_h$.

Soit $v_h \in \mathcal{V}_h$. Montrons que v_h sécrit comme une combinaison linéaire des $(\varphi_i)_{i\in\{1,...,N\}}$. Plus précisément montrons que

$$v_h = \sum_{i=1}^{N} v_h(x_i)\varphi_i. \tag{VI.9}$$

Tout d'abord, notons $w_h := \sum_{i=1}^{N} v_h(x_i)\varphi_i$. On sait que w_h est un élément de $\mathcal{V}_h$ puisque pour tout $i \in \{1, ..., N\}$, $\varphi_i \in \mathcal{V}_h$ et que $\mathcal{V}_h$ est un sous-espace vectoriel de $\mathcal{V}$.

De plus pour chaque $i \in \{1, ..., N\}$ $w_h(x_i) = v_h(x_i)$ vu la définition des $(\varphi_i)_{i\in\{1,...,N\}}$. Enfin, pour $i \in \{0, ..., N\}$ sur chaque I_i, $w_h - v_h$ est un polynôme de degré au plus 1 qui prend la valeur 0 en x_i et x_{i+1}. $w_h - v_h$ est donc un polynôme de degré au plus 1 qui admet deux racines distinctes, c'est donc le polynôme nul. Donc pour tout $i \in \{0, ..., N\}$, $w_h = v_h$ sur I_i. En conclusion $v_h = w_h$. La conclusion suit. □

VI.2.4 Formulation variationnelle discrète et résolution

Formulation variationnelle discrète et système linéaire associé

La formulation variationnelle discrète s'écrit : *Trouver $u_h \in \mathcal{V}_h$ tel que pour tout $v_h \in \mathcal{V}_h$,*

$$a(u_h, v_h) = l(v_h). \tag{VI.10}$$

Comme $(\varphi_i)_{i\in\{1,...,N\}}$ est une base de $\mathcal{V}_h$, on a vu dans le chapitre précédent que la formulation variationnelle discrète est équivalente à : *Trouver $u_h \in \mathcal{V}_h$ tel que pour tout $i \in \{1, ..., N\}$,*

$$a(u_h, \varphi_i) = l(\varphi_i). \tag{VI.11}$$

Et on a enfin vu que cela revenait à résoudre le système linéaire (V.36).

À l'aide des fonctions de base, on forme la matrice du système linéaire. On a vu dans la preuve de la proposition V.3.1 que $\mathcal{A}_h$ est donnée par $(a(\varphi_j, \varphi_i))_{(i,j)\in\{1,...,N\}}$.

Remarque VI.2.4 *On remarque que a est symétrique, donc* $a(\varphi_j, \varphi_i) = a(\varphi_i, \varphi_j)$, $\forall (i,j) \in \{1, ..., N\}^2$.

On va maintenant chercher un moyen de calculer la matrice du système linéaire de façon efficace [3].

Notion d'intervalle de référence

On pourrait calculer directement les termes de la matrice en utilisant les expressions de chaque fonction de base sur $[\alpha, \beta]$ (*Exercice*). Mais on va privilégier une autre méthode qui montrera surtout tous ses avantages lorsqu'on utilisera des polynômes de degré supérieur ou en dimension supérieure. Pour calculer les intégrales qui interviennent dans les coefficients de $\mathcal{A}_h$, on va utiliser une notion d'intervalle de référence et s'y ramener par le changement de variable.

En effet, à un changement de variable près, tous les intervalles I_i pour $i \in \{0, ..., N\}$ peuvent se ramener à l'intervalle $[0, 1]$. En effet pour $i \in \{0, ..., N\}$, définissons l'application

$$F_i : [0,1] \to I_i, \xi \mapsto (x_{i+1} - x_i)\xi + x_i.$$

Elle envoie bien $[0, 1]$ dans I_i et de plus $F_i(0) = x_i$ et $F_i(1) = x_{i+1}$. On remarque également que F_i est une fonction affine. Elle est de plus inversible et $F_i^{-1} : I_i \to [0, 1], x \mapsto \frac{x - x_i}{x_{i+1} - x_i}$.

On appellera l'intervalle $[0, 1]$, l'*intervalle de référence* (ou *élément de référence*).

Sur un intervalle donné par $i \in \{0, ..., N-1\}$, I_i, seules φ_i et φ_{i+1} sont non nulles. Les restrictions de ces deux fonctions sur I_i peuvent de plus être obtenues à partir de deux fonctions élémentaires polynomiales sur $[0, 1]$ et de degré ≤ 1. On les note $\hat{\varphi}_0$ et $\hat{\varphi}_1$ définies sur l'intervalle $[0, 1]$. Ces deux fonctions sont définies par $\hat{\varphi}_0(\xi) = 1 - \xi$ et $\hat{\varphi}_1(\xi) = \xi$ pour $\xi \in [0, 1]$. Une représentation graphique est donnée en Figure VI.2.

Et on a pour $i \in \{1, ..., N\}$, $(\varphi_i)_{/I_i} = \hat{\varphi}_0 \circ F_i^{-1}$, $(\varphi_{i+1})_{/I_i} = \hat{\varphi}_1 \circ F_i^{-1}$. Ce qui peut s'illustrer par la Figure VI.3.

Calcul des termes de la matrice à partir de l'élement de référence

Tout d'abord, pour $i \in \{1, ..., N\}$, tous les termes $a(\varphi_i, \varphi_j) = \int_\alpha^\beta \varphi_i' \varphi_j'$ sont nuls si $|i - j| \geq 2$, puisque nous avons vu que le support de φ_i est réduit à $[x_{i-1}, x_{i+1}]$ et que $]x_{i-1}, x_{i+1}[\cap]x_{j-1}, x_{j+1}[= \emptyset$ si $|i - j| \geq 2$. Il n'y a donc à calculer que les termes $(a(\varphi_i, \varphi_i))_{i \in \{1,...,N\}}$ et $(a(\varphi_i, \varphi_{i+1}))_{i \in \{1,...,N-1\}}$, $(a(\varphi_i, \varphi_{i-1}))_{i \in \{2,...,N\}}$. Plus précisément, on a :

$$a(\varphi_i, \varphi_i) = \int_\alpha^\beta (\varphi_i'(x))^2 dx = \int_{I_{i-1}} (\varphi_i'(x))^2 dx + \int_{I_i} (\varphi_i'(x))^2 dx. \tag{VI.12}$$

Chacun de ces deux termes peut être calculé individuellement. Tout d'abord, on a sur I_i, $\varphi_i = \hat{\varphi}_0 \circ F_i^{-1}$ et $\varphi_{i+1} = \hat{\varphi}_1 \circ F_i^{-1}$ et donc $\varphi_i'(x) = \hat{\varphi}_0'(F_i^{-1}(x)) F_i^{-1'}(x)$. On a donc

$$\int_{I_i} (\varphi_i'(x))^2 dx = \int_{I_i} (\hat{\varphi}_0'(F_i^{-1}(x)) F_i^{-1'}(x))^2 dx. \tag{VI.13}$$

3. l'efficacité sera plus parlante lorsque l'on passera à des polynômes d'ordre supérieur à 2 ou en dimension supérieure

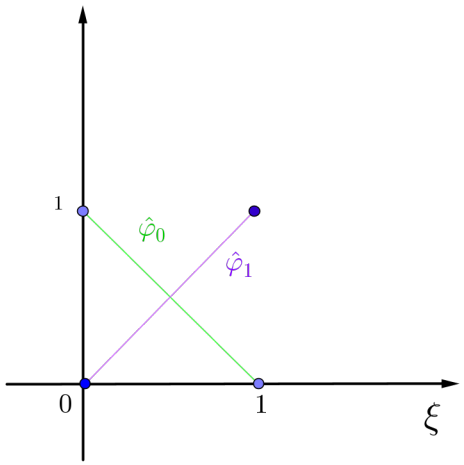

FIGURE VI.2 – Représentation graphique des fonctions de référence $\mathbb{P}_1$.

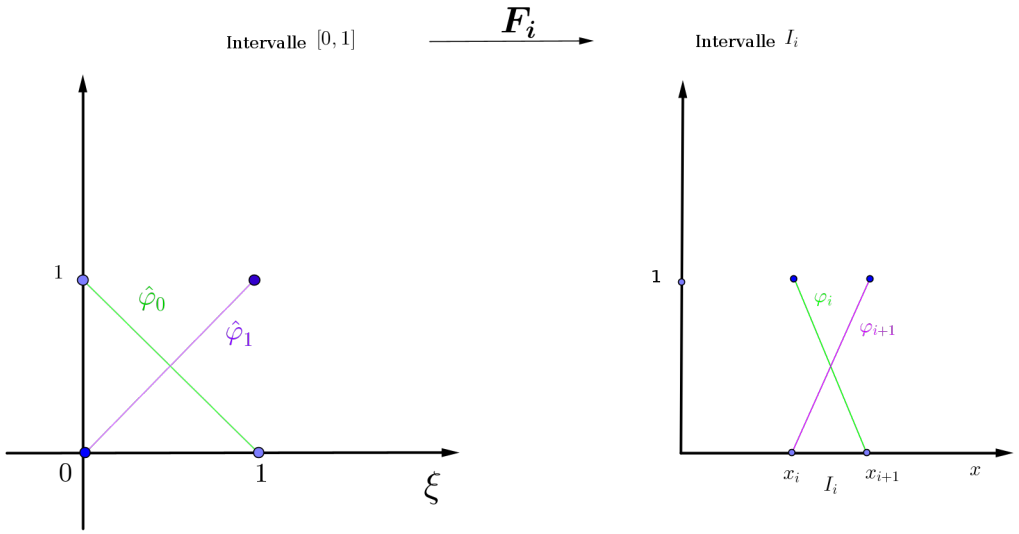

FIGURE VI.3 – Passage de l'élément de référence à l'intervalle I_i.

On définit un changement de variable en posant $\xi = F_i^{-1}(x)$, ce qui donne $d\xi = F_i^{-1'}(x)dx$ avec $F_i^{-1'}(x) = \frac{1}{x_{i+1}-x_i} = \frac{1}{h}$.

On trouve alors

$$\int_{I_i} (\varphi_i'(x))^2 dx = \frac{1}{h}\int_0^1 (\hat{\varphi}_0'(\xi))^2 d\xi = \frac{1}{h}\int_0^1 d\xi = \frac{1}{h}. \tag{VI.14}$$

De même,

$$\int_{I_{i-1}} (\varphi_i'(x))^2 dx = \int_{I_{i-1}} (\hat{\varphi}_1'(F_{i-1}^{-1}(x))F_{i-1}^{-1'}(x))^2 dx = \frac{1}{h}\int_0^1 (\hat{\varphi}_1'(\xi))^2 d\xi = \frac{1}{h}. \tag{VI.15}$$

Donc

$$\int_\alpha^\beta (\varphi_i'(x))^2 dx = \frac{2}{h}. \tag{VI.16}$$

Ensuite de la même façon, puisque l'intersection du support de φ_i et φ_{i-1} est I_{i-1}, on a

$$\int_\alpha^\beta \varphi_i'(x)\varphi_{i-1}'(x)dx = \int_{I_{i-1}} \varphi_i'(x)\varphi_{i-1}'(x)dx. \tag{VI.17}$$

Et en passant à l'élément de référence, on a

$$\int_{I_{i-1}} \varphi_i'(x)\varphi_{i-1}'(x)dx = \int_{I_{i-1}} \hat{\varphi}_1'(F_{i-1}^{-1}(x))\hat{\varphi}_0'(F_{i-1}^{-1}(x))(F_{i-1}^{-1'}(x))^2 dx. \tag{VI.18}$$

Ce qui donne

$$\int_{I_{i-1}} \varphi_i'(x)\varphi_{i-1}'(x)dx = \frac{1}{h}\int_0^1 \hat{\varphi}_1'(\xi)\hat{\varphi}_0'(\xi)d\xi. \tag{VI.19}$$

Et donc

$$\int_{I_{i-1}} \varphi_i'(x)\varphi_{i-1}'(x)dx = -\int_0^1 \frac{1}{h}d\xi = -\frac{1}{h}. \tag{VI.20}$$

De façon analogue, on a

$$\int_\alpha^\beta \varphi_i'(x)\varphi_{i+1}'(x)dx = -\frac{1}{h}. \tag{VI.21}$$

Finalement,

$$\mathcal{A}_h = \frac{1}{h}\begin{pmatrix} 2 & -1 & 0 & ... & ... & 0 \\ -1 & 2 & -1 & 0 & ... & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ 0 & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & 0 & -1 & 2 & -1 \\ 0 & ... & ... & ... & -1 & 2 \end{pmatrix} \tag{VI.22}$$

Remarque VI.2.5 *On reconnait avec cette matrice une matrice connue que vous avez vue en cours de Différence Finies, c'est la matrice du Laplacien (à un facteur $\frac{1}{h}$ près).*

Remarque VI.2.6 *On voit que tous les termes de la matrice $\mathcal{A}_h$ peuvent être calculés à partir de la matrice sur l'élément de référence (regardez* (VI.14), (VI.15), (VI.19)*), on l'appellera matrice élémentaire. Elle est donnée par :*

$$\hat{\mathcal{A}} = \begin{pmatrix} \int_0^1 (\hat{\varphi}_0'(\xi))^2 d\xi & \int_0^1 \hat{\varphi}_0'(\xi)\hat{\varphi}_1'(\xi) d\xi \\ \int_0^1 \hat{\varphi}_0'(\xi)\hat{\varphi}_1'(\xi) d\xi & \int_0^1 (\hat{\varphi}_1'(\xi))^2 d\xi \end{pmatrix}$$

Calcul du second membre

Reste maintenant à calculer le second membre du système linéaire (V.36). On a vu dans la preuve de la proposition V.3.1 que $L_h = (l(\varphi_i))_{i\in\{1,\dots,N\}}$. Il faut donc, pour avoir le second membre du système, calculer les valeurs $\int_\alpha^\beta f\varphi_i(x)dx$. On ne sait pas forcément calculer exactement ce terme. On va donc utiliser des *formules de quadratures* (Gauss) qui permettront d'approcher ces intégrales (*cf. TP*).

Numérotation

Pour pouvoir calculer efficacement les intégrales, on met en place plusieurs numérotations et un moyen simple de repérer les degrés de libertés puisqu'on a vu qu'en passant à l'élément de référence, il est essentiel d'être capable de relier la fonction de base associée à un degré de liberté (φ_i) à la bonne fonction de base sur l'élément de référence ($\hat{\varphi}_0$ ou $\hat{\varphi}_1$?).

Ainsi, les degrés de libertés pour la méthode d'éléments finis de cette section sont associés aux sommets internes du maillage $(x_i)_{i\in\{1,\dots,N\}}$. On attribue à chacun de ces noeuds un numéro : c'est ce que l'on appelle *la numérotation globale*. De la même façon, au niveau local (i.e. sur une cellule), on peut compter le nombre de degré de liberté (noté $nloc_{ddl}$) qui sont associés à des noeuds qui appartiennent à cette cellule. Dans cette section on a 2 degrés de liberté dans chaque cellule. Ce nombre est appelé le *nombre de degrés de liberté local* (par opposition au *nombre de degré de liberté global*). On donne également un numéro à chaque cellule du maillage global.

On veut ainsi trouver pour un degré de liberté, repéré par son numéro local à la cellule et son numéro de cellule, quel est son numéro global dans le maillage. Cela se matérialise par un tableau T de dimension $(nloc_{ddl} * n_c)$ qui à un numéro local de degré de liberté et de numéro global de cellule renvoie un numéro qui est le numéro du degré de liberté dans la numérotation globale. *Les détails de la mise en œuvre seront vus en TP.*

Retour sur le calcul à partir des matrices élémentaires

Ce paragraphe peut être passé en première lecture. Comme mentionné en remarque VI.2.6, on voit que tous les termes de la matrice $\mathcal{A}$ peuvent être calculés à partir de la matrice sur l'élément de référence (regardez (VI.14), (VI.15), (VI.19)), on l'appellera *matrice élémentaire*. Elle est donnée

par :

$$\hat{\mathcal{A}} = \begin{pmatrix} \int_0^1 (\hat{\varphi}_0'(\xi))^2 d\xi & \int_0^1 \hat{\varphi}_0'(\xi)\hat{\varphi}_1'(\xi) d\xi \\ \int_0^1 \hat{\varphi}_0'(\xi)\hat{\varphi}_1'(\xi) d\xi & \int_0^1 (\hat{\varphi}_1'(\xi))^2 d\xi \end{pmatrix}$$

Ce qui donne ici

$$\hat{\mathcal{A}} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

Une fois que l'on a calculé cette matrice, on a accès à tous les termes calculés plus haut. On voit que cette stratégie marche car F_i' est constante sur I_i et donc on peut "sortir" le facteur $\frac{1}{h}$ de l'intégrale. En utilisant la numérotation, on peut alors contruire la matrice $\mathcal{A}$ par une technique appelée *technique d'assemblage*. La stratégie est alors la suivante. On initialise une matrice $\mathcal{A}$ à la matrice nulle. Ensuite, on parcourt toutes les cellules du maillage en utilisant la numérotation ($l \in \{1, ..., n_c\}$). Sur la cellule l, on parcourt tous les degrés de libertés de la cellule : ici il y en a 2 par cellule, disons un de numéro 0 (extrémité gauche de la cellule c), et un de numéro 1 (extrémité droite de la cellule c). On a alors plusieurs couples (numéro de cellule,numéro de degré de liberté local), qui, grâce au tableau T de correspondance permet de repérer les degrés de liberté concernés dans la numérotation globale (ici, dans la cellule l, il y aurait donc 2 nombres k_0 et k_1 dans la numérotation globale correspondant aux deux degrés de libertés de la cellule l, respectivement de numéros locaux 0 et 1). On va alors remplir la matrice $\mathcal{A}$ à l'aide des termes de la matrice $\hat{\mathcal{A}}$. On écrit [4]

$$\mathcal{A}_{k_0 k_0} = \mathcal{A}_{k_0 k_0} + \frac{1}{h}\hat{\mathcal{A}}_{00} \tag{VI.23}$$

$$\mathcal{A}_{k_0 k_1} = \mathcal{A}_{k_0 k_1} + \frac{1}{h}\hat{\mathcal{A}}_{01} \tag{VI.24}$$

$$\mathcal{A}_{k_1 k_0} = \mathcal{A}_{k_1 k_0} + \frac{1}{h}\hat{\mathcal{A}}_{10} \tag{VI.25}$$

$$\mathcal{A}_{k_1 k_1} = \mathcal{A}_{k_1 k_1} + \frac{1}{h}\hat{\mathcal{A}}_{11} \tag{VI.26}$$

$$\tag{VI.27}$$

Et ensuite on itère sur l.

La technique d'assemblage prend donc la forme (algorithmiquement) de plusieurs boucles imbriquées (cellule puis degrés de liberté locaux).

4. Cela peut se formaliser par une double boucle.

Chapitre VII

Éléments finis $\mathbb{P}_2$ en 1D pour l'équation de Poisson.

Pour construire un second exemple d'espace d'éléments finis, on garde le même espace $\mathcal{V}$, mais on change l'espace $\mathcal{V}_h$. On choisit des approximations **quadratiques** par morceaux pour construire l'espace de discrétisation $\mathcal{V}_h$. On peut reprendre exactement le même formalisme que dans la section précédente. Puisque l'espace $\mathcal{V}_h$ change, il va falloir adapter les fonctions de bases, le système linéaire etc...

On formalise cela dans la suite.

VII.1 Equation modèle et formulation variationnelle

On garde ici la même équation modèle, la même formulation variationnelle que dans le chapître précédent. Elle s'écrit donc toujours : *Trouver $u \in \mathcal{V}$ tel que pour tout $v \in \mathcal{V}$,*

$$\int_\alpha^\beta u'(x)v'(x)dx = \int_\alpha^\beta f(x)v(x)dx. \tag{VII.1}$$

VII.2 Discrétisation par éléments finis $\mathbb{P}_2$.

On suit les mêmes étapes que pour la discrétisation par éléments finis $\mathbb{P}_1$.

VII.2.1 Maillage du domaine

On commence par effectuer ***un maillage* de ce domaine** comme pour le cas $\mathbb{P}_1$. On utilisera donc une discrétisation de l'intervalle $[\alpha, \beta]$ obtenue à l'aide d'une subdivision de $[\alpha, \beta]$. Choisissons une subdivision uniforme par exemple (on aurait pu prendre plus général) de $N+2$ points ($N \in \mathbb{N}^*$) et de pas $h = \frac{1}{N+1} > 0$. On a donc choisi $(x_i)_{i\in\{0,..,N+1\}}$ telle que $x_0 = \alpha$, $x_{N+1} = \beta$ et $x_{i+1} - x_i = h$ pour tout $i \in \{0, .., N\}$. On notera par la suite pour $i \in \{0, ..., N\}$, $I_i = [x_i, x_{i+1}]$.

Les éléments (ou cellules) du maillage sont tous les segments I_i et les sommets sont tous les x_i.

Ici, le nombre de sommets du maillage est $N+2$ ($n_s = N+2$ points dans la subdivision) et le nombre de cellules est $N+1$ ($n_c = N+1$ intervalles dans le maillage). De plus, il y a 2 points (ou sommets) de bord (x_0 et x_{N+1}) et N points intérieurs ($(x_i)_{i\in\{1,...,N\}}$). Enfin, il y a 2 sommets par cellules.

VII.3 Espace d'élément fini $\mathbb{P}_2$

VII.3.1 Définition de l'espace $\mathcal{V}_h$

L'espace discret de dimension finie $\mathcal{V}_h$ est maintenant **différent** du cas $\mathbb{P}_1$. On considère ici des approximation continues dont la restriction aux cellules du maillage sont des polynômes de degré au plus **2**. Cela donne :

$$\mathcal{V}_h := \left\{u \in \mathcal{C}^0([\alpha,\beta]),\ \text{tels que pour tout } i \in \{0,...,N\}, u_{/I_i} \in \mathbb{P}_{\mathbf{2}}(I_i),\ u(\alpha) = u(\beta) = 0\right\}, \quad \text{(VII.2)}$$

où $\mathbb{P}_{\mathbf{2}}(J)$ (avec J intervalle de $\mathbb{R}$) est l'espace des polynômes sur J de degré inférieur ou égal à **2**. Comme pour le chapitre précédent, on doit alors déterminer une base de $\mathcal{V}_h$.

Commençons par faire un raisonnement formel. Sur chaque sous-intervalle I_i, un élément de $\mathcal{V}_h$ est déterminé de façon unique dès que **trois** valeurs de cette fonction sont connues. En effet, si $u \in \mathcal{V}_h$, on sait que sur chaque I_i, u est un polynôme de degré au plus **2**, donc s'écrit $\mathbf{a_i x^2 + b_i x + c_i}$ avec $(\mathbf{a_i}, \mathbf{b_i}, \mathbf{c_i}) \in \mathbb{R}^{\mathbf{3}}$ à déterminer. Il y a donc **trois** inconnues à déterminer qui peuvent l'être en fixant **trois** valeurs de la fonction en **trois** points donnés. Il y a donc **3** inconnues à déterminer pour déterminer entièrement la fonction sur cet intervalle (on peut toujours appeler ces inconnues des *degrés de liberté* ou *noeuds*). De plus, pour les intervalles I_0 (resp. I_N), une valeur est déjà connue : $u(\alpha) = 0$ (resp. $u(\beta) = 0$).

Un choix naturel pour déterminer ces **trois** degrés de liberté est d'utiliser sur chaque intervalle I_i les valeurs aux noeuds x_i, $x_{i+\frac{1}{2}} := \frac{x_i + x_{i+1}}{2}$ et x_{i+1} et on détermine donc la fonction en utilisant la valeur de la fonction aux extrémités de chaque intervalle I_i ainsi qu'en son milieu. On pourrait donc penser qu'il y a $\mathbf{3} * (\mathbf{N+1}) - \mathbf{2}$ degrés de libertés pour déterminer la fonction (**3** par intervalle moins les deux du bord). Mais les fonctions de $\mathcal{V}_h$ doivent être continues, donc pour assurer la continuité, il suffit d'attribuer en chaque x_i (extrémités des intervalles) la même valeur à droite et à gauche (i.e. si $v \in \mathcal{V}_h$, $v(x_i^-) = v(x_i^+)$ pour tout $i \in \{1,...,N\}$, ou encore $v_{/I_{i-1}}(x_i) = v_{/I_i}(x_i)$), pour tout $i \in \{1,...,N\}$. Il n'y a donc en fait que $\mathbf{N} + (\mathbf{N+1}) = \mathbf{2N+1}$ degrés de liberté à fixer pour déterminer l'expression de la fonction de $\mathcal{V}_h$ recherchée. Ce qui correspond au nombre de points qui sont les extrémités des intervalles (moins les points de bord) + le nombre de points qui sont milieux des intervalles. **Une fonction de $\mathcal{V}_h$ est en fait entièrement déterminée par ses valeurs aux noeuds $(x_i)_{i\in\{1,...,N\}}$ (qui sont les sommets internes) et aux milieux des intervalles $(x_{i+\frac{1}{2}})_{i\in\{0,N\}}$.** En formalisant ce raisonnement, on voit que la dimension de $\mathcal{V}_h$ va être $N_h = \mathbf{2N+1}$.

VII.3.2 Détermination d'une base de $\mathcal{V}_h$: base de Lagrange

Nous allons maintenant déterminer une base de l'espace $\mathcal{V}_h$. Comme pour le cas $\mathbb{P}_1$, nous allons donner une base de l'espace utilisant les fonctions dites *fonctions de Lagrange*. Ces fonctions sont définies comme suit. On commence par renommer les noeuds utilisés pour évaluer les degrés de libertés pour que ce soit plus lisible. Pour $i \in \{1, ..., 2N+1\}$, on note $\tilde{x}_i = x_{\frac{i}{2}}$ ces $2N+1$ points. [1]. On note également $\tilde{x}_0 = x_0 = \alpha$ et $\tilde{x}_{2N+2} = x_{N+1} = \beta$. Pour $i \in \{1, ..., 2N+1\}$, on définit $\varphi_i : [\alpha, \beta] \to \mathbb{R}$ telle que $\varphi_i \in \mathcal{V}_h$ et pour tout $j \in \{1, ..., 2N+1\}$, $\varphi_i(\tilde{x}_j) = \delta_{i,j}$ (où $\delta_{i,j}$ est le symbole de Kronecker, i.e. $\delta_{i,j} = 0$, si $i \neq j$ et $\delta_{i,i} = 1$). Déterminons les $(\varphi_i)_{i\in\{1,...,2N+1\}}$. On peut avoir une expression explicite de ces fonctions. Pour cela, on écrit que sur chaque I_i (avec $i \in \{0, .., N\}$), et $x \in I_i$, $\varphi_i(x)$ s'écrit $a_i x^2 + b_i x + c_i$ avec $(a_i, b_i, c_i) \in \mathbb{R}^3$ à déterminer. En écrivant les égalités imposées par la définition de φ_i, on obtient [2]

Si $j \in \{1, ..., 2N+1\}$ est pair,

$$\varphi_j(x) = \begin{cases} \frac{(x-\tilde{x}_{j-1})(x-\tilde{x}_{j-2})}{(\tilde{x}_j-\tilde{x}_{j-1})(\tilde{x}_j-\tilde{x}_{j-2})}, & \text{pour } \tilde{x}_{j-2} \leq x \leq \tilde{x}_j, \\ \frac{(\tilde{x}_{j+1}-x)(\tilde{x}_{j+2}-x)}{(\tilde{x}_{j+1}-\tilde{x}_j)(\tilde{x}_{j+2}-\tilde{x}_j)}, & \text{pour } \tilde{x}_j \leq x \leq \tilde{x}_{j+2}, \\ 0 \text{ sinon.} \end{cases} \tag{VII.3}$$

Si $j \in \{1, ..., 2N+1\}$ est impair,

$$\varphi_j(x) = \begin{cases} \frac{(\tilde{x}_{j+1}-x)(x-\tilde{x}_{j-1})}{(\tilde{x}_{j+1}-\tilde{x}_j)(\tilde{x}_j-\tilde{x}_{j-1})}, & \text{pour } \tilde{x}_{j-1} \leq x \leq \tilde{x}_{j+1}, \\ 0 \text{ sinon.} \end{cases} \tag{VII.4}$$

Une représentation graphique est donnée en figure VII.1.

Remarque VII.3.1 *On peut faire plusieurs remarques.*

- *Le support d'une fonction φ_j pour $j \in \{1, ..., 2N+1\}$ donné est $[\tilde{x}_{j-2}, \tilde{x}_{j+2}]$, si j est pair et $[\tilde{x}_{j-1}, \tilde{x}_{j+1}]$, si j est impair.*
- *On voit qu'ici, on distinge les expressions des fonctions de base associées aux extrémités des intervalles de celles associées aux milieux des intervalles.*
- *Les fonctions associées aux milieux des intervalles sont souvent appelées fonctions bulles du fait de leur forme.*

Avec le même raisonnement que dans le cas $\mathbb{P}_1$, on peut montrer que la famille est libre et génératrice.

Proposition VII.3.2 *La famille $(\varphi_i)_{i\in\{1,...,2N+1\}}$ constitue une base de $\mathcal{V}_h$. La dimension de $\mathcal{V}_h$ est donc $2N+1$.*

1. De cette façon, les indices pairs réfèrent à une extrémité d'intervalle et les indices impairs à un milieu d'intervalle.
2. Faire le calcul en exercice. Pour ceux qui se souviennent de leur cours d'interpolation polynomiale, on reconnaît là des polynômes de Lagrange.

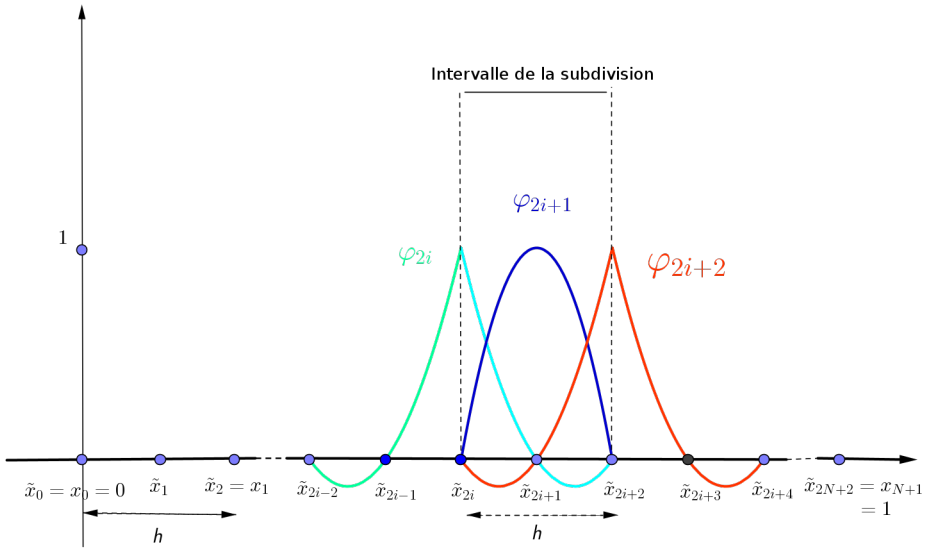

FIGURE VII.1 – Représentation graphique des fonctions de base $\mathbb{P}_2$.

Preuve : Comme pour le cas $\mathbb{P}_1$, on montre que la famille est libre. Le raisonnement étant identique, on ne le reproduit pas ici. De même on peut montrer que tout élément $v_h \in \mathcal{V}_h$ sécrit

$$v_h = \sum_{i=1}^{2N+1} v_h(\tilde{x}_i)\varphi_i. \tag{VII.5}$$

Ici encore, comme le raisonnement est très similaire au cas $\mathbb{P}_1$, on ne le redétaille pas, mais **faites le comme un exercice**. □

VII.3.3 Formulation variationnelle discrète et résolution

Formulation variationnelle discrète et système linéaire associé

La formulation variationnelle discrète s'écrit : *Trouver* $u_h \in \mathcal{V}_h$ *tel que pour tout* $v_h \in \mathcal{V}_h$,

$$a(u_h, v_h) = l(v_h). \tag{VII.6}$$

Comme $(\varphi_i)_{i\in\{1,\dots,2N+1\}}$ est une base de $\mathcal{V}_h$, on a vu dans le chapitre précédent que la formulation variationnelle discrète est équivalente à : *Trouver* $u_h \in \mathcal{V}_h$ *tel que pour tout* $i \in \{1, ..., 2N+1\}$,

$$a(u_h, \varphi_i) = l(\varphi_i). \tag{VII.7}$$

Et on a enfin vu que cela revenait à résoudre le système linéaire (V.36).

À l'aide des fonctions de base, on forme la matrice du système linéaire. On a vu dans la preuve de la proposition V.3.1 que $\mathcal{A}_h$ est donnée par $(a(\varphi_j, \varphi_i))_{(i,j)\in\{1,\dots,2N+1\}}$.

Remarque VII.3.3 *On remarque que a est symétrique, donc* $a(\varphi_j, \varphi_i) = a(\varphi_i, \varphi_j)$, $\forall (i,j) \in \{1, ..., 2N+1\}^2$.

On va maintenant chercher un moyen de calculer la matrice du système linéaire de façon efficace.

Calcul via l'intervalle de référence

On va maintenant chercher un moyen de calculer la matrice du système linéaire de façon efficace. Tout d'abord, comme pour le cas $\mathbb{P}_1$, à un changement de variable près tous les intervalles $I_i = [\tilde{x}_{2i}, \tilde{x}_{2i+2}] (= [x_i, x_{i+1}])$ pour $i \in \{0, ..., N\}$ peuvent se ramener à l'intervalle $[0,1]$. Pour $i \in \{0, ..., N\}$, on conserve la même application F_i comme pour le cas $\mathbb{P}_1$ qui envoie bien $[0,1]$ dans I_i.

On appellera toujours l'intervalle $[0,1]$, l'*intervalle de référence* (ou *élément de référence*).

Sur un intervalle donné par $i \in \{1, ..., N-1\}$, I_i, seules φ_{2i}, φ_{2i+1} et φ_{2i+2} sont non nulles. Sur I_0, seules φ_1 et φ_2 sont non nulles. Sur I_N, seules φ_{2N} et φ_{2N+1} sont non nulles. La restriction de toutes ces fonctions sur I_i peuvent de plus être obtenues à partir de trois fonctions élémentaires polynomiales sur $[0,1]$ et de degré ≤ 2. On les note $\hat{\varphi}_0$, $\hat{\varphi}_1$ et $\hat{\varphi}_2$ définies sur l'intervalle $[0,1]$. Ces trois fonctions sont définies par $\hat{\varphi}_0(\xi) = (1-\xi)(1-2\xi)$, $\hat{\varphi}_1(\xi) = 4(1-\xi)\xi$ et $\hat{\varphi}_2(x) = \xi(2\xi - 1)$ pour $\xi \in [0,1]$. Une représentation graphique est donnée en Figure VII.2.

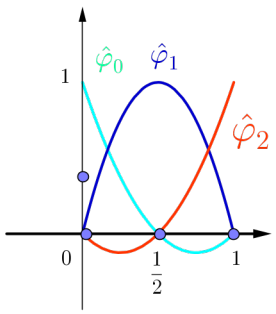

FIGURE VII.2 – Représentation graphique des fonctions de référence $\mathbb{P}_2$.

Et on a $(\varphi_{2i})_{/I_i} = \hat{\varphi}_0 \circ F_i^{-1}$, $(\varphi_{2i+1})_{/I_i} = \hat{\varphi}_1 \circ F_i^{-1}$, $(\varphi_{2i+2})_{/I_i} = \hat{\varphi}_2 \circ F_i^{-1}$.

Remarque VII.3.4 *La forme bilinéaire a est symétrique, donc* $a(\varphi_j, \varphi_i) = a(\varphi_i, \varphi_j)$, $\forall (i,j) \in \{1, ..., 2N+1\}$ *et donc* $\mathcal{A}$ *est symétrique.*

On peut calculer directement les termes de la matrice en utilisant les expressions de chaque fonction de base sur $[\alpha, \beta]$ (*Exercice*). Mais, comme dans le cas $\mathbb{P}_1$, on va priviliégier la méthode de calcul utilisant l'élément de référence. Pour calculer les intégrales qui interviennent dans les coefficients de $\mathcal{A}_h$, on va utiliser l'intervalle de référence et s'y ramener par le changement de variable utilisant F_i.

Pour simplifier la lecture, on commence par précalculer les termes de référence. On a [3] :

$$\int_0^1 (\hat{\varphi}_0'(\xi))^2 d\xi = \frac{14}{6}, \tag{VII.8}$$

$$\int_0^1 (\hat{\varphi}_1'(\xi))^2 d\xi = \frac{16}{3}, \tag{VII.9}$$

$$\int_0^1 (\hat{\varphi}_2'(\xi))^2 d\xi = \frac{14}{6}, \tag{VII.10}$$

$$\int_0^1 \hat{\varphi}_0'(\xi)\hat{\varphi}_1'(\xi) d\xi = -\frac{8}{3}, \tag{VII.11}$$

$$\int_0^1 \hat{\varphi}_0'(\xi)\hat{\varphi}_2'(\xi) d\xi = \frac{1}{3}, \tag{VII.12}$$

$$\int_0^1 \hat{\varphi}_1'(\xi)\hat{\varphi}_2'(\xi) d\xi = -\frac{8}{3}, \tag{VII.13}$$

Calcul des termes.

Tout d'abord, pour $j \in \{1, ..., 2N+1\}$, on est sûr que tous les termes $a(\varphi_j, \varphi_l) = \int_\alpha^\beta \varphi_j'\varphi_l'$ sont nuls si $|j-l| \geq 4$, puisque nous avons vu que le support de φ_j est réduit à $[\tilde{x}_{j-2}, \tilde{x}_{j+2}]$, si j pair, et $[\tilde{x}_{j-1}, \tilde{x}_{j+1}]$, si j impair et que $]\tilde{x}_{j-2}, \tilde{x}_{j+2}[\cap]\tilde{x}_{l-2}, \tilde{x}_{l+2}[= \emptyset$ si $|j-l| \geq 4$. De plus, l'intersection des supports des fonctions φ_{2i} et φ_{2i-3} est vide et l'intersection des supports de φ_{2i+1} et φ_{2i+4} est lui aussi vide [4]. Il n'y a donc à calculer que les termes $(a(\varphi_j, \varphi_j))_{j\in\{1,...,2N+1\}}$ et $(a(\varphi_j, \varphi_{j+1}))_{j\in\{1,...,2N\}}$, $(a(\varphi_j, \varphi_{j-1}))_{i\in\{2,...,2N+1\}}$, $(a(\varphi_j, \varphi_{j+2}))_{j\in\{1,...,2N-1\}}$, $(a(\varphi_j, \varphi_{j-2}))_{j\in\{3,...,2N+1\}}$. Plus précisément, on a ce qui suit.

Si $j \in \{1, ..., 2N+1\}$ est pair, alors il existe $i \in \{1, ..., N\}$, tel que $j = 2i$ et

$$a(\varphi_j, \varphi_j) = a(\varphi_{2i}, \varphi_{2i}) \tag{VII.14}$$

$$= \int_\alpha^\beta (\varphi_{2i}'(x))^2 dx \tag{VII.15}$$

$$= \int_{I_{i-1}} (\varphi_{2i}'(x))^2 dx + \int_{I_i} (\varphi_{2i}'(x))^2 dx. \tag{VII.16}$$

3. Calculs à faire en exercice

4. on utilise que le support des fonctions d'indice impair est plus petit.

Chacun de ces deux termes peut être calculé individuellement. Tout d'abord, on a sur I_i, $\varphi_{2i} = \hat{\varphi}_0 \circ F_i^{-1}$ et donc $\varphi'_{2i}(x) = \hat{\varphi}'_0(F_i^{-1}(x))F_i^{-1'}(x)$. On a donc

$$\int_{I_i} (\varphi'_{2i}(x))^2 dx = \int_{I_i} (\hat{\varphi}'_0(F_i^{-1}(x))F_i^{-1'}(x))^2 dx. \tag{VII.17}$$

On définit un changement de variable en posant $\xi = F_i^{-1}(x)$, ce qui donne $d\xi = F_i^{-1'}(x)dx$ avec $F_i^{-1'}(x) = \frac{1}{\tilde{x}_{2i+2} - \tilde{x}_i} = \frac{1}{h}$.
On trouve alors

$$\int_{I_i} (\varphi'_{2i}(x))^2 dx = \frac{1}{h}\int_0^1 (\hat{\varphi}'_0(\xi))^2 d\xi, \tag{VII.18}$$

$$= \frac{14}{6h}. \tag{VII.19}$$

De même,

$$\int_{I_{i-1}} (\varphi'_{2i}(x))^2 dx = \int_{I_{i-1}} (\hat{\varphi}'_2(F_{i-1}^{-1}(x))F_{i-1}^{-1'}(x))^2 dx = \frac{1}{h}\int_0^1 (\hat{\varphi}'_2(\xi))^2 d\xi = \frac{14}{6h}. \tag{VII.20}$$

Donc

$$\int_\alpha^\beta (\varphi'_{2i}(x))^2 dx = \frac{14}{3h}. \tag{VII.21}$$

Ensuite de la même façon, puisque l'intersection du support de φ_{2i} et φ_{2i-1} est I_{i-1}, on a

$$\int_\alpha^\beta \varphi'_{2i}(x)\varphi'_{2i-1}(x)dx = \int_{I_{i-1}} \varphi'_{2i}(x)\varphi'_{2i-1}(x)dx. \tag{VII.22}$$

Et donc

$$\int_{I_{i-1}} \varphi'_{2i}(x)\varphi'_{2i-1}(x)dx = \int_{I_{i-1}} \hat{\varphi}'_2(F_{i-1}^{-1}(x))\hat{\varphi}'_1(F_{i-1}^{-1}(x))(F_{i-1}^{-1'}(x))^2 dx. \tag{VII.23}$$

Et donc, en utilisant le même changement de variable que précédemment, on a

$$\int_{I_{i-1}} \varphi'_{2i}(x)\varphi'_{2i-1}(x)dx = \frac{1}{h}\int_0^1 \hat{\varphi}'_2(\xi)\hat{\varphi}'_1(\xi)d\xi = -\frac{8}{3h}. \tag{VII.24}$$

De façon analogue, on a

$$\int_\alpha^\beta \varphi'_{2i}(x)\varphi'_{2i+1}(x)dx = \frac{1}{h}\int_0^1 \hat{\varphi}'_0(\xi)\hat{\varphi}'_1(\xi)d\xi = -\frac{8}{3h}. \tag{VII.25}$$

Puis,

$$\int_\alpha^\beta \varphi'_{2i}(x)\varphi'_{2i+2}(x)dx = \frac{1}{h}\int_0^1 \hat{\varphi}'_0(\xi)\hat{\varphi}'_2(\xi)d\xi = \frac{1}{3h}. \tag{VII.26}$$

Enfin

$$\int_\alpha^\beta \varphi'_{2i}(x)\varphi'_{2i-2}(x)dx = \frac{1}{h}\int_0^1 \hat{\varphi}'_2(\xi)\hat{\varphi}'_0(\xi)d\xi = \frac{1}{3h}. \tag{VII.27}$$

Si j est impair, alors il existe $i \in \{1,..,N\}$, telle que $j = 2i+1$. On a alors : $\int_\alpha^\beta \varphi'_{2i+1}(x)\varphi'_{2i}(x)dx$ a déjà été calculé. Il reste encore plusieurs termes à calculer :

$$\int_\alpha^\beta (\varphi'_{2i+1}(x))^2 dx = \frac{1}{h}\int_0^1 (\hat{\varphi}'_1(\xi))^2 d\xi = \frac{16}{3h}, \tag{VII.28}$$

$$\int_\alpha^\beta \varphi'_{2i+1}(x)\varphi'_{2i+2}(x)dx = \frac{1}{h}\int_0^1 \hat{\varphi}'_1(\xi)\hat{\varphi}'_2(\xi)d\xi = -\frac{8}{3h}, \tag{VII.29}$$

Finalement,

$$\mathcal{A}_h = \frac{1}{3h}\begin{pmatrix}
16 & -8 & 0 & & & & & & \\
-8 & 14 & -8 & 1 & & & & & \\
0 & -8 & 16 & -8 & 0 & & & 0 & \\
 & 1 & -8 & 14 & -8 & 1 & & & \\
 & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\
 & & & \ddots & \ddots & \ddots & \ddots & \ddots & \\
 & & 0 & & \ddots & \ddots & \ddots & \ddots & 0 \\
 & & & & & 1 & -8 & 14 & -8 \\
 & & & & & & 0 & -8 & 16
\end{pmatrix}$$

Remarque VII.3.5 *La matrice est pentadiagonale et symétrique.*

On peut faire des remarques similaires concernant la numérotation, le calcul du second membre et l'assemblage, ces points étant légèrement plus techniques.

Une fois la matrice construite, il "suffit" de résoudre le système linéaire. On trouve alors en sortie le vecteur U_h contenant les coordonnées du vecteur dans la base $(\varphi_i)_{i\in\{1,...,2N+1\}}$ qui correspondent aux valeurs de l'approximation $u_h \in \mathcal{V}_h$ au points $(\tilde{x}_i)_{i\in\{1,...,2N+1\}}$.

Chapitre VIII

Résultats de convergence des deux méthodes et extension.

VIII.1 Résultats de convergence de la méthode

VIII.1.1 Stratégie globale

On rappelle les deux problèmes variationnels qui nous intéressent :

- Le problème continu $(\mathcal{F})$: *Trouver* $u \in \mathcal{V}$ *tel que pour tout* $v \in \mathcal{V}$,

$$a(u, v) = l(v).$$

- Le problème discret $(\mathcal{F}_h)$: *Trouver* $u_h \in \mathcal{V}_h$ *tel que pour tout* $v_h \in \mathcal{V}_h$,

$$a(u_h, v_h) = l(v_h).$$

La question naturelle qui se pose est de savoir si la méthode des éléments finis *converge*. Autrement dit, la question qui se pose est : est-ce que l'approximation $u_h \in \mathcal{V}_h$ trouvée est une bonne approximation de $u \in \mathcal{V}$. Plus précisément on se demande si $u_h \underset{h\to 0}{\longrightarrow} u$ (en un sens à préciser). Autrement dit : *est-ce que si on met de "plus en plus" de points dans la subdivision (i.e. on raffine le maillage), la solution approchée se "rapproche" de la solution exacte ?*

Tout d'abord, on se met dans le cas où toutes les formulations variationnelles (resp. continues et discrètes) admettent une unique solution (dans resp. $\mathcal{V}$ et $\mathcal{V}_h$). **On fait donc l'hypothèse** que a est continue et coercive sur $\mathcal{V} \times \mathcal{V}$ et que l est continue sur $\mathcal{V}$. Pour l'existence et unicité d'une solution au problème variationnel continu, on peut utiliser Lax-Milgram. L'existence et unicité d'une solution au problème variationnel discret peut s'obtenir de la même façon, mais on peut aussi le montrer directement en utilisant la proposition V.3.1.

On aimerait maintenant pouvoir montrer que la solution $u_h \in \mathcal{V}_h$ converge vers u lorsque le paramètre h tend vers 0 dans une norme à préciser. En d'autres termes, plus on raffine le maillage (i.e. plus on met de points dans la subdivision), plus la solution calculée $u_h \in \mathcal{V}_h$ est "proche" (en un sens à préciser) de la solution $u \in \mathcal{V}$. Et on aimerait également pouvoir préciser la vitesse de convergence.

Dans les deux méthodes d'éléments finis présentées dans les deux chapitres précédents ($\mathbb{P}_1$ et $\mathbb{P}_2$), on a ce genre de résultats de convergence.

Les résultats de convergence sont obtenus en partant du Lemme de Céa (Lemme V.3.4). Rappelons le Lemme de Céa V.3.4.

Lemma VIII.1.1 ***Lemme de Céa.*** *On suppose que $(\mathcal{V}, \langle\cdot,\cdot\rangle)$ est un espace de Hilbert réel, $\mathcal{V}_h$ un sous espace de dimension finie de $\mathcal{V}$. Soit $a : \mathcal{V}\times\mathcal{V} \to \mathbb{R}$ une forme bilinéaire continue (avec constante de continuité $M > 0$) et coercive (avec constante de corecivité $\alpha > 0$) sur $\mathcal{V}\times\mathcal{V}$ et $l : \mathcal{V} \to \mathbb{R}$ une forme linéaire continue sur $\mathcal{V}$. On considère $u \in \mathcal{V}$ la solution de $(\mathcal{F})$ et u_h la solution de $(\mathcal{F}_h)$. On a alors l'estimation suivante*

$$\|u - u_h\| \leq \frac{M}{\alpha} \inf_{v_h\in\mathcal{V}_h} \|u - v_h\|. \tag{VIII.1}$$

Pour montrer que $\|u-u_h\| \underset{h\to 0}{\longrightarrow} 0$, en utilisant le Lemme de Céa, on voit qu'il suffit de montrer que

$$\inf_{w_h\in\mathcal{V}_h} \|u - w_h\| \to 0,$$

lorsque $h \to 0$. Pour cela, il suffit de trouver un seul élément $\tilde{w}_h$ de $\mathcal{V}_h$ tel que

$$\|u - \tilde{w}_h\| \to 0,$$

lorsque $h \to 0$. En effet, dans ce cas, puisque $\inf_{w_h\in\mathcal{V}_h} \|u - w_h\| \leq \|u - \tilde{w}_h\|$, on a le résultat voulu.

VIII.1.2 Dans le cas des éléments finis de Lagrange $\mathbb{P}_1$ sur l'équation modèle

On peut montrer que dans le cas de l'équation modèle de Poisson les hypothèses pour appliquer Lax-Milgram sont vérifiées [1]. Donnons une intuition d'un tel $\tilde{w}_h$ dans le cas d'une approximation éléments finis $\mathbb{P}_1$ et on se place dans le cadre de l'équation modèle. On reprend les notations du chapitre VI. Pour $u \in \mathcal{V}$ donnée [2], on introduit son *interpolation* dans l'espace $\mathcal{V}_h$. C'est l'unique fonction de $\mathcal{V}_h$ qui prend les mêmes valeurs que u aux noeuds associés aux degrés de liberté. Cette fonction s'appelle l'*interpolée* de u dans $\mathcal{V}_h$ et on la note $\mathcal{I}_h(u)$. Comme elle appartient à $\mathcal{V}_h$, on peut l'exprimer dans la base $(\varphi_i)_{i\in\{1,\dots,N\}}$. Elle s'écrit pour tout $u \in \mathcal{V}$, $x \in [\alpha, \beta]$:

$$I_h(u)(x) = \sum_{i=1}^{N} u(x_i)\varphi_i(x). \tag{VIII.2}$$

En d'autres termes :

$$I_h : \mathcal{V} \to \mathcal{V}_h, \tag{VIII.3}$$

$$u \mapsto \sum_{i=1}^{N} u(x_i)\varphi_i. \tag{VIII.4}$$

Une illustration est donnée en figure VIII.1.

1. **Pour les MPA.** Si on se met dans le cadre de l'espace $H_0^1([\alpha,\beta])$, la continuité de a se montre en utilisant la norme de cet espace. Ensuite, pour la coercivité de a, on utilise une inégalité appelée inégalité de Poincaré.

2. dans $\mathcal{V}$, les conditions de bords sont prises en compte.

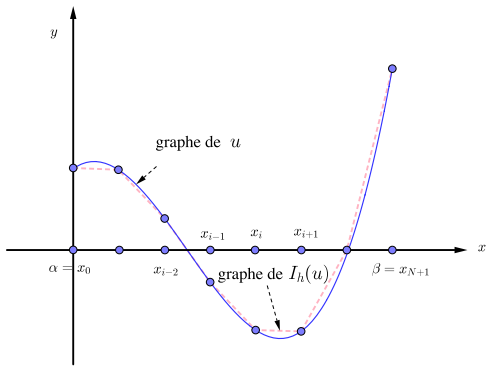

FIGURE VIII.1 – Illustration de l'interpolation de fonction par des fonctions affines par morceaux. On voit qu'entre x_{i-2} et x_{i-1}, la courbe représentative de u et celle de $I_h(u)$ semblent confondues : ce n'est pas le cas, elles sont très proches, mais pas confondues !

On a le résultat suivant (**que l'on admettra**) [3] :

Proposition VIII.1.2 *Soit $v \in \mathcal{V}$ et $I_h(v) \in \mathcal{V}_h$ son interpolée dans $\mathcal{V}_h$. On a*

$$\lim_{h\to 0} \|v - I_h(v)\| = 0. \tag{VIII.5}$$

Si on considère que la fonction v est [4] $\mathcal{C}^2([\alpha,\beta])$, alors il existe $C > 0$ (indépendant de h), tel que pour h suffisamment petit,

$$\|v - I_h(v)\| \leq Ch\|v''\|_{L^2([\alpha,\beta])}. \tag{VIII.6}$$

De cette proposition découle directement la convergence de la méthode en combinant ce résultat au Lemme de Céa comme motivé plus haut. On obtient alors que

$$\|u - u_h\| \to 0, \tag{VIII.7}$$

lorsque $h \to 0$ et sous les hypothèses supplémentaires ($\mathcal{C}^2([\alpha,\beta])$ comme dans le théorème [5]), on a $\exists C > 0$ (indépendante de h) tel que

$$\|u - u_h\| \leq \tilde{C}h\|u''\|_{L^2([\alpha,\beta])}. \tag{VIII.8}$$

On dit alors que la méthode converge à **l'ordre** (au moins) 1. **L'ordre** est donné par la puissance de h dans l'inégalité ci-dessus. Il quantifie la vitesse à laquelle la solution u_h converge vers u dans $\mathcal{V}$.

3. On voit qu'ici pouvoir définir l'interpolée suppose que si $u \in \mathcal{V}$, alors on est capable d'évaluer u en x_i. Ce qui est le cas si u est continue et c'est aussi le cas si $u \in H_0^1([0,1])$ (**pour les MPA**).

4. voir la remarque sur les espace possibles lors de l'obtention de la formulation variationnelle.

5. **Pour les MPA.** En fait, l'espace adapté est là aussi un espace de Sobolev ; $\mathcal{V} = H_0^1([\alpha,\beta])$ et l'hypothèse de régularité supplémentaire est reliée à l'espace $H^2([\alpha,\beta]) = \{v \in L^2(\Omega),\ v' \in H^1([\alpha,\beta])\}$ avec les normes correspondantes.

Remarque VIII.1.3 *On appelle $u - u_h$ l'erreur et $\|u - u_h\|$ l'erreur en norme $\mathcal{V}$ (ou norme $\mathcal{V}$ de l'erreur). Le résultat de convergence est obtenu en utilisant la norme de $\mathcal{V}$ de cette erreur. On peut continuer et prouver un résultat portant sur la norme L^2 qui permet d'écrire (**admis**)*

$$\|u - u_h\|_{L^2([\alpha,\beta])} \leq Ch^2 \|u''\|_{L^2([\alpha,\beta])}. \tag{VIII.9}$$

On dira alors qu'on converge à l'ordre 2 (cf. puissance de h) pour la norme $L^2([\alpha, \beta])$.

On a le même genre de résultat pour une approximation élément finis $\mathbb{P}_2$. On peut montrer qu'on a **convergence à l'ordre 2 en norme $\mathcal{V}$** (sous de bonnes hypothèses de régularité) et 3 **en norme** $L^2([\alpha, \beta])$.

J'invite les personnes intéressées et voulant aller plus loin sur ce cours à lire les quelques pages du livre de A. Quarteroni, R. Sacco, F. Saleri, Méthodes Numériques, Analyse, Algorithmes et Applications, Éditions Springer, mises en ligne sur Moodle, à partir de la section 11.3. ***MPA fortement incités à le lire.***

VIII.2 Extension à la dimension supérieure : cas bidimensionnel

Dans cette section, nous proposons l'extension de la méthode des éléments finis au cas bidimensionnel.

La généralisation de l'équation $-u'' = f$, $u(\alpha) = u(\beta) = 0$ s'écrit comme suit. On se donne Ω un ouvert borné de $\mathbb{R}^2$ (vous pouvez penser à l'exemple qui est dans la Feuille de TD5, dans le dernier exercice). L'équation de Poisson avec conditions aux bords de Dirichlet s'écrit :

$$-\Delta u = f \text{ sur } \Omega, \tag{VIII.10}$$
$$u = 0 \text{ sur } \partial\Omega. \tag{VIII.11}$$

où $\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$ est l'opérateur Laplacien appliqué à la fonction u.

Dans la suite, on considèrera $\Omega =]0, 1[^2$.

On suit les mêmes grandes étapes pour définir une méthode éléments finis (qui sont généralisables au cas où Ω n'est pas le carré $]0, 1[^2$).

(1) L'équation est posée sur le domaine Ω. L'opérateur différentiel est de degré 2. Il y a un second membre : f.

(2) On identifie la formulation variationnelle. Comme dans l'exercice de TD, on obtient

$$a : \mathcal{V} \times \mathcal{V} \to \mathbb{R} \tag{VIII.12}$$
$$(u, v) \mapsto \int_\Omega (\nabla u, \nabla v), \tag{VIII.13}$$

où ∇ est l'opérateur gradient, $(\cdot, \cdot)$ le produit scalaire euclidien, et

$$l : \mathcal{V} \to \mathbb{R} \tag{VIII.14}$$

$$v \mapsto \int_\Omega fv. \tag{VIII.15}$$

Pour ce qui est de l'espace $\mathcal{V}$, on peut penser à prendre les fonctions qui sont $\mathcal{C}^2(\Omega)$ et nulles sur le bord de Ω. Là encore, le cadre correct est le cadre des espaces de Hilbert déjà mentionné en dimension 1. Les espaces qui vont convenir sont les espaces de Sobolev et en particulier ici $H_0^1(\Omega)$ qui est la généralisation de l'espace du même nom en dimension 1 . Dans ces conditions, on a existence et unicité au problème variationnel grâce au théorème de Lax-Milgram et à des propriétés fines de l'espace $\mathcal{V}$ (**admis**).

La formulation variationnelle continue est donc $(\mathcal{F})$: *Trouver $u \in \mathcal{V}$ tel que pour tout $v \in \mathcal{V}$,*

$$a(u, v) = l(v).$$

(3) La prochaine étape est de construire une discrétisation (ou maillage) du domaine. Celle-ci peut-être très générale. Dans les cas les plus standards, on envisage une discrétisation dite cartésienne (formée de carrés ou rectangles, voir un exemple en figure VIII.2) ou une discrétisation triangulaire (formée de triangles, voir illustration en figure VIII.3). Dans le cas du carré $]0, 1[^2$, les deux discrétisations semblent adaptées. Si on imagine un cercle, on imagine que la discrétisation à partir de triangles va être plus flexible pour pouvoir coller "au mieux" et plus facilement à la géométrie du domaine.

Dans ce qui suit, on va privilégier la discrétisation triangulaire. On note $\mathcal{T}_h$ cette discrétisation (au sens où $\mathcal{T}_h$ représente la réunion de tous les triangles). Ici h correspondra à la taille maximale d'un élément (au lieu du pas de la subdivision en dimension 1). Une fois le maillage $\mathcal{T}_h$ construit, on construit l'espace de discrétisation $\mathcal{V}_h \subset \mathcal{V}$ et une base associée. Là aussi le principe consiste en une généralisation des espaces en dimension 1. On peut écrire cette méthode de façon très générale en se laissant la possibilité d'avoir des polynômes de degrés autres que 1 et 2. Ainsi, on peut construire une méthode d'élément finis $\mathbb{P}_k$ (avec $k \in \mathbb{N}^*$) en définissant $\mathcal{V}_h^k$ comme

$$\mathcal{V}_h^k = \left\{ v \in \mathcal{C}^0(\bar{\Omega}),\ (v_h)_{/T} \in \mathbb{P}_k(T),\ \forall T \in \mathcal{T}_h \text{ et } (v_h)_{/\partial\Omega} = 0 \right\}, \tag{VIII.16}$$

où $\mathbb{P}_k(T)$ est l'espace des polynômes de degré $\leq k$ sur T, ce qui s'écrit

$$\mathbb{P}_k(T) = \left\{ \sum_{i,j=0,\, i+j\leq k}^{k} a_{i,j} x^i y^j,\ (x, y) \in T \right\}. \tag{VIII.17}$$

Lorsque $k = 1$, $\mathcal{V}_h^1$ permettra de construire une méthode d'élément finis $\mathbb{P}_1$, si $k = 2$, $\mathcal{V}_h^2$, une méthode d'éléments finis $\mathbb{P}_2$ etc...

Les bases d'éléments finis seront construites de "la même façon" qu'en dimension 1. On aura accès à la dimension de $\mathcal{V}_h^k$, que l'on note N_h^k, et on identifie des degrés de libertés pour la méthode. Ainsi pour $k = 1$, un degré de libertés est associé aux sommets internes (i.e. en enlevant les points du bords) des triangles. Pour $k = 2$, on choisit comme noeuds associés aux degrés de liberté les sommets internes et les milieux des arêtes internes des triangles (i.e. on a enlevés les milieux des arêtes des triangles qui

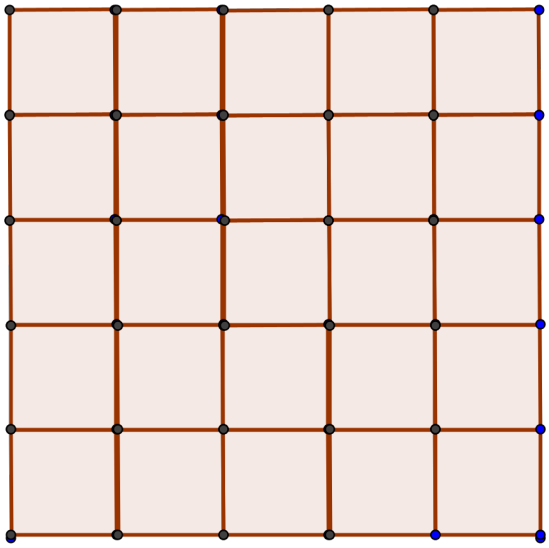

FIGURE VIII.2 – Un maillage cartésien du carré.

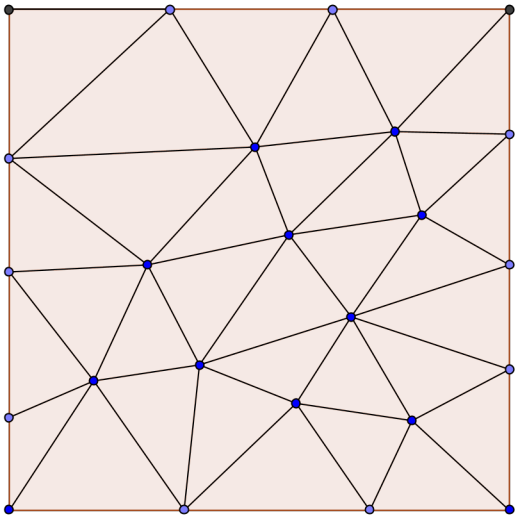

FIGURE VIII.3 – Un maillage triangulaire du carré.

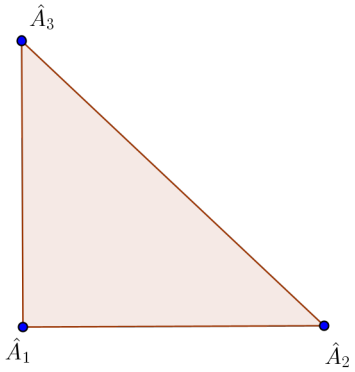

FIGURE VIII.4 – Illustration du triangle de référence $\hat{T}$.

touchent le bord) du maillage. Les fonctions de base seront alors définies par analogie au cas $1D$ de telle sorte qu'il y a ait une fonction de base associée à chaque degré de liberté et que cette fonction de base prenne la valeur 1 en le noeud associé et 0 sur tous les autres noeuds.

(4) La formulation variationnelle discrète est alors donnée par $(\mathcal{F}_h^k)$: *Trouver* $u_h \in \mathcal{V}_h^k$, *telle que pour tout* $v_h \in \mathcal{V}_h^k$,

$$a(u_h, v_h) = l(v_h). \tag{VIII.18}$$

(5) On forme ensuite le système linéaire correspondant. La matrice $\mathcal{A}_h$ est toujours donnée par $(a(\varphi_j, \varphi_i))_{i,j\in\{1,\dots,N_h^k\}^2}$. Comme a est symétrique, $\mathcal{A}_h$ est elle aussi symétrique. Le second membre est donné par $(l(\varphi_i))_{i\in\{1,\dots,N_h^k\}}$.

Cette matrice $\mathcal{A}_h$ est creuse, i.e. elle a beaucoup de zéros (comme pour la dimension 1). Ceci est dû au support des fonctions de base. Par exemple, pour une approximation $\mathbb{P}_1$, le support de chaque fonction de base ne sera localisé que sur les triangles ayant le noeud associé au degré de liberté comme sommet pour la fonction de base associée à ce degré de liberté.

Pour le calcul des termes de la matrice, on va là aussi se ramener à un élément (une cellule) de référence, ici un triangle de référence que l'on notera $\hat{T}$. On considère le triangle rectangle d'arêtes principales de longueur 1, comme illustré en figure VIII.4. Ce triangle a pour sommets $\hat{A}_1 = (0,0)$, $\hat{A}_2 = (1,0)$ et $\hat{A}_3 = (0,1)$.

Là aussi, on utilisera une fonction affine pour chaque $T \in \mathcal{T}_h$, $F_T : \hat{T} \to T$ pour ramener toutes les intégrales à calculer sur l'élément de référence.

On définira alors également des fonctions de base élémentaires sur le triangle de référence. Par

exemple, pour une méthode $\mathbb{P}_1$, il y en a 3, chacune associée à un sommet. Ces m fonctions de base élémentaires seront utilisées pour simplifier le calcul des termes de la matrice. En utilisant des formules de changements de variable

Comme pour la dimension 1, les remarques concernant la numérotation, le calcul du second membre se généralisent et restent vraies.

Enfin, tout ce qui vient d'être fait ici, se généralise au cas tridimensionnel (i.e. $\Omega \subset \mathbb{R}^3$) qui est souvent le cas pertinent en physique.

Les principales difficultés inhérentes aux méthode éléments finis sont :

(a) Le fait que pour chaque domaine, il faut être capable de construire un maillage. Pour cela, on a recourt à des *mailleurs*. Pour les curieux, vous pouvez aller voir le site du logiciel gmsh [6].

(b) Il faut être capable de construire les tableaux de connectivité (lien entre les numérotations locales à une cellule donnée et la numérotation globale d'un degré de liberté).

(c) Il faut pouvoir assembler la matrice $\mathcal{A}_h$.

(d) Enfin il faut être capable de résoudre le système linéaire associé. Tout en sachant que plus on raffine (i.e. plus on met d'éléments dans le maillage), plus la matrice est mal conditionnée et le système linéaire dur à inverser et de taille de plus en plus grande.

Pour les points (b), (c), (d), il existe maintenant beaucoup de logiciels de résolution efficaces, qui vont faire tout ce travail pour vous dans le cas des éléments finis standards. Un de ceux-ci est FreeFem++ [7].

Enfin, on dispose également de résultats de convergence en combinant là encore le Lemme de Céa et un résultat d'interpolation. On a la propriété suivante **que l'on admettra**.

Proposition VIII.2.1 *Soient $k \in \mathbb{N}^*$, $u \in \mathcal{V}$ la solution du problème variationnel continu $(\mathcal{F})$, et $u_h \in \mathcal{V}_h^k$ solution du problème variationnel discret $(\mathcal{F}_h^k)$. On suppose que $u \in W = \mathcal{C}^s(\Omega)$ (avec $s \geq 2$)* [8]*. On a alors l'estimation d'erreur qui suit. Il existe une constante $C > 0$ indépendante de h, et telle que*

$$\|u - u_h\| \leq C h^l \|u\|_W, \tag{VIII.19}$$

où $l = \min(k, s-1)$ et $W = \mathcal{C}^{l+1}(\Omega)$ [9]*. Sous les mêmes hypothèses, on peut aussi montrer que*

$$\|u - u_h\| \leq C h^{l+1} \|u\|_W. \tag{VIII.20}$$

On a donc convergence de la méthode à **l'ordre l en norme $\mathcal{V}$** et à **l'ordre $l+1$ en norme** $L^2(\Omega)$.

6. http ://gmsh.info/
7. https ://freefem.org/
8. **Pour les MPA.** On considère en fait $u \in W = H^s(\Omega)$.
9. **Pour les MPA.** On prend $\tilde{W} = H^{l+1}(\Omega)$

Annexe A

Quelques rappels de calcul différentiel

A.1 Cas de la dimension 1 ($d = 1$)

A.1.1 Rappel des formules de Taylor

On dispose des formules de Taylor.

Proposition A.1.1 *Formule de Taylor-Lagrange. Soit $f : [a, b] \to \mathbb{R}$, une application de classe $\mathcal{C}^n$, avec $n \in \mathbb{N}^*$ donné et telle que $f^{(n+1)}$ existe sur $]a, b[$. Alors*

$$\exists c \in]a, b[,\ f(b) = f(a) + (b-a)f'(a) + ... + \frac{(b-a)^n}{n!} f^{(n)}(a) + \frac{(b-a)^{n+1}}{(n+1)!} f^{(n+1)}(c).$$

Proposition A.1.2 *Formule de Taylor-Young. Soit $f : [a, b] \to \mathbb{R}$, une application de classe $\mathcal{C}^n$, avec $n \in \mathbb{N}^*$ donné. Alors lorsque $h \to 0$, on a*

$$f(a + h) = f(a) + hf'(a) + ... + \frac{h^n}{n!} f^{(n)}(a) + o(h^n).$$

A.2 Rappels de résultats classiques en dimension supérieure

On peut généraliser tout ce qu'on vient de dire à la dimension supérieure, i.e. $d \geq 1$. On notera $\langle \cdot, \cdot \rangle$ le produit scalaire euclidien sur $\mathbb{R}^d$.

Attention ! On manipule alors des matrices et des vecteurs, il faut être prudents !

La visualisation, relativement aisée en dimension 1 (on trace le graphe d'une fonction, quand on le peut) devient beaucoup plus difficile et moins intuitive en dimension supérieure.

A.2.1 Rappel de la notion de différentiabilité et des formules de Taylor à l'ordre 2

On rappelle que l'on a une notion de **différentiabilité** d'une fonction de $U \subset \mathbb{R}^d$ dans $\mathbb{R}$ avec U ouvert de $\mathbb{R}^d$.

Definition A.2.1 *Soit U un ouvert de $\mathbb{R}^d$ et $a \in U$. Une application $F : U \to \mathbb{R}$ est dite différentiable en a, s'il existe une application linéaire et continue de $\mathbb{R}^d$ dans $\mathbb{R}$ (espace $\mathcal{L}_c(\mathbb{R}^d, \mathbb{R})$), L, telle que lorsque $h \to 0$ (dans $\mathbb{R}^d$),*

$$F(a+h) = F(a) + L(h) + o(\|h\|).$$

Si L existe, on l'appelle la différentielle de F en a et on la note $DF(a)$.

Si F est *différentiable en tout point de U*, on dit que F est *différentiable sur U* et l'application $DF : U \to \mathcal{L}_c(\mathbb{R}^d, \mathbb{R})$ est appelée application différentielle de F. Si DF est continue, on dit que F est de classe $\mathcal{C}^1$ (et ainsi de suite pour $\mathcal{C}^p$, $p \in \mathbb{N}$).

Gradient et Hessienne. Soit U un ouvert de $\mathbb{R}^d$, $a \in U$ et $F : U \to \mathbb{R}$ différentiable en a. On a aussi une notion de *dérivée partielle* qui est une dérivée directionnelle par rapport à chaque direction canonique $\dfrac{\partial}{\partial x_i}$, $i \in \{1, .., d\}$ (qu'on ne rappellera pas dans ce cours) et on a

$$DF(a).h = \sum_{i=1}^{d} \frac{\partial F}{\partial x_i}(a) h_i.$$

Une fonction $F : U \to \mathbb{R}$ est dite de *classe $\mathcal{C}^p$*, si toutes ses dérivées partielles jusqu'à l'ordre p existent et sont continues sur U.

On appelle **gradient** de F en un point a de $\mathbb{R}^d$, le vecteur de $\mathbb{R}^d$ dont les coordonnées dans la base canonique de $\mathbb{R}^d$ sont données par les dérivée partielles, autrement dit :

$$\nabla F(a) = \begin{pmatrix} \dfrac{\partial F}{\partial x_1}(a) \\ \dfrac{\partial F}{\partial x_2}(a) \\ \vdots \\ \dfrac{\partial F}{\partial x_d}(a) \end{pmatrix}$$

On a donc en particulier pour tout $h \in \mathbb{R}^d$:

$$DF(a).h = \langle \nabla F(a), h \rangle = {}^t \nabla F(a) h.$$

Si F est $\mathcal{C}^2$ en $a \in U$, on peut définir la différentielle seconde de F en a, $D^2F(a)$. Dans ce cas, la matrice $A = \left(\dfrac{\partial^2 F}{\partial x_i \partial x_j}(a) \right)_{(i,j)\in\{1,..,d\}^2} \in M_d(\mathbb{R})$ est appelée **matrice Hessienne** de F en a, on pourra la noter $Hess_a(F)$. On remarque que comme $\dfrac{\partial^2 F}{\partial x_i \partial x_j}(a) = \dfrac{\partial^2 F}{\partial x_j \partial x_i}(a)$ (théorème de Schwartz), la matrice est symétrique.

De plus, pour tout $(h, l) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$D^2F(a).(h,l) = \sum_{i,j} (Hess_a(F))_{i,j} h_i l_j = \langle Hess_a(F) h, l \rangle.$$

On dispose aussi de formules de Taylor.

Proposition A.2.2 *Soit $F : U \subset \mathbb{R}^d \to \mathbb{R}$ (où U est **un ouvert** de $\mathbb{R}^d$), une application de classe $\mathcal{C}^2$. Soit $u \in \mathbb{R}^d$ et $h \in \mathbb{R}^d$ tel que $\{u + th, t \in [0, 1]\} \subset U$. Alors :*

— ***Formule de Taylor-Lagrange à l'ordre 2.*** *Il existe $\theta \in]0, 1[$ tel que*

$$F(u + h) = F(u) + DF(u).h + \frac{1}{2}D^2F(u + \theta h)(h, h).$$

— ***Formule de Taylor avec reste intégral.*** *On a*

$$F(u + h) = F(u) + DF(u).h + \int_0^1 (1 - t)D^2F(U + th).(h, h).$$

— ***Formule de Taylor-Young.*** *Lorsque $h \to 0$,*

$$F(u + h) = F(u) + DF(u).h + \frac{1}{2}D^2F(u) \cdot (h, h) + o(\|h\|^2).$$

Remarque A.2.3 *Si vous n'êtes pas à l'aise avec les notations DF et D^2F, transcrivez tous les résutalts avec le gradient et la matrice Hessienne, avec les correspondances données plus haut.*

Annexe B

Conditions nécessaires et suffisantes d'optimalité.

On commence par le cas simple de la dimension 1 et on généralisera au cas de la dimension supérieure ensuite. On va voir que les caractérisations que l'on a valent généralement pour un minimum local.

B.1 Cas de la dimension 1 ($d = 1$)

B.1.1 Condition nécessaire d'optimalité (rappels)

On se donne $I = [a, b]$ avec $(a, b) \in \mathbb{R}$ et $a < b$, avec éventuellement $I = [a, +\infty[$, $I =]-\infty, b]$ ou $I = \mathbb{R}$. On a le résultat suivant

Théorème B.1.1 *Si $\mathcal{J} : I \to \mathbb{R}$ est $\mathcal{C}^1$ sur I et $x^* \in]a, b[= \mathring{I}$ est un minimum **local** de $\mathcal{J}$, alors $\mathcal{J}'(x^*) = 0$. Si de plus $\mathcal{J}$ est deux fois dérivable, alors $\mathcal{J}''(x^*) \geq 0$.*

Preuve : Comme x^* est un minimum local de $\mathcal{J}$, alors il existe $\delta > 0$ tel que si $x \in I$ et $|x - x^*| < \delta$, alors $\mathcal{J}(x) - \mathcal{J}(x^*) \geq 0$. Autrement dit, il existe $\delta > 0$ tel que si $x \in]x^* - \delta, x^* + \delta[\cap [a, b]$, alors $\mathcal{J}(x) - \mathcal{J}(x^*) \geq 0$. Quitte à choisir δ suffisamment petit, on peut supposer que $]x^* - \delta, x^* + \delta[\cap [a, b] =]x^* - \delta, x^* + \delta[$ (i.e. $]x^* - \delta, x^* + \delta[\subset [a, b]$), puisque $x^* \in]a, b[$.

⚠ *Si x^* n'était pas dans l'ouvert $]a, b[$, on ne pourrait pas faire ce que l'on vient de faire !*

On peut donc écrire pour tout les $0 < t < \delta$,

$$\frac{\mathcal{J}(x^* + t) - \mathcal{J}(x^*)}{t} \geq 0.$$

En passant à la limite quand $t \to 0^+$, on a $\mathcal{J}'(x^*) \geq 0$ (puisque $\mathcal{J}$ est dérivable en x^*).

De même pour tout $0 < t < \delta$,

$$\frac{\mathcal{J}(x^* - t) - \mathcal{J}(x^*)}{-t} \leq 0.$$

En passant à la limite quand $t \to 0^+$, on a $\mathcal{J}'(x^*) \leq 0$ (puisque $\mathcal{J}$ est dérivable en x^*).

Ce qui ne nous laisse donc que la seule possibilité $\mathcal{J}'(x^*) = 0$.

Pour la seconde partie de la preuve, procédons l'absurde. Supposons que $\mathcal{J}''(x^*) < 0$. Alors dans ce cas, pour $0 < t < \delta$, on a par une formule de Taylor-Young à l'ordre 2, et puisque $\mathcal{J}'(x^*) = 0$,

$$\mathcal{J}(x^* + t) = \mathcal{J}(x^*) + \frac{1}{2}t^2\mathcal{J}''(x^*) + o(t^2).$$

Donc pour tout $\varepsilon > 0$, il existe $\eta > 0$ tel que si $|t| < \eta$,

$$|\mathcal{J}(x^* + t) - \mathcal{J}(x^*) - \frac{1}{2}t^2\mathcal{J}''(x^*)| \leq \varepsilon t^2.$$

Ce qui donne pour $0 < t < \min(\delta, \eta)$,

$$\mathcal{J}(x^* + t) - \mathcal{J}(x^*) \leq \frac{1}{2}t^2\mathcal{J}''(x^*) + \varepsilon t^2.$$

Choisissons $\varepsilon = -\mathcal{J}''(x^*)/4 > 0$. On a alors

$$\mathcal{J}(x^* + t) - \mathcal{J}(x^*) \leq \frac{1}{4}t^2\mathcal{J}''(x^*) < 0.$$

Contradiction avec la définition de x^*. □

⚠**Le fait que x^* soit dans l'ouvert $\mathring{I} =]a, b[$ joue un rôle essentiel.**

Remarque B.1.2 ⚠*La condition donnée dans le théorème précédent n'est qu'une condition* ***nécessaire****. En effet, on peut avoir $\mathcal{J}'(x^*) = 0$, sans pour autant avoir ni de minimum, ni de maximum (Pensez simplement à la fonction $x \mapsto x^3$ sur $\mathbb{R}$).*

Un point x qui vérifie la condition $\mathcal{J}'(x) = 0$ est appelé un ***point critique de la fonction $\mathcal{J}$***.

B.1.2 Condition suffisante d'optimalité (rappels)

On cherche maintenant à avoir aussi une condition **suffisante** d'optimalité.

Théorème B.1.3 *Soit $\mathcal{J} : [a, b] \to \mathbb{R}$ $\mathcal{C}^2$ sur $[a, b]$ et $x^* \in]a, b[$ un point critique de $\mathcal{J}$ (i.e. $\mathcal{J}'(x^*) = 0$). Si $\mathcal{J}''(x^*) > 0$, alors x^* est un minimum local de $\mathcal{J}$ sur $\mathbb{R}$.*

Preuve : Supposons que x^* est un point critique de $\mathcal{J}$. On a donc par une formule de Taylor-Young :

$$\mathcal{J}(x) = \mathcal{J}(x^*) + \frac{1}{2}\mathcal{J}''(x^*)(x - x^*)^2 + o((x - x^*)^2).$$

On sait donc que pour tout $\varepsilon > 0$, il existe $\eta > 0$ tel que pour tout $x \in]x^* - \eta, x^* + \eta[\cap I$, $|\mathcal{J}(x) - \mathcal{J}(x^*) - \frac{1}{2}\mathcal{J}''(x^*)(x - x^*)^2| \leq \varepsilon(x - x^*)^2$.

Ce qui donne

$$\mathcal{J}(x) - \mathcal{J}(x^*) - \frac{1}{2}\mathcal{J}''(x^*)(x-x^*)^2 \geq -\varepsilon(x-x^*)^2.$$

Ainsi,

$$\mathcal{J}(x) - \mathcal{J}(x^*) \geq (\frac{1}{2}\mathcal{J}''(x^*) - \varepsilon)(x-x^*)^2.$$

Fixons $\varepsilon = \frac{\mathcal{J}''(x^*)}{4} > 0$. On a ainsi

$$\frac{1}{2}\mathcal{J}''(x^*) - \varepsilon > 0,$$

et donc pour tout $x \in]x^* - \eta, x^* + \eta[\cap I$, $\mathcal{J}(x) - \mathcal{J}(x^*) \geq 0$.

On en conclut donc que x^* est un minimum local de $\mathcal{J}$.

□

B.2 Conditions nécessaires et/ou suffisantes d'optimalité en dimension supérieure.

On se place dans le cas où $V = \mathbb{R}^d$, avec $d \geq 1$. On a accès là aussi à des conditions nécessaires ou suffisantes d'optimalité.
On suppose que $U \subset \mathbb{R}^d$ est un ouvert de $\mathbb{R}^d$. Commençons par la condition nécessaire.

Proposition B.2.1 *Si $F : U \subset \mathbb{R}^d \to \mathbb{R}$ admet un minimum local en un point a de U et si F est différentiable en a, alors $DF(a) = 0$ (autrement dit $\frac{\partial F}{\partial x_i}(a) \equiv 0$, pour tout $i \in \{1, .., d\}$ ou encore $\nabla F(a) \equiv 0$).*

Remarque B.2.2 *Cette équation vérifiée au point de minimum, s'appelle aussi équation d'Euler.*

Preuve : La preuve généralise le cas de la dimension 1.

Si F admet un minimum local en $a \in U$, alors $\exists \eta > 0$ tel que $B(a, \eta) \subset U$ et $\forall x \in B(a, \eta)$, $F(a) \leq F(x)$.

⚠*Là aussi on a un besoin* ***crucial*** *que U soit ouvert pour pouvoir y insérer une boule ouverte de rayon η.*

Soit $z \in \mathbb{R}^d$, $z \neq 0$ quelconque fixé et t suffisamment petit tel que $a + tz \in B(a, \eta)$ (il suffit de choisir t tel que $|t|\|z\| < \eta$, i.e. $|t| \leq \frac{\eta}{\|z\|}$.)

On a donc $F(a) \leq F(a + tz)$ pour tout $t \in \mathbb{R}$ tel que $|t| < \frac{\eta}{\|z\|}$.

Puis comme F est différentiable en a, lorsque $t \to 0$,

$$F(a + tz) = F(a) + tDF(a) \cdot z + o(t).$$

Ainsi

$$F(a) \leq F(a) + tDF(a) \cdot z + o(t).$$

Donc pour $t \in]0, \frac{\eta}{\|z\|}[$, on obtient $DF(a) \cdot z + o(1) \geq 0$.

En passant à la limite lorsque $t \to 0^+$, on obtient :

$$DF(a) \cdot z \geq 0.$$

Si on choisit $t \in]-\frac{\eta}{\|z\|}, 0[$, on obtient $DF(a) \cdot z + o(1) \leq 0$. Et donc $DF(a) \cdot z \leq 0$.

En conclusion, on a $DF(a) \cdot z = 0$. Cette égalité est aussi vraie si $z = 0$. On a donc le résultat voulu. □

On continue avec un analogue du résultat en dimension 1 qui porte sur la différentielle seconde et qui contient une condition suffisante d'optimalité si on a un point critique.

Théorème B.2.3 *Soit $F : U \subset \mathbb{R}^d \to \mathbb{R}$ une fonction de classe $\mathcal{C}^2$ et supposons qu'il existe $a \in U$, tel que $DF(a) = 0$. Alors :*

(i) Si F admet un minimum local en a, $Hess_a(F)$ est une matrice symétrique positive (condition nécessaire).

(ii) Si

$$Hess_a(F)$$

est une matrice symétrique définie positive, alors F admet un minimum local en a (condition suffisante).

Preuve : (i) Comme a est un minimum local de F sur U, on sait que pour tout $z \in \mathbb{R}^d$, $z \neq 0$, il existe $\eta > 0$, tel que $\forall |t| < \frac{\eta}{\|z\|}$, on ait

$$F(a) \leq F(a + tz) \tag{B.1}$$

(voir la preuve précédente). De plus, comme F est $\mathcal{C}^2$ sur U, on a lorsque $t \to 0$,

$$F(a + tz) = F(a) + tDF(a) \cdot z + \frac{t^2}{2} D^2F(a) \cdot (z, z) + o(t^2\|z\|^2).$$

On sait de plus que $DF(a) \equiv 0$. On en déduit que $F(a + tz) - F(a) = \frac{t^2}{2} D^2F(a).(z, z) + o(t^2)$. Ainsi lorsque $t \to 0$, $\frac{t^2}{2} D^2F(a).(z, z) + o(t^2) \geq 0$. En divisant cette inégalité par t^2 pour $t \neq 0$ puis en faisant tendre t vers 0, on aboutit à

$$\underbrace{D^2F(a).(z, z)}_{\langle Hess_a(F)z, z\rangle} \geq 0. \tag{B.2}$$

Cela signifie exactement que la matrice Hessienne en a est positive.

(ii) On suppose que $\forall z \in \mathbb{R}^d$, $z \neq 0$, $D^2F(a) \cdot z = \underbrace{{}^t z Hess_a(F) z}_{\langle z, Hess_a(F)z\rangle} > 0$.

Comme F est $\mathcal{C}^2$, on sait que pour tout $a \in U$, $Q : h \mapsto D^2F(a) \cdot (h, h) = \langle h, Hess_a(F)h\rangle$ est une forme quadratique continue sur $\mathbb{R}^d$. En particulier, Q est continue sur la sphère unité qui est compacte.

On en déduit donc que Q y est bornée et atteint ses bornes. Notons $\alpha = \min_{h\in\mathbb{R}^d, \|h\|=1} Q(h) > 0$. Par formule de Taylor-Lagrange et comme par hypothèse $DF(a) \equiv 0$, on a $\forall h \in \mathbb{R}^d$,

$$\begin{aligned} F(a+h) - F(a) &= \frac{1}{2}Q(h) + o(\|h\|^2), && \text{(B.3)} \\ &= \frac{\|h\|^2}{2}\left[Q(\frac{h}{\|h\|}) + o(1)\right] && \text{(B.4)} \\ &\geq \frac{\|h\|^2}{2}(\alpha + o(1)) && \text{(B.5)} \end{aligned}$$

De plus, on sait qu'il existe $\eta > 0$, tel que $\alpha + o(1) > 0$ pour $\|h\| < \eta$. On en déduit donc qu'il existe $\eta > 0$ tel que $F(a+h) - F(a) \geq 0$ et donc F admet un minimum local. □

⚠Si la matrice Hessienne n'est que positive (et pas **définie** positive) au point critique, on n'est pas assuré d'avoir un minimum (exemple classique de x^3 en dimension 1).

Bibliographie

[1] Grégoire Allaire. *Analyse Numérique et Optimisation*. École polytechnique edition.