Optimisation et calcul des variations

Olivier Lafitte[1][2]

[1]Institut Galilée, Université de Paris XIII
[2]Commissariat à l'Energie Atomique, Centre d'études de Saclay, lafitte@cea.fr

Contents

Chapter 1

Introduction et exemples

1.1 Introduction

Le but de ce cours est d'introduire quelques unes des méthodes de la théorie de l'optimisation. La méthode employée dans ce cours consiste essentiellement à présenter une suite (non exhaustive) d'exemple simples issu en majeure partie de la physique et de l'économie pour mettre en valeur une question que l'on se pose dans le cadre de l'optimisation: trouver la meilleure quantité ou le meilleur choix pour un problème lié à la physique ou à l'économie. Ce cours présentera peu de résultats (les théorèmes principaux sont peu nombreux). Nous avons essayé de traiter explicitement ici des exemples modèles simples, qui peuvent nous permettre d'introduire des notions et de pouvoir les généraliser.

Les théories liées à l'optimisation sont très variées. On rencontre par exemple (et cela est le plus courant) des problèmes de minimisation sons contraintes, des résolutions d'équations aux dérivées partielles sous forme variationnelle, des problèmes de contrôle, des problèmes de commande. Elles ont en commun la minimisation d'un **critère**, c'est-à-dire d'une fonction chargée de mesurer le **coût** d'un problème, en fonction de variables dites d'état (caractérisant la position d'une particule par exemple) et de variables dites de commande (qui modélisent les paramètres par lesquels on peut agir sur un système). Nous évoquerons ainsi dans le cours la notion de commande optimale, dans les cas où, à partir de variables d'état x et de commandes u, on souhaite soit minimiser un critère, soit atteindre un état fixe.

Un des atouts de l'optimisation est la facilité d'obtention d'algorithmes numériques qui convergent, et nous en aborderons certains: algorithmes d'optimisation sans contrainte, comme un algorithme où on recherche un optimum sur N variables en résolvant, à chaque étape, N algorithmes d'optimisation sur chaque variable, des algorithmes dit de gradient (à pas fixe ou à pas optimal, c'est à dire une généralisation de la méthode de Newton de recherche de zéros), des algorithmes de minimisation avec contraintes, l'algorithme d'Uzawa.

Pour l'instant, nous allons donner une liste non exhaustive d'exemples, provenant des références [2], [3], [1]. Certains pourront être résolus dans cette introduction sans utiliser de théorèmes nouveaux, d'autres non, et nous voulons, dans la suite de ce cours, pouvoir résoudre les problèmes abordés ici.

On peut, très sommairement, diviser les résultats en conditions nécessaires et en conditions nécessaires et suffisantes d'optimalité. Par exemple, x^2 est minimum en $x = 0$, où sa dérivée s'annule, mais la dérivée de $1 - x^2$ est dans le même cas, alors que

$1 - x^2$ est maximum en $x = 0$. La condition "la dérivée s'annule" est une condition nécessaire de minimum, mais n'est pas une condition suffisante.

1.2 Exemples

1. Résolution d'un système matriciel.

   Soit A une matrice symétrique $N \times N$ définie positive et b un vecteur de $\mathbb{R}^N$. La solution du système linéaire $Ax = b$ est donnée par le point de minimum suivant

   $$\inf_{x \in \mathbf{R}^N} \frac{1}{2}(Ax, x) - (b, x)$$

   Preuve On désigne par x_0 la solution de $Ax = b$. On vérifie alors que

   $$\frac{1}{2}(A(x - x_0), x - x_0) = \frac{1}{2}(Ax, x) - \frac{1}{2}(b, x) - \frac{1}{2}(Ax, x_0) + \frac{1}{2}(b, x_0).$$

   Comme $(Ax, x_0) = (x, {}^tAx_0) = (x, Ax_0) = (x, b)$ car A est symétrique

   $$\frac{1}{2}(Ax, x) - (b, x) = -\frac{1}{2}(b, x_0) + \frac{1}{2}(A(x - x_0), x - x_0).$$

   On diagonalise A qui est symétrique définie positive, on écrit $x = x_0 + \sum_i y_i e_i$, où les e_i sont les vecteurs orthonormés qui diagonalisent A, alors

   $$\frac{1}{2}(Ax, x) - (b, x) = -\frac{1}{2}(b, x_0) + \frac{1}{2}\sum_{i=1}^{i=N} \lambda_i y_i^2.$$

   L'expression ci-dessus est minimum lorsque tous les y_i sont nuls, car tous les λ_i sont strictement positifs, donc lorsque $x = x_0$. Le résultat est démontré.

   Je vais décrire sommairement un algorithme dans ce cas: l'algorithme qui consiste à minimiser sur chaque coordonnée. On vérifie que $(A(1, 0...0), (1, 0...0)) = a_{11}$ donc $a_{11} > 0$ (matrice définie positive). Ainsi le minimum, $x_2, ..x_n$ étant fixés, de la fonction quadratique en x_1 est obtenu pour $a_{11}x_1 + \sum_{i=2}^{i=N} a_{i1}x_i - b_1 = 0$, et sa valeur est

   $$f(x_2, ..x_n) = \frac{1}{2}\sum_{i,j \geq 2} a_{ij}x_i x_j - \sum_{i \geq 2} b_j x_j - \frac{1}{2a_{11}}(b_1 - \sum_{j \geq 2} a_{1j}x_j)^2.$$

   Il s'agit à nouveau d'une forme quadratique que l'on peut minimiser en x_2. On itère le procédé.

2. Soit f une application de $\mathbb{R}^M$ dans $\mathbb{R}^N$. On appelle solution de l'équation $f(x) = 0$ une solution du problème

   $$\inf_{x \in \mathbf{R}^M} |f(x)|^2.$$

   Par exemple, soit B une matrice $N \times M$, et c un élément de $\mathbb{R}^N$. On appelle solution de $Bx = c$ au sens des moindres carrés (remarquons qu'une solution de $Bx = c$ n'existe pas forcément) un point de minimum de $|Bx - c|^2$. Nous allons identifier de telles solutions.

En effet, on cherche un point minimum de $(Bx-c, Bx-c) = (Bx, Bx)-(c, Bx)-(Bx, c)+(c, c)$, c'est à dire de $({}^tBBx, x) - 2({}^tBc, x) + (c, c)$. La matrice ${}^tBB$ est symétrique, et son noyau est le noyau de B (ceci car $tBBx = 0$ implique $|Bx|^2 = 0$, soit $Bx = 0$).

On vérifie que $\mathrm{Im}{}^tBb \subset \mathrm{Im}{}^tB$. De plus, pour $y \in (\mathrm{Im}{}^tB)^\perp$, on a

$$\forall x \in \mathbb{R}^N, (y, {}^tBx) = 0$$

ce qui implique $(By, x) = 0 \forall x \in \mathbb{R}^N$. Ainsi $By = 0$, donc $y \in \ker B$. La réciproque est claire. Par le théorème du rang on a $dim(\ker {}^tBB)+dim(\mathrm{Im}{}^tBB) = M = dim(\ker B)+dim(\mathrm{Im}B) = M$. On trouve donc que l'image de ${}^tBB$ est confondue avec l'image de ${}^tB$. L'équation donnant le minimum étant ${}^tBBx = {}^tBc$, on en conclut que x existe nécessairement, puisqu'il existe $d \in \mathbb{R}^N$ tel que ${}^tBBd = {}^tBc$. Le système d'équations ainsi écrit s'appelle le système d'équations normales. On remarque que c'est un espace affine passant par d dirigé par $\ker {}^tBB = \ker B$.

Une autre méthode plus directe: on diagonalise ${}^tBB$ dans une base orthonormée, les valeurs propres étant $0 \leq \lambda_1 \leq ... \leq \lambda_M$ associées aux vecteurs propres $(e_1, ...e_M)$. Alors on introduit p (éventuellement il n'existe pas) tel que $\lambda_p = 0$ et $\lambda_{p+1} > 0$. Alors $(e_1, ...e_p)$ forme une base de $\ker {}^tBB$, donc de $\ker B$. On constate alors qu'en écrivant $x = \sum_i y_i e_i$, on trouve

$$({}^tBBx, x) - 2({}^tBc, x) = \sum_{i>p} \lambda_i y_i^2 - 2\sum_i ({}^tBc, e_i)y_i.$$

Vérifiant alors que pour $i \leq p$, $({}^tBc, e_i) = (c, Be_i) = 0$, on en déduit que la fonction ne dépend que des $y_i, i > p$. On applique le résultat précédent et l'ensemble des solutions est un espace affine dirigé par $\ker B$.

Ce résultat se retrouve en considérant la projection de c sur l'hyperespace $\mathrm{Im}B$. Alors on réalise le minimum de la distance au sous espace fermé $\mathrm{Im}B$. Soit $p(c)$ cette projection. Le minimum de $|Bx-c|$ est alors l'ensemble des points tels que $Bx = p(c)$. En effet, par caractérisation de la projection, on a, pour tout $z \in \mathrm{Im}B$, $(Bx, z) = (p(c), z) = (c, z)$, ce qui équivaut à $\forall y$, $(Bx, By) = (p(c), By)$, soit utilisant $c - p(c)$ orthogonal à $\mathrm{Im}B$, $({}^tBBx, y) - (tBc, y)$ pour tout y. On vérifie immédiatement que si x_0 vérifie $Bx_0 = p(c)$, alors $(B(x - x_0), B(x - x_0)) = |Bx - c|^2 + (Bx_0, Bx_0) - (c, c)$, ce qui indique le résultat de minimum.

3. Recherche de la plus petite valeur propre d'une matrice symétrique.

   La plus petite valeur propre d'une matrice symétrique A de $\mathbb{R}^N \times \mathbb{R}^N$ est

   $$\lambda_1 = \inf_{v \in \mathbf{R}^N, ||v||=1} (Av, v) = \inf_{\mathbf{R}^N - \{0\}} \frac{(Av, v)}{(v, v)}.$$

   La matrice A est symétrique donc diagonalisable. On écrit $(Av, v) = \sum_i \lambda_i v_i^2$. Pour $\sum v_i^2 = 1$, on trouve $(Av, v) \geq \lambda_1$, avec égalité si $v_i = 0$ si $\lambda_i \neq \lambda_1$. Ceci permet de conclure sur l'existence d'un inf, qu'il est atteint, et que le minimum est λ_1. Le lieu des points réalisant le minimum est la sphère unité dans le sous-espace propre associé à λ_1. Quant à l'autre terme de l'égalité, il provient du fait que $\frac{v}{(v,v)^{\frac{1}{2}}}$ est de norme 1 lorsque $v \neq 0$.

4. On se donne $A = \{a \in L^\infty([0, 1]), 0 < \alpha \leq a(x) \leq \beta \forall x\}$. On se donne aussi $f_i, \bar{u}_i$ des fonctions (à préciser sur $[0, 1]$). On cherche à trouver a et u_i de sorte que

$$-\frac{d}{dx}(\frac{1}{a(x)}\frac{du_i}{dx}) = f_i(x), \forall x, u_i(0) = u_i(1) = 0 \tag{1.2.1}$$

$$\inf_{a\in A}\sum_i \int_0^1 |u_i(x) - \bar{u}_i(x)|^2 dx. \tag{1.2.2}$$

C'est un problème modèle pour certains problèmes de la physique. Ici, on cherche une équation de la chaleur (caractérisée par sa distribution a) telle que les résultats théoriques de l'observation (pour chaque donnée extérieure f_i on construit mathématiquement une solution de (1.2.1)) soient les plus proches possible de ce l'on observe $(\bar{u}_i)$.

Dans un premier temps, on peut résoudre explicitement (8.2.1) en introduisant $A(x) = \int_0^x a(s)ds$, mais trouver le meilleur a n'est pas encore à notre portée. On peut le faire quand $a(x)$ est une constante.

Dans le cas général, on trouve

$$\frac{du_i}{dx} = CA'(x) + A'(x)\int_0^x f_i(t)dt = \frac{d}{dx}(CA(x) + A(x)\int_0^x f_i(t)dt) - A(x)f_i(x),$$

soit

$$u_i(x) = CA(x) + A(x)\int_0^x f_i(t)dt - \int_0^x A(t)f_i(t)dt$$

en ayant utilisé $u_i(0) = 0$. On identifie C grâce à $u_i(1) = 0$, ce qui donne

$$u_i(x) = \frac{A(x)}{A(1)}(\int_0^1 A(t)f_i(t)dt - A(1)\int_0^1 f_i(t)dt) + \int_0^x (A(x) - A(t))f_i(t)dt.$$

Dans le cas $a(x) = a$, on trouve $u_i(x) = au_i^1(x)$, avec

$$u_i^1(x) = x\int_0^1 (t-1)f_i(t)dt + \int_0^x (x-t)f_i(t)dt.$$

Il est immédiat que le critère s'écrit

$$J(a) = a^2\int_0^1 (u_i^1(t))^2 dt - 2a\int_0^1 u_i^1(x)\bar{u}_i(x)dx + \int_0^1 (\bar{u}_i(x))^2 dx$$

et qu'il est minimum en $a_0 = \frac{\sum_{i=1}^{i=N}\int_0^1 u_i^1(t)\bar{u}_i(t)dt}{\sum_{i=1}^{i=N}\int_0^1 (u_i^1(t))^2 dt}$. Son minimum, d'après les inégalités de Cauchy-Schwarz, est positif ou nul et n'est nul que si tous les u_i^1 sont égaux à un coefficient foit $\bar{u}_i$.

5. Projection sur un convexe.

   Soit K un ensemble convexe fermé dans un espace de Hilbert V. On appelle projection de u_0 sur K, et on note $p(u_0)$, le point de K le plus proche de u_0, soit $||p(u_0) - u_0|| = \inf_{v\in K}||v - u_0||$. On note que, de la relation $\forall v \in K, ||v-u_0||^2 \geq ||p(u_0) - u_0||^2$, et, plus précisément de $\forall v \in K, \forall \lambda \in]0,1[, ||\lambda v + (1-\lambda)p(u_0) - u_0||^2 \geq ||p(u_0) - u_0||^2$, on tire

$$\lambda^2||v - p(u_0)||^2 + 2\lambda(v - p(u_0), p(u_0) - u_0) \geq 0.$$

   Faisant tendre λ vers 0, on en déduit l'inégalité

$$(v - p(u_0), p(u_0) - u_0) \geq 0 \forall v \in K.$$

Notons que cette égalité, dans le cas du plan, implique que $(v - p(u_0), u_0 - p(u_0)) \leq 0$, c'est-à-dire l'angle entre les vecteurs joignant la projection à u_0 et à un élément quelquonque de K est obtus.

Réciproquement, si cette inégalité est vérifiée, alors

$$||v-u_0||^2 = ||v-p(u_0)||^2+||p(u_0)-u_0||^2+2(v-p(u_0), p(u_0)-u_0) \geq ||v-p(u_0)||^2.$$

Il y a unicité de la projection. En effet, si on désigne par v_0 une autre projection, on a

$$(v - v_0, u_0 - v_0) \leq 0, (v - p(u_0), u_0 - p(u_0)) \leq 0.$$

Dans la premiére inégalité on considère $v = p(u_0)$ et dans la deuxième on considère $v = v_0$. Alors

$$(p(u_0) - v_0, u_0 - v_0) \leq 0, (-v_0 + p(u_0), -u_0 + p(u_0)) \leq 0.$$

Additionnant les deux égalités, on obtient

$$(p(u_0) - v_0, p(u_0) - v_0) \leq 0$$

ce qui implique $v_0 = p(u_0)$. Il y a unicité de la projection sur un convexe.

Ceci est la redémonstration du théorème de Hahn-Banach.

6. Gain minimum pour un turfiste.

   On suppose qu'un tiercé présente N chevaux au départ, chacun étant côté avec un rapport r_i. Montrer que la condition nécessaire et suffisante pour qu'un joueur récupère au moins sa mise est $\sum_i \frac{1}{r_i} \leq 1$.

   Posons les inconnues de ce problème. On suppose que le joueur joue x_i sur chaque cheval. Son gain est alors $y_{i_0} = x_{i_0} r_{i_0}$ si le cheval i_0 l'emporte. Pour simplifier notre analyse, on suppose $\sum x_i = 1$ (on mise 1) et on veut qu'il existe une combinaison de sorte que chaque y_i soit plus grand que 1. Ainsi on a

   $$\sum_i \frac{y_i}{r_i} = 1, y_i \geq 1 \forall i \Rightarrow 1 = \sum \frac{y_i}{r_i} \geq \sum_i \frac{1}{r_i}.$$

   Ainsi la condition $1 \geq \sum \frac{1}{r_i}$ est nécessaire pour que le gain soit au moins égal à la mise.

   Réciproquement, on suppose $1 \geq \sum \frac{1}{r_i}$, et on veut y_i pour tout i plus grand que i. Le cas limite est obtenu pour tous les y_i égaux, et cette valeur commune est $y_i = \frac{1}{\sum \frac{1}{r_p}}$, ce qui impose de choisir $x_i = \frac{1}{r_i} \frac{1}{\sum \frac{1}{r_p}}$. Dans ce cas, le gain est $\frac{1}{\sum \frac{1}{r_p}}$ pour tout i; il est donc plus grand que 1.

7. Un exemple de programme linéaire en recherche opérationnelle

   On considère M entrepôts, chacun présentant s_i unités d'un stock. On connait les N destinations, et on doit livrer r_j unités à la destination j. Les coûts de

transport unitaire c_{ij} de l'entrepôt i à la destination j sont connus, et on les appelle c_{ij}. Comment livrer au meilleur coût?

Pour formaliser le problème, on appelle v_{ij} la quantité livrée à j à partir de l'entrepôt i. On a comme conditions:

$$v_{ij} \geq 0, \sum_{j=1}^{j=N} v_{ij} \leq s_i, \sum_{i=1}^{i=M} v_{ij} \geq r_j$$

et le coût de livraison est $\sum_{i,j} c_{ij} v_{ij}$. On cherche l'inf de cette fonction.

Notons tout d'abord que, si l'on désigne par c_j le min pour $i = 1..M$ des c_{ij}, on trouve

$$\sum_{i,j} c_{ij} v_{ij} \geq \sum_{j=1}^{j=N} c_j (\sum_{i=1}^{i=M} v_{ij}) \geq \sum_j c_j r_j.$$

Ainsi l'inf existe et est strictement positif. Il faut voir si cette valeur est atteinte. Pour cela, il faut $c_j r_j = \sum_i c_{ij} v_{ij}$, donc si les c_{ij} sont ordonnés et distincts, tous les v_{ij} sont nuls sauf celui correspondant au plus petit des c_{ij}, où il vaut r_j.

On peut écrire la solution explicite dans le cas $M = N = 2$ et sous la **condition de compatibilité** $r_1 + r_2 \leq s_1 + s_2$ (on ne peut pas livrer plus que ce que l'on a). On trouve alors

$$\begin{array}{l} \min(c_{11}, c_{12}) = c_{12} \Rightarrow v_{12} = r_1, v_{11} = 0 \\ \min(c_{11}, c_{12}) = c_{11} \Rightarrow v_{12} = 0, v_{11} = r_1 \\ \min(c_{21}, c_{21}) = c_{21} \Rightarrow v_{21} = r_2, v_{22} = 0 \\ \min(c_{21}, c_{22}) = c_{22} \Rightarrow v_{21} = 0, v_{22} = r_2 \end{array}$$

On n'a même pas besoin de se poser les questions de v_{ij} entier. D'autre part, lorsque deux sont égaux, on peut choisir les quantités arbitrairement. On note ainsi que l'on se trouve donc sur le bord du domaine défini par les contraintes.

8. Un exemple de contrôle optimal

On considère $y^0 \in \mathbb{R}^N$, $T > 0$, $f \in L^1(]0, T[, \mathbb{R}^N)$ et A matrice $N \times N$, B matrice $N \times M$ données. On considère, pour chaque $v \in L^2(]0, T[, K)$, la solution $y(v)$ du système

$$\frac{dy(v)}{dt}(t) = Ay(v)(t) + Bv + f(t)$$

avec $y(v)(0) = y^0$. On cherche à minimiser le critère, qui peut s'exprimer par "avec un v aussi petit que possible sur $]0, T[$, trouver $y(v)$ aussi proche que possible de g aussi bien pondéré sur $]0, T[$ qu'en $t = T$" Le critère que j'écris est

$$\begin{array}{rl} J(v) = & \int_0^T (v(t), v(t)) dt + \int_0^T (Q(y(v)(t) - g(t)), y(v)(t) - g(t)) dt \\ & + (R(y(v)(T)) - g(T), y(v)(T) - g(T)) \end{array}$$

On note pour l'instant que $y(v)$ peut être calculée, par exemple à l'aide de $y(0)$ puis de l'exponentielle de A dans une base où par exemple A est diagonalisable, mais cela ne sera pas de grande aide pour calculer et minimiser le critère. On aura un principe dans la suite du cours.

9. Commande en temps minimal

   Dans ce cas, le critère s'écrit de la manière suivante: "atteindre une cible donnée C dans le temps le plus petit possible". On introduit alors le temps d'arrivée à la cible:

   $$J(v) = +\infty \text{ si } y(v) \notin C \forall t$$
   $$J(v) = \inf\{t \geq 0, y(v)(t) \in C\} \text{ si il existe } t_0 \text{ tel que } y(v)(t_0) \in C.$$

   Commander le système en temps minimal est trouver $\inf J$ pour v dans l'espace de commande et trouver un v_0 tel que $J(v_0) = \inf J$.

10. Equilibre d'un fil pesant.

    On se place dans le champ de pesanteur $\vec{g} = -g\vec{j}$, et on se donne deux points (x_0, y_0) et (x_1, y_1). On se place dans une situation suffisamment simple pour qu'un fil placé entre ces deux points puisse être représenté par $y(x)$, avec $y(x_0) = y_0, y(x_1) = y_1$. La longueur de ce fil est supposée fixe, égale à l, ce qui se traduit par l'égalité (basée sur la notion d'abscisse curviligne, $s = 0$ au point (x_0, y_0) et $s = l$ au point (x_1, y_1))

    $$l = \int_0^l ds = \int_{x_0}^{x_1} (1 + (y'(x))^2)^{\frac{1}{2}} dx.$$

    Il est en équilibre lorsque son énergie potentielle est minimum. L'origine de l'énergie potentielle est placée en y_1. Alors, si on désigne par ρ sa masse linéique, l'énergie potentielle du fil est

    $$\rho g \int_0^l (y(x(s)) - y_1) ds = -\rho g y_1 l + \rho g \int_{x_0}^{x_1} y(x)(1 + (y'(x))^2)^{\frac{1}{2}} dx.$$

    L'énergie totale, qui est constante, fait intervenir la vitesse, qui est donc nulle. On a donc le problème

    $$\inf_{y \in C^0} \int_{x_0}^{x_1} y(x)(1+(y'(x))^2)^{\frac{1}{2}} dx, \int_{x_0}^{x_1} (1+(y'(x))^2)^{\frac{1}{2}} dx = l, y(x_0) = y_0, y(x_1) = y_1.$$

11. Le problème de Pappus, ou comment Didon a pu construire Carthage.

    "Parmi toutes les courbes de longueur donnée joignant $(0, 0)$ à $(\xi, 0)$, trouver celle qui conduit à l'aire maximum"

    On se donne l'équation de cette courbe $y = v(x)$. On a les conditions

    $$v \geq 0, v(0) = v(\xi) = 0, \int_0^\xi (1 + (v'(x))^2)^{\frac{1}{2}} dx = l$$

    et on recherche à minimiser $-\int_0^\xi v(x)dx$. Notons ici l'emploi du signe $-$ lorsqu'on a à trouver un maximum et non un minimum.

12. Principe de Fermat et de Huyghens

On veut trouver la trajectoire reliant en temps minimum les points (x_0, y_0) et (x_1, y_1), en sachant qu'en (x, y), la vitesse est $c(x, y)$. Alors on cherche v (que l'on précisera) telle que $v(x_0) = y_0$, $v(x_1) = y_1$ et $\int_0^s \frac{ds}{c(x(s),y(s))}$ soit minimum, c'est-à-dire

$$\inf \int_{x_0}^{x_1} \frac{(1 + (v'(x))^2)^{\frac{1}{2}}}{c(x, v(x))} dx.$$

Lorsque on veut par exemple évaluer le rayon entre deux milieux de vitesse c_1 et c_2, tels que $c(x, y) = c_1 1_{x>0} + c_2 1_{x>0}$, on a donc, appliquant ce qui est écrit ci-dessus à trouver le lieu de

$$\inf[\int_{x_0}^{0} \frac{(1 + (v'(x))^2)^{\frac{1}{2}}}{c_1} dx + \int_0^{x_1} \frac{(1 + (v'(x))^2)^{\frac{1}{2}}}{c_2} dx].$$

13. Problèmes d'équilibre en mécanique des milieux continus

Soit Ω un ouvert de $\mathbb{R}^n$ et Γ sa frontière. On se donne les trois énergies

$$U_1(v) = \frac{1}{2}\lambda \int_\Omega |\nabla v|^2 dx$$

$$U_2(v) = \frac{1}{2} k \int_\Omega |v|^2 dx$$

$$U_3(v) = -\int_\Omega f(x)v(x)dx$$

qui sont respectivement l'énergie potentielle de déformation, l'énergie potentielle élastique, l'énergie d'une force extérieure constante dans le temps.

On étudie deux fonctionnelles $J_1 = U_1 + U_2 + U_3$ et $J_2 = U_1 + U_3$. On écrira quatre types de problèmes:

$$\inf_{v \in H_0^1(\Omega)} J_2(v), \inf_{v \in H^1(\Omega)} J_1(v), \inf_{v \in H^1(\Omega), v|_\Gamma \geq 0} J_1, \inf_{v \in H_0^1(\Omega), v \geq \psi} J_2$$

qui sont respectivement les problèmes de Dirichlet, Neumann, élasticité avec contraintes unilatérales, équilibre avec obstacle.

Pour introduire certaines des méthodes de ce cours, traitons le premier problème. Nous allons le faire à l'aide de ce que nous avons utilisé pour le théorème de Hahn-Banach. On suppose que u existe. Alors, pour toute fonction ϕ dans $C_0^\infty(\Omega)$, on remarque que $u + \phi \in H_0^1(\Omega)$, ainsi on a

$$J_2(u + \phi) \geq J_2(u).$$

Cette inégalité se traduit par

$$\forall \phi \in C_0^\infty(\Omega), \lambda \int_\Omega \nabla u \nabla \phi + J_2(\phi) \geq 0.$$

On choisit alors ψ et on considère $\phi = \varepsilon\psi$, où ε tend vers 0. Alors on en déduit, au passage à la limite, l'inégalité $\lambda \int_\Omega \nabla u \nabla\psi - \int f\psi \geq 0$ pour toute $\psi \in C_0^\infty(\Omega)$. On choisit alors $-\psi$, pour obtenir $\lambda \int_\Omega \nabla u \nabla\psi - \int f\psi = 0 \forall \psi \in C_0^\infty(\Omega)$. Un résultat d'intégrations par parties indique que, au sens des distributions de $H^{-1}(\Omega)$ (dual, rappelons le, des distributions de $H_0^1(\Omega)$), on a la relation

$$-\lambda\Delta u = f$$

Réciproquement, lorsque u est dans $H_0^1(\Omega)$ solution dans $H^{-1}(\Omega)$ de ce problème, alors par écriture du produit scalaire qui correspond à la dualité des distributions, on trouve

$$J_2(v) - J_2(u) = \frac{1}{2}\lambda \int (\nabla v - \nabla u)^2 dx.$$

14. Un exemple simple avec contraintes.

On veut trouver $\min(\frac{1}{2}v^2 - cv)$ sous la contrainte $v \leq b$. Pour cela, on voit que, si $b \leq c$, $\min_{v\leq b}(\frac{1}{2}v^2 - cv) = (\frac{1}{2}v^2 - cv)|_{v=b}$ et si $b > c$, $\min_{v\leq b}(\frac{1}{2}v^2 - cv) = (\frac{1}{2}v^2 - cv)|_{v=c}$. Dans le premier cas, la contrainte est saturée, dans le deuxième cas elle est insaturée.

15. Problème de Neumann avec contrainte.

Nous étudions ici le cas du problème $\inf J_1(u), u|_\Gamma \geq 0$, où $u \in H^1(\Omega)$, $\partial\Omega = \Gamma$. On prend d'abord $\phi \in C_0^\infty(\Omega)$, ainsi, pour tout $\varepsilon > 0$, $u + \varepsilon\phi$ est dans le domaine K défini par $K = \{u \in H^1(\Omega), u|_\Gamma \geq 0\}$ dès que $u \in K$. On applique alors la même méthode que précédemment, de faire tendre ε vers 0 après avoir divisé l'inégalité déduite de $J_1(u + \varepsilon\phi) \geq J_1(u)$ par ε. Ainsi on a

$$\forall \phi \in C_0^\infty(\Omega), \lambda \int \nabla u \nabla\phi dx + k \int u\phi dx = \int f\phi dx.$$

On en déduit, dans $\mathcal{D}'(\Omega)$, l'égalité

$$-\lambda\Delta u + ku = f.$$

Désormais, on considère $v \in H^1(\Omega), v|_\Gamma \geq 0$. Ainsi, de $J_1(v) \geq J_1(u)$, écrivant $v = u + (v - u)$, on déduit

$$J_1(v) - J_1(u) = U_1(v-u) + U_2(v-u) + \int_\Omega [\lambda \nabla u \nabla(v-u) + ku(v-u) - f(v-u)]dx \geq 0 \tag{1.2.3}$$

D'une part, si $v = cu, c \geq 0$, alors $v \in K$. On trouve alors

$$(c-1)\int_\Omega (\lambda(\nabla u)^2 + k(u)^2 - fu)dx \geq 0.$$

Comme $c \in]0, +\infty[$, alors $c - 1 \in]-1, +\infty[$. On peut prendre une valeur négative et une valeur positive de $c - 1$, ce qui implique la relation

$$\int_\Omega (\lambda(\nabla u)^2 + k(u)^2 - fu)dx = 0.$$

Remplaçant alors cette égalité dans l'inégalité (1.2.3), on trouve, pour tout $v \in K$:

$$U_1(v-u) + U_2(v-u) + \int_\Omega [\lambda \nabla u \nabla v + kuv - fv] dx \geq 0$$

On remplace f par $-\lambda\Delta u + ku$ et on utilise la relation $\int \Delta uv dx = -\int_\Omega \nabla u \nabla v dx + \int_\Omega \partial_n uv d\sigma$ (qui est une manière de définir $\partial_n u$ pour $u \in H^1(\Omega)$ et $v \in H^1(\Omega)$ comme le résultat d'un théorème de Riesz)[1].

La relation obtenue est alors $\forall v \in K, \int_\Gamma \partial_n uv|_\Gamma d\sigma \geq 0$.

Nous avons pu ici étudier le problème facilement car la fonctionnelle est une forme quadratique. Dans le cas où elle ne l'est pas, il s'agit d'étudier $u + \psi$, et on vérifie que si $x \in \Gamma_\alpha$ où Γ_α est la partie du bord où u est supérieur ou égal à α, alors on peut prendre ψ tel que $\psi = 0$ sur $\Gamma - \Gamma_\alpha$ et $|\psi| \leq \frac{\alpha}{2}$ sur Γ_α, ψ identiquement égale à 1 sur le bord dans un voisinage d'un point x_0 de Γ_α. On peut alors vérifier que $u + \psi$ et que $u - \psi$ sont dans K, ce qui permet d'obtenir directement, avec $v - u = \pm\psi$, la relation au bord $\int_\Gamma \partial_n u \psi d\sigma = 0$, ce qui donne $\partial_n u = 0$ sur Γ_α. On a donc

$$\forall \alpha > 0, \partial_n u \Gamma_\alpha = 0, \int_\Gamma u \partial_n u d\sigma = 0$$

ce qui permet de partitionner Γ en $\Gamma_1 = \{x, u(x) = 0\}$ et $\Gamma_2 = \Gamma_0 = \Gamma - \Gamma_\alpha$, sur lequel $\partial_n u = 0$, et on a, par la condition $\int_\Gamma \partial_n uv d\sigma \geq 0$ pour tout $v, v|_\Gamma \geq 0$, la condition $\partial_n u \geq 0$.

16. Cas de non existence d'un minimum.

On se place dans l'espace $H^1(]0,1[)$ muni de la norme usuelle, et on définit $J(v) = \int_0^1 [(|v'(x)| - 1)^2 + (v(x))^2] dx$. On note que $J(v) \geq 0$ et qu'il n'existe pas de u tel que $J(u) = 0$. En effet, si il en existe un, $|u'| = 1$ p.p. et $u = 0$ impossible dans H^1. D'autre part, si on construit $u_n(x) = \frac{1}{2n} - |x - \frac{2k+1}{2n}|$ sur l'intervalle $[\frac{k}{n}, \frac{k+1}{n}]$ pour $0 \leq k \leq n-1$, on trouve $\int_{\frac{k}{n}}^{\frac{k+1}{n}} (u_n(x))^2 = 2\int_0^{\frac{1}{2n}} x^2 = \frac{1}{6n^3}$ et $\int_{\frac{k}{n}}^{\frac{k+1}{n}} (|u'(x)| - 1)^2 dx = 0$. Ainsi

$$J(u_n) = \frac{1}{6n^2}$$

et $\inf J = 0$, alors qu'il n'existe pas de u tel que $J(u) = \inf J$.

[1] On introduit la fonctionnelle $v \to \int_\Omega \nabla u \nabla v + < \Delta u, v >$. Lorsque $v \in C^\infty(\Omega)$, il est clair que cette fonctionnelle est continue et que, par dualité, comme $u \in H^1(\Omega)$, $\Delta u \in H^{-1}(\Omega)$ lorsque le bord est régulier, on trouve

$$|\int_\Omega \nabla u \nabla v + < \Delta u, v > | \leq C||v||_{H^1(\Omega)}.$$

Pour $v = \phi \in C_0^\infty(\Omega)$, on trouve 0, donc c'est une distribution qui ne considére que les valeurs au bord de $v = \phi$. D'autre part, lorsque $u \in H^2(\Omega)$, on trouve que cette fonctionnelle permet de définir la dérivée normale de u, $\partial_n u$ par la formule de Green usuelle.

Finalement, pour $u \in H^2(\Omega)$ et $v \in C^\infty(\Omega)$, il existe C_1 telle que (on améliore la relation précédente)

$$|\int_\Omega \nabla u \nabla v + < \Delta u, v > | \leq C_1 ||v|_\Gamma||_{H^{\frac{1}{2}}(\Gamma)}.$$

17. Minimisation quadratique dans $\mathbb{R}^2$.

On introduit la fonctionnelle $J(y_1, y_2) = \frac{1}{2}(y_1^2 + y_2^2) - b_1 y_1 - b_2 y_2$ et on cherche à résoudre les deux problèmes

$$\inf J(y), a_1 y_1 + a_2 y_2 = 0$$

$$\inf J(y), a_1 y_1 + a_2 y_2 \leq 0$$

Dans le premier cas, on a plusieurs méthodes à notre disposition. La plus évidente est de supposer $a_1 \neq 0$, ainsi $y_1 = -\frac{a_2}{a_1} y_1$, et on se ramène à

$$\inf \frac{1}{2}(1 + \frac{a_1^2}{a_2^2}) y_1^2 - \frac{b_2 a_1 - b_1 a_2}{a_1} y_2$$

qui est atteint au point $y_2 = a_1 \frac{b_2 a_1 - b_1 a_2}{a_1^2 + a_2^2}$ et donc $y_1 = -a_2 \frac{b_2 a_1 - b_1 a_2}{a_1^2 + a_2^2}$.

On peut simplifier les expressions en vérifiant que, dans y_2, le coefficient de b_2 s'écrit avec $a_1^2/(a_1^2 + a_2^2)$, ainsi

$$(y_1, y_2) = (b_1, b_2) - \frac{a_1 b_1 + a_2 b_2}{a_1^2 + a_2^2}(a_1, a_2).$$

Cette méthode n'est pas instructive, mais son résultat l'est: le minimum est obtenu au point $b + \lambda a$. Le réel λ est nul lorsque $a.b = 0$.

Distinguons les deux cas. Notons avant cela que le minimum absolu de la fonctionnelle se situe au point b. Si b est dans la contrainte, alors ce minimum absolu est atteint sur la contrainte, et donc le problème

$$\inf J, a.y = 0$$

admet comme solution $y = b$, de même que le problème

$$\inf J, a, y \leq 0.$$

Si b n'est pas dans la contrainte égalité, on désigne par b_0 la projection de b sur la droite $a.y = 0$. On a bien sûr $J(y) = -\frac{1}{2}b^2 + \frac{1}{2}(y - b)^2$, donc minimiser J revient donc à minimiser la distance de b à la droite $a.y = 0$. Le point qui réalise ceci est bien sûr $y = b_0$. On vérifie alors que $y = b + (b_0 - b)$, et, avec $b_0 - b = -\lambda a$, on a l'égalité $y = b - \lambda a$. Le minimum est solution de $y - b + \lambda a = 0$, ce qui sera dans le cours l'égalité de définition du point selle et du multiplicateur de Lagrange. On note que, par $b_0.a = 0$, on a $\lambda = \frac{a.b}{a^2}$.

On étudie maintenant la contrainte inégalité $a.y \leq 0$.

Si on considère b tel que $a.b \leq 0$, on n'a besoin de rien d'autre, le minimum absolu est dans l'espace des contraintes, donc le minimum de la fonctionnelle est atteint en $y = b$. On suppose donc que b est dans la zone $a.y > 0$. Grâce à l'égalité $b_0 = b - \lambda a$ et à l'égalité $b_0.a = 0$, on trouve que $\lambda a^2 > 0$, et donc $\lambda > 0$ et le minimum est en b_0.

Lorsque on suppose que b n'est pas dans la zone $a.y > 0$, on trouve que $b_0 = b - \lambda a$ avec $\lambda a^2 \leq 0$ et $\lambda \leq 0$. Le minimum est alors obtenu en b et on a $b = b + 0a$.

On voit sur cet exemple et sur la notion de projection que l'on forme $y - b + \lambda a$ et $a.y = 0$. Lorsque la résolution de ce système conduit à $\lambda \leq 0$, on dit que la contrainte est insaturée et on a $y = b$ comme minimum. Le point de minimum est dans l'espace des contraintes. Lorsque la résolution du système conduit à $\lambda \geq 0$, la contrainte est saturée et $y = b - \lambda a$ convient.

Chapter 2

Minimum dans $\mathbb{R}^N$ ou dans un espace de Hilbert, conditions d'Euler et de Legendre

2.1 Condition générale d'existence (suffisante)

Nous allons d'abord donner des conditions suffisantes d'existence d'un minimum. Le théorème le plus classique, que l'on trouve au début de chaque cours d'optimisation, est

Théorème 2.1 *Soit $K \subset \mathbb{R}^N$, soit J une fonctionnelle continue sur Ω contenant K, et K fermé.*

Si K est compact, ou si J est ∞ à l'∞ (c'est-à-dire, pour toute suite v_n telle que $|v_n| \to +\infty$, $J(v_n) \to +\infty$), alors J a au moins un minimum sur K.

On peut extraire de toute suite minimisante sur K une sous-suite convergeant vers un point de minimum sur K.

Preuve Toute partie de $\mathbb{R}$ admet une borne inférieure l, éventuellement $-\infty$. Si il s'agit de $-\infty$, on a immédiatement l'existence d'une suite u_n telle que $J(u_n) \to -\infty$. Si l est fini, et si K est compacte, d'une suite u_n telle que $J(u_n)$ tend vers l, on peut extraire (car $u_n \in K$ compact), une sous-suite convergente $u_{n'} \to a$. Comme J est continue, $J(u_{n'})$ tend vers $J(a)$, et donc $J(a) = l$. Si K n'est pas compacte, on vérifie cependant que la suite est bornée (si elle ne l'était pas, on trouverait une sous-suite extraite $u_{n'}$ telle que $|u_{n'}| \to +\infty$, auquel cas $J(u_{n'}) \to +\infty$ par l'hypothèse sur le comportement de J, et donc $J(u_{n'})$ ne converge pas vers l). Soit B une boule fermée contenant tous les termes de la suite. Alors $u_n \in K \cap B$ est une suite dans un compact, une suite extraite converge donc vers une valeur minimisante.

On note que dans l'exemple 16 de l'introduction, la fonctionnelle vérifie la condition à l'infini, mais il n'y a pourtant pas de minimum car dans un espace de dimension infinie, un fermé borné n'est pas necessairement compact.

Il s'agit maintenant d'être capable, comme dans les exemples traités précédemment, de calculer les solutions. Nous allons faire cela, en écrivant des conditions très anciennes, nécessaires pour certaines, suffisantes pour d'autres.

2.2 Condition d'Euler, condition de Legendre

Du traitement des exemples 13 et 15, on déduit un certain nombre de notions. Nous reviendrons sur certaines d'entre elles plus loin. Pour l'instant, intéressons nous à deux notions:

- la notion de dérivée dont nous avons besoin
- la notion de direction admissible.

La notion de dérivée que nous cherchons à obtenir s'obtient en comparant (ce qui a été fait dans les exemples 13 et 15), $J(u+\varepsilon v)$ et $J(u)$ après avoir divisé par ε et fait tendre ε vers 0. On voit ainsi que la bonne notion est de considérer

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}[J(u+\varepsilon v)-J(u)]$$

et d'écrire l'inégalité, valable pour tout v tel que $u+\varepsilon v$ est dans le domaine étudié

$$\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}[J(u+\varepsilon v)-J(u)]\geq 0.$$

2.2.1 Dérivabilité au sens de Fréchet et au sens de Gâteaux

La dérivée d'une fonction d'une variable élément d'un espace vectoriel de dimension finie doit être généralisée aux fonctionnelles, application d'un espace vectoriel de dimension infinie dans $\mathbb{R}$. Il faut se placer dans un espace normé, et un espace pour lequel l'espace dual est isomorphe à l'espace (on verra plus loin que cela permettra de définir une application gradient). On se place sur un espace de Hilbert V, dans lequel on a isomorphisme entre V et V', et donc le théorème de Riesz.

Définition 2.1 *Lorsque, pour tout w, la limite $\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}[J(u+\varepsilon v)-J(u)]$ existe, on la note $J'(u;w)$ et on l'appelle dérivée directionnelle de J en u dans la direction w, qui est une fonction définie de $V\times V$ dans $\mathbb{R}$, homogène de degré 1 dans la variable w.*

Lorsque, de plus, la fonction $w\to J'(u;w)$ est une fonction linéaire continue, alors il existe, par le théorème de Riesz, un élément de l'espace de Hilbert V, que l'on appelle la dérivée de Gâteaux de J en u et que l'on note $J'(u)$. On notera souvent de la même façon la forme linéaire et son représentant dans le produit scalaire, soit $(J'(u),w)=J'(u;w)$.

On peut aussi définir la dérivée seconde J"(u) si elle existe, lorsque la limite

$$\lim_{\delta\to 0}\frac{1}{\delta}[J'(u+\delta w_1;w_2)-J'(u;w_2)]$$

existe pour tout (w_1,w_2) et est une forme bilinéaire continue sur $V\times V$. La limite est alors $(J"(u)w_1,w_2)$ par représentation des formes bilinéaires continues.

On rappelle la définition de la dérivée au sens de Fréchet, qui n'est plus cette fois une forme linéaire définie sur chaque direction:

Définition 2.2 *J est dérivable au sens de Fréchet en u si*

$$J(u+v)=J(u)+L_u(v)+\varepsilon(v)$$

avec L_u forme linéaire continue sur V et $\frac{\varepsilon(v)}{||v||}\to 0$ quand $v\to 0$.

Lorsque J est dérivable au sens de Fréchet, elle est dérivable au sens de Gâteaux, mais la réciproque est fausse, car l'écriture de la dérivabilité au sens de Fréchet correspond à $\frac{\varepsilon(v)}{||v||}$ tend vers 0, alors que la dérivabilité au sens de Gateaux correspond à $\frac{\varepsilon(\lambda w)}{\lambda}$ tend vers 0 lorsque λ tend vers 0 et on perd l'uniformité de w.

On peut alors écrire des formules de Taylor sur v a l'ordre 2 si J est deux fois différentiable au sens de Fréchet:

$$J(u+v) = J(u) + (J'(u), v) + \frac{1}{2}(J"(u)v, v) + o(||v||^2) \tag{2.2.1}$$

Si J est diff'erentiable au sens de Fréchet et si sa dérivée est différentiable au sens de Gateaux, alors on a aussi une formule de Taylor:

$$J(u+tw) = J(u) + t(J'(u), w) + \frac{1}{2}t^2(J"(u)w, w) + o(t^2). \tag{2.2.2}$$

Lorsque J" est continue, on peut écrire la formule de Taylor avec reste intégral

$$J(u+tw) = J(u) + t(J'(u), w) + t^2 \int_0^1 (1-x)(J"(u+xtw)w, w)dx. \tag{2.2.3}$$

La démonstration de ces égalités de Taylor peut par exemple se faire en considérant la fonction de la variable réelle

$$\phi(t) = J(u+tw).$$

On vérifie que

$$\frac{\phi(t+h) - \phi(t)}{h} \to (J'(u+tw), w)$$

ainsi $\phi'(t) = (J'(u+tw), w)$.

On voit alors que $\frac{\phi'(t)-\phi'(0)}{t} = \frac{(J'(u+tw),w)-(J'(u),w)}{t}$ tend vers $\phi"(0) = (J"(u)w, w)$. Ainsi on peut écrire la formule de Taylor

$$\phi(t) = \phi(0) + t\phi'(0) + \frac{t^2}{2}\phi"(0) + o(t^2)$$

et on a obtenu la formule de Taylor pour une fonction différentiable, qui admet une dérivée seconde au sens de Gateaux.

D'autre part, si J est deux fois différentiable au sens de Fréchet dans un voisinage de u

$$\phi"(t) = (J"(u+tw)w, w)$$

ainsi la formule de Taylor avec reste intégral pour la fonction ϕ conduit à l'égalité (2.2.3).

Avec les outils de differentiabilité ainsi définis, on peut donner les résultats d'optimalité connus soul le nom de condition d'Euler et de Legendre.

2.2.2 Conditions necessaires d'optimalité. Conditions suffisantes d'optimalité

On écrit des conditions nécessaires dans le

Théorème 2.2 *Soit V un espace de Hilbert et J une fonctionnelle différentiable (1 ou 2 fois) au sens des définitions précédentes*

Pour que $u \in V$ soit solution de

$$\begin{cases} \inf J(v) \\ v \in V \end{cases} \tag{2.2.4}$$

il **FAUT** *que $J'(u) = 0$* **(condition d'Euler)**.

(c'est-à-dire former cette équation, appelée équation d'Euler, donne tous les minima, entre autres points (elle donne aussi tous les maxima locaux)).

Si J est différentiable deux fois, on a, de plus **nécessairement**

$$\forall w \in V, (J"(u)w, w) \geq 0.$$

(condition de Legendre)

Démonstration:

On vérifie que, si u est un point d'optimum de J, alors, pour tout $v \in V$ on a

$$J(u + v) \geq J(u).$$

Si on utilise la dérivée de Fréchet de J, on en déduit que

$$\forall v \in V, L_u(v) + o(v) \geq 0.$$

On écrit $v = tw$, et on fait tendre t vers 0, $t > 0$. On en déduit , par passage à la limite, $L_u(w) \geq 0$. On choisit alors $v = -tw$, $t > 0$ et on en déduit $L_u(-w) \geq 0$. On a alors, $\forall w, L_u(w) = 0$. Ceci équivaut à $J'(u) = 0$.

Pour la condition de Legendre, on suppose que la fonctionnelle est dérivable au sens de Fréchet et que sa dérivée de Fréchet est différentiable au sens de Gateaux.

On utilise alors la formule de Taylor (2.2.2), ce qui donne, si u est un minimum, utilisant $J'(u) = 0$:

$$J(u + tw) = J(u) + \frac{t^2}{2}(J"(u)w, w) + o(t^2)$$

et l'inégalité $J(u + tw) \geq J(u)$ conduit à $(J"(u)w, w) \geq 0$ pour tout w. Le théorème est démontré.

Ce théorème est complété par une écriture de conditions suffisantes, valables pour un minimum local

Théorème 2.3 *Un ensemble de conditions suffisantes pour que u soit solution du problème du théorème précédent est*

$$J'(u) = 0$$

et *pour tout $\tilde{u}$ dans un voisinage de u_0, on ait la condition $(J"(\tilde{u})w, w) \geq 0$.* ***(condition forte de Legendre**)*

De manière opératoire, on peut aussi écrire une condition plus forte que la condition forte sous la forme

Il existe $\alpha > 0$ tel que $(J"(u)w, w) \geq \alpha(w, w)$[1].

Démontrons le théorème. On suppose que $J'(u) = 0$ et $(J"(\tilde{u}w, w) \geq 0$ pour tout $\tilde{u}$ dans un voisinage de u, et J deux fois Fréchet différentiable. Alors en utilisant la formule de Taylor avec reste intégral

$$J(u + tw) = J(u) + t^2 \int_0^1 (1 - x)(J"(u + txw)w, w)dx$$

et l'hypotèse sur la dérivée seconde qui implique que, pour tout $\tilde{u}$ dans ce voisinage de u, on choisit $t = 1$ et $w = \tilde{u} - u$ de sorte que $u + txw = x\tilde{u} + (1 - x)u$ est dans ce même voisinage, alors $J(\tilde{u}) \geq J(u)$ et u est un point de minimum local, ce qu'il fallait démontrer.

Notons que l'on n'a pas ainsi de condition nécessaire et suffisante. En effet, si on considère dans $V = \mathbb{R}$ $J(x) = x^6(1 + \sin\frac{1}{x})$, et $J(0) = 0$, on vérifie que $J(x) \geq 0$ car $\sin u \geq -1$. Ainsi $J(x) \geq J(0)$ pour tout x et 0 est un point de minimum absolu. On vérifie que J est continue en 0 (car $\lim x \sin\frac{1}{x} = 0$). Sa dérivée est $J'(x) = 6x^5(1+\sin\frac{1}{x}) - x^4\cos\frac{1}{x}$, elle vérifie $J'(x) \to 0$ lorsque x tend vers 0, et de plus, $\frac{J(x)-J(0)}{x}$ tend vers 0, donc J est dérivable et sa dérivée est continue. Alors $J"(x) = -x^2[\sin\frac{1}{x} - 30x^2(1+\sin\frac{1}{x}) - 10x\cos\frac{1}{x}]$. On vérifie que $J"(0) = 0$ et que $J"(\frac{1}{(n+\frac{1}{2})\pi}) = -(\frac{1}{(n+\frac{1}{2})\pi})^2[(-1)^n - 30(\frac{1}{(n+\frac{1}{2})\pi})^2(1+(-1)^n)]$, dont le signe est alternativement + et − pour n pair ou impair assez grand (par exemple $n \geq 4$). Ceci prouve que J ne vérifie pas la condition forte de Legendre et pourtant J admet un minimum absolu en 0.

2.3 Inéquation d'Euler dans un problème avec contraintes

Les problèmes avec contrainte s'écrivent aussi problème d'optimum liés. Il s'agit en particulier de l'exemple 15. On voit, dans ce problème, que la remarque utilisée généralement est que l'on doit pouvoir avoir $u+\varepsilon\phi$ dans le domaine K si u est donnée, afin d'écrire les conditions $J(u + \varepsilon\psi) \geq J(u)$. Il faut alors que ψ soit positive sur le bord lorsque $u|_\Gamma$ est nulle en ce point du bord, alors que, modulo le fait que ε soit choisi assez petit, ψ peut être prise arbitraire sur le bord hors des points où u est nulle.

Lorsque K est l'ensemble des contraintes, et lorsque $u \in K$, on définit les **directions admissibles de u dans K** par

Définition 2.3 *L'espace des directions admissibles au sens de Fréchet est l'ensemble des w de V est une direction admissible pour u sur K si il existe une suite w_n de V tendant vers w et une suite $e_n \geq 0$ telle que $u + e_n w_n \in K$. L'ensemble des directions admissibles est noté $K(u)$.*

Définition 2.4 *L'espace des directions admissibles au sens de Gâteaux est l'ensemble des w tels que, pour ε assez petit, $u + \varepsilon w$ soit dans K. L'ensemble de telles directions w est aussi appelé ensemble de directions admissibles intérieures et noté $\dot{K}(u)$.*

[1]Notons que dans un Hilbert de dimension finie, cette inégalité est équivalente à l'inégalité $(J"(u)w, w) > 0$ pour tout w non nul, puisque dans ce cas là la matrice $J"(u)$ n'a pas de vecteur propre nul, et α est sa plus petite valeur propre

On note que les deux ensembles ainsi définis sont des cônes, et que $\dot{K}(u) \subset K(u)$..

On a alors les conditions nécessaires suivantes sur un minimum de la fonctionnelle sous contraintes:

Théorème 2.4 *(Inéquations d'Euler)*

Si J est dérivable au sens usuel (de Fréchet), pour que u soit solution de (2.2.4), il **faut** *que*

$$\forall w \in K(u), (J'(u), w) \geq 0.$$

Si J est dérivable au sens de Gâteaux, il **faut** *que*

$$\forall w \in \dot{K}(u), (J'(u), w) \geq 0.$$

Soit u une solution de (2.2.4). Alors, comme $u + e_n w_n \in K$, on a $J(u + e_n w_n) \geq J(u)$. Ainsi on en déduit

$$\frac{1}{e_n}[J(u + e_n w_n) - J(u)] \geq 0 \forall n$$

puisque $e_n \geq 0$. Ainsi, en passant à la limite dans l'égalité de définition de la dérivée de Fréchet, on obtient $\frac{1}{e_n}[J(u + e_n w_n) - J(u) - (J'(u), e_n w_n)] \to 0$, ainsi, écrivant $(J'(u), w_n) - (J'(u), w) = (J'(u), w_n - w) \to 0$, on a

$$(J'(u), w) \geq 0.$$

Pour le deuxième, on vérifie que $J(u + \varepsilon w) - J(u) \geq 0$, ainsi, en divisant par ε et en faisant tendre ε vers 0 pour $w \in \dot{K}(u)$, on trouve

$$\forall w \in \dot{K}(u), (J'(u), w) \geq 0.$$

2.4 Multiplicateurs de Lagrange

Nous appliquons les résultats de la section précédente à des contraintes particulières, qui sont les plus simples que nous rencontrons. Les contraintes les plus simples sont les contraintes égalités et les contraintes inégalités. Par exemple, on peut écrire

$$K = \{u \in V, F_1(u) = 0, F_2(u) = 0, ..F_m(u) = 0\}$$

les fonctions $F_1, ...F_m$ étant continues.

Par exemple, lorsque $V = \mathbb{R}^3$, on peut donner comme condition l'appartenance à la sphère unité, qui s'écrit $x^2 + y^2 + z^2 - 1 = 0$. Ici $F(x, y, z) = x^2 + y^2 + z^2 - 1$.

Nous traitons le cas particulier de la contrainte égalité $x^2 + y^2 + z^2 = 1$.

Commençons par l'ensemble ouvert $\dot{K}((x, y, z))$. On trouve que $(x + \epsilon w_1)^2 + (y + \epsilon w_2)^2 + (z + \epsilon w_3)^2 = 1$ et $x^2 + y^2 + z^2 = 1$. Ainsi, en utilisant ces deux égalités et en divisant par ϵ, on obtient

$$(*)(xw_1 + yw_2 + zw_3) = -\frac{\epsilon}{2}||w||^2.$$

En faisant tendre ϵ vers 0, on trouve que $xw_1 + yw_2 + zw_3 = 0$ car (x, y, z) et (w_1, w_2, w_3) sont indépendants de ϵ. D'autre part, en remplacant cette égalité dans (*), on trouve $\epsilon||w||^2 =$

0. Comme on prend ϵ quelconque assez petit, la norme de w est nulle donc $w = 0$. On trouve $\dot{K}((x, y, z)) = \{(0, 0, 0)\}$.

D'autre part, considérons maintenant la définition de $K((x, y, z))$. Alors $w \in K((x, y, z))$ lorsqu'il existe une suite e_n tendant vers 0 et une suite $w^n = (w_1^n, w_2^n, w_3^n)$ tendant vers w telles que $(x, y, z) + e_n w^n$ soit dans la sphère. On cherche des conditions nécessaires pour que cela soit le cas. Comme précédemment, on écrit les deux égalités et on obtient

$$xw_1^n + yw_2^n + zw_3^n = -\frac{e_n}{2}||w^n||^2.$$

En considérant la limite lorsque n tend vers l'infini, le membre de gauche tend vers $xw_1 + yw_2 + zw_3$ et le membre de droite tend vers 0, donc une condition nécessaire est $xw_1 + yw_2 + zw_3 = 0$.

Montrons que cette condition est suffisante. On se donne un élément (w_1, w_2, w_3) tel que $u.w = 0$, $u = (x, y, z)$. On considère alors une suite quelconque w^n qui tend vers w (c'est toujours possible à définir, ce serait-ce qu'en prenant $w + \frac{1}{n}e$, où e est un vecteur fixe quelconque). On sait alors que $x.w_n$ tend vers 0. On construit alors $\tilde{w}^n = w^n - 2|u.w^n|(x, y, z)$ (ceci veut dire $\tilde{w}_1^n = w_n^1 - 2|xw_n^1 + yw_n^2 + zw_n^3|x, \tilde{w}_2^n = w_n^2 - 2|xw_n^1 + yw_n^2 + zw_n^3|y$). Il en découle que $\tilde{w}^n$ tend vers w car w^n tend vers w et $u.w^n$ tend vers 0. De plus, $\tilde{w}^n.(x, y, z) = \tilde{w}^n.u = w^n.u - 2|w^n.u| \leq 0$. On construit alors $e_n = -\frac{2u\tilde{w}^n}{||\tilde{w}^n||^2} \geq 0$. La suite $(e_n, \tilde{w}^n)$ vérifie les conditions de la définition, donc $(w_1, w_2, w_3) \in K(u)$ (exemple 1).

Exemple1

Si $K = \{(x, y, z), x^2 + y^2 + z^2 \leq 1\}$, alors $K(u) = \dot{K}(u) = \mathbb{R}^3$ pour $u = (x, y, z)$ tel que $x^2 + y^2 + z^2 < 1$ (en effet, il suffit, pour toute direction non nulle w, de considérer $u + \frac{1}{2}(1 - ||u||)\frac{w}{||w||}$, qui est dans la sphère unité, donc on vérifie que pour $\epsilon_0 = \frac{1}{2}\frac{(1-||u||)}{||w||}$ et $\epsilon < \epsilon_0$, $u + \epsilon w$ est dans la sphère). Pour un point du bord $u^2 = 1$, on aboutit, en divisant par e_n ou par ϵ, à l'inégalité

$$u.w \leq -\frac{\epsilon}{2}||w||^2, u.w_n \leq \frac{e_n}{2}||w^n||^2$$

ce qui aboutit aux relations $\dot{K}(u) = \{u.w < 0\}$ et $K(u) = \{u.w \leq 0\}$.

Nous généralisons ces expressions. Commençons par une contrainte égalité $F(v) = 0$ (exemple 1). Ainsi w est une direction admissible pour u si il existe une suite w_n tendant vers w et une suite $e_n > 0$ tendant vers 0 telles que $F(u + e_n w_n) = 0$. Alors on en déduit, en supposant que F est différentiable

$$F(u) + (F'(u), e_n w_n) + o(e_n|w_n|) = 0.$$

Faisant tendre e_n vers 0 après avoir utilisé $F(u) = 0$ et avoir divisé par e_n conduit à $(F'(u), w) = 0$.

Réciproquement, supposons $(F'(u), w) = 0$. On introduit la fonction $\phi(\lambda, \varepsilon) = \frac{1}{\varepsilon}F(u + \varepsilon w + \varepsilon\lambda F'(u))$, $\phi(\lambda, 0) = (F'(u), w + \lambda F'(u))$. On a

$$\frac{\phi(\lambda + h, \varepsilon) - \phi(\lambda, \varepsilon)}{h} = \frac{1}{\varepsilon h}(F(u + \varepsilon w + \varepsilon\lambda F'(u) + \varepsilon h F'(u)) - F(u + \varepsilon w + \varepsilon\lambda F'(u)))$$

donc

$$\phi'_\lambda(\lambda, \varepsilon) = (F'(u + \varepsilon w + \varepsilon\lambda F'(u)), F'(u)).$$

On suppose que F' est Lipschitz et que $F'(u) \neq 0$. On souhaite trouver $\lambda(\varepsilon)$ tel que $\phi(\lambda(\varepsilon), \varepsilon) = 0$. On écrit l'équation sous la forme

$$\phi(\lambda, \varepsilon) - \phi(0, \varepsilon) = -\phi(0, \varepsilon)$$

De l'égalité $(F'(u), w) = 0$, on déduit $\phi(0, \varepsilon) = o(1)$. De la relation $F'(u) \neq 0$, on tire que la dérivée de $\phi(\lambda, \varepsilon) - \phi(0, \varepsilon)$ est $||F'(u)||^2 > 0$, et, de plus, $\phi(0, 0) = 0$. On est dans le cas d'application du théorème des fonctions implicites et il existe ε_0 et une fonction continue $\lambda(\varepsilon)$ telle que, pour $\varepsilon < \varepsilon_0$ on ait

$$\phi(\lambda(\varepsilon), \varepsilon) - \phi(0, \varepsilon) = -\phi(0, \varepsilon).$$

La fonction $\lambda(\varepsilon)$ tend vers 0 lorsque ε tend vers 0. On peut aussi voir ce résultat en écrivant l'équation sous la forme

$$\lambda \int_0^1 \phi'_\lambda(\lambda x, \varepsilon) dx = -\phi(0, \varepsilon)$$

ce qui donne, par approximation de la dérivée première

$$\lambda[||F'(u)||^2 + O(\varepsilon)] = -\phi(0, \varepsilon)$$

soit

$$\lambda = -\frac{\phi(0, \varepsilon)}{||F'(u)||^2}(1 + O(\varepsilon)),$$

d'où une expression de $\lambda(\varepsilon)$ (dont on a montré l'existence et l'unicité ci-dessus). Ainsi on a trouvé $w_\varepsilon = w + \lambda_0 F'(u)$ tel que $F(u + \varepsilon w_\varepsilon) = 0$ et $w_\varepsilon \to w$. La direction w est une direction admissible. Lorsque $F'(u) = 0$, w est quelconque, mais cela n'assure pas l'existence d'un w non nul qui soit une direction admissible. Par exemple, $F(x) = x^2$ conduit, dans la définition, à écrire le cône des directions admissibles à $\{0\}$ dans $\mathbb{R}$, qui correspond à $\{0\}$, car dans ce cas $0 + e_n w_n = 0$ ce qui implique $w_n = 0$, et non pas tout l'axe réel.

Lemme 2.1 *Le cône $K(u)$ associé à u tel que $F(u) = 0$ est, dans le cas $F'(u) \neq 0$ l'ensemble des $w \in V$ tels que $(F'(u), w) = 0$.*

On en déduit la représentation suivante

Définition 2.5 *Soit $K = \{u, F_1(u) = 0, F_2(u) = 0, ...F_m(u) = 0\}$. Lorsque les vecteurs $(F'_1(u), F'_2(u), ..F'_m(u))$ sont linéairement indépendants, on dit que les contraintes sont régulières en u.*

Lemme 2.2 *Si les contraintes sont régulières en u, alors* $K(u) = \{w \in V, (F_i'(u), w) = 0 \forall i = 1..m\}$.

L'implication directe est facile. L'implication réciproque est une conséquence du théorème des fonctions implicites matriciel. On choisit donc, pour un w tel que $(F_j'(u), w) = 0$ pour tout j, de regarder une perturbation de $u + \varepsilon w$ et de déterminer $(\mu_1, .., \mu_m)$ tels que

$$\forall j F_j(u + \varepsilon w + \sum_{k=1}^{k=m} \varepsilon \mu_k F_k'(u)) = 0.$$

On regarde alors ce système comme une application de $\mathbb{R}^M$ dans lui même. Le jacobien de cette application est, pour $\varepsilon = 0$, la matrice des produits scalaires $(F_j'(u), F_k'(u))$. La famille est libre, donc cette matrice est inversible et cette propriété est vraie pour $\varepsilon < \varepsilon_0$ lorsque les μ_j appartiennent à un compact. On applique alors le théorème des fonctions implicites de $\mathbb{R}^M$ dans $\mathbb{R}^M$ et on conclut. Lorsque les vecteurs $F_i'(u)$ ne forment pas une famille libre, on a le même problème que précédemment dans le cas $F'(u) = 0$. On ne peut pas assurer l'existence de directions admissibles. Par exemple, si on considère l'ensemble $x^2 + y^2 = 1, x^3 + y^3 = 1$ admet comme solutions $(1, 0), (0, 1)$ et ces points sont isolés donc leurs directions admissibles sont réduites à $\{0\}$. On peut aussi considérer l'exemple d'une sphère S et d'un de ses plans tangents P. Au point d'intersection, les deux vecteurs $F_i'(u)$ sont égaux à la direction normale à la sphère, et l'intersection est réduite au point.

Lorsque le cône $K(u)$ est facile à évaluer, le théorème 2.4 permet de calculer ce que l'on appelle les multiplicateurs de Lagrange.

Théorème 2.5 *Pour que u tel que* $(F_j'(u))_j$ *forme une famille libre (on dit que* **les contraintes** $F_j(v), 1 \leq j \leq m$ **sont régulières en** *u), soit solution de (2.2.4), il* **faut** *qu'il existe m réels* $\lambda_1, ...\lambda m$ *tels que*

$$J'(u) + \lambda_1 F_1'(u) + \lambda_2 F_2'(u) + ... + \lambda_m F_m'(u) = 0$$

Preuve La partie difficile de la preuve a été faite. En effet, si u est régulier, on identifie aisément le cône $K(u)$ des directions admissibles; c'est l'espace vectoriel orthogonal à l'espace vectoriel F engendré par la famille $(F_j'(u))_{j=1..m}$. Le théorème (2.4) se traduit alors par

$$\forall w \in K(u), (J'(u), w) \geq 0.$$

Comme $K(u)$ est un espace vectoriel, $-w \in K(u)$ lorsque $w \in K(u)$, ce qui se traduit par

$$\forall w \in K(u), (J'(u), w) = 0.$$

Ainsi $J'(u)$ est dans l'espace vectoriel orthogonal à $F^{\perp}$, c'est-à-dire F, et l'égalité du théorème est vraie.

On peut aussi le vérifier comme suit. Il existe des scalaires λ_j et un vecteur r, orthogonal à tous les $F_j'(u)$, tels que $J'(u) = -\sum_{j=1}^m \lambda_j F_j'(u) + r$. Alors $r \in K(u)$ et $(J'(u), r) = 0$, ce qui s'écrit $(r, r) = 0$ soit $r = 0$.

Un travail identique peut être fait pour les contraintes inégalités. On suppose donc $F(u) \leq 0$ une contrainte donnée de V dans $\mathbb{R}$. Soit $u \in K$, vérifiant ainsi $F(u) \leq 0$. Une direction w de $K(u)$ est alors telle que $F(u + \varepsilon w) \leq 0$ pour ε assez petit, soit $F(u) + \varepsilon(F'(u), w) + o(\varepsilon w) \leq 0$.

Deux cas sont alors à envisager:

• soit $F(u) < 0$, auquel cas, dès que ε est assez petit, tout élément w est admissible. La contrainte $F(u) \leq 0$ n'ajoute donc pas de condition dans le théorème 2.4, la condition nécessaire est donc l'égalité d'Euler $J'(u) = 0$ qui provient de $(J'(u), w) \geq 0 \forall w \in K(u)$. On dit pour cette raison que la contrainte est inactive (on dira aussi de temps en temps insaturée).

• soit $F(u) = 0$, auquel cas, comme $\varepsilon > 0$, il faut et il suffit, dans le cas $F'(u) \neq 0$, que $(F'(u), w) \leq 0$.

On note tout de suite que si $(F'(u), w) < 0$, alors il est clair que, pour ε assez petit, $F(u + \varepsilon w) = \varepsilon(F'(u), w) + o(\varepsilon) < 0$. Le problème se pose lorsque $(F'(u), w) = 0$ pour trouver un élément de l'espace des contraintes. On doit donc introduire une notion de plus grande régularité des contraintes.

Par exemple la condition $F'(u) \neq 0$ est assurée lorsqu'il existe w tel que $(F'(u), w) < 0$.

D'autre part, lorsqu'il y a plusieurs contraintes inégalités, on veut pouvoir montrer que l'ensemble des directions admissibles n'est pas vide.

Pour cela, il faut trouver un w_0 tels que, pour toutes les contraintes F_j saturées, on a $(F'_j(u), w_0) \leq 0$.

Cette condition n'est pas assez restrictive. En effet, la définition des directions admissibles w conduit à la relation $(F'_j(u), w) \leq 0$. En revanche, si on ne peut trouver un w_0 que dans le cas où il existe un couple (j_1, j_2) tels que $(F'_{j_1}(u), w_0) = (F'_{j_2}(u), w_0) = 0$, on pourrait se trouver dans la situation où les deux hypersurfaces $F_{j_1} \leq 0$ et $F_{j_2} \leq 0$ sont tangentes en u, de vecteur normal w_0, et (par exemple) de concavité stricte opposée (exemple 2):

Exemple 2

Dans ce cas, l'intersection des contraintes $F_{j_1} \leq 0$ et $F_{j_2} \leq 0$ est réduite à $\{u\}$, et on ne peut plus parler de direction admissible.

Une condition pour que l'ensemble des directions admissibles soit non vide est alors la condition:

Il existe w_0 tel que, $\forall j, (F_j(u), w_0) < 0$.

Cette condition est peu utilisable, car trop restrictive; en particulier une contrainte affine pourra donner une direction admissible avec uniquement l'égalité. On utilise alors plutôt la condition suivante:

Il existe w_0 tel que $\forall j, (F_j(u), w_0) < 0$ (contraintes non affines) et $(F'_j(u), w_0) = 0$ si la contrainte est affine, car on sait que dans ce cas l'intersection entre le demi

hyperplan défini par la contrainte affine et les autres conditions est non vide.

Enfin, on élimine grâce à cela la condition d'indépendance des $(F_j'(u))$ que l'on avait utilisé pour caractériser les directions admissibles (qui est non pas automatique, mais inutile: voir exemple 3). Exemple 3

Cette étude induit une définition de contraintes qualifiées, qui est une hypothèse technique mais qui est l'hypothèse la plus classique en théorie des multiplicateurs de Lagrange:

Définition 2.6 *Soit* $K = \{u, F_j(u) \leq 0, j = 1..m\}$.

• *On dit qu'une contrainte* F_j *est active si* $F_j(u) = 0$, *et elle est inactive si* $F_j(u) < 0$. *On note* $I(u)$ *l'ensemble des indices des contraintes actives.*

• *On dit que l'ensemble des contraintes* (F_j) *est qualifié si il existe* $w_0 \in V$ *tel que pour tout* $j \in I(u)$ *(pour les contraintes actives),* $(F_i'(u), w_0) \leq 0$, *et* $(F_j'(u), w_0) = 0$ *uniquement pour* F_j *affine.*

Commençons par ranger les contraintes actives affines pour $j \in I'(u)$. On prend w_0 dans l'orthogonal de l'espace vectoriel F_0 engendré par les $F_j'(u)$, $j \in I'(u)$, qui est indépendant de u. Il suffit alors de voir que, pour tout $w_0 \in F_0$ et pour tout $j \in I'(u)$, on a $F_j(u + w_0) = F_j(u) = 0$. Il suffit alors de regarder, pour les autres conditions, $(j \in I(u) - I'(u))$, $(F_j'(u), w_0)$ et $K(u)$ est non vide lorsque w_0 existe.

Une notion moins restrictive mais plus abstraite est la notion de **contraintes qualifiables**:

Définition 2.7 *On dit que les contraintes inégalités* $\{F_j(u) \leq 0\}$ *sont qualifiables en* u *si*

$$K(u) = \{w, (F_j'(u), w) \leq 0 \text{ pour } j \in I(u)\}.$$

On a alors le lemme suivant

Lemme 2.3 *On suppose que les contraintes* $F_j, 1 \leq j \leq m$, *sont qualifiées en* $u \in K$. *Alors elles sont qualifiables en* u.

La preuve de ce lemme s'appuie sur l'existence de w_0 pour la démonstration de la réciproque; en effet l'implication directe est une conséquence de la dérivabilité et du fait de faire tendre e_n vers 0.

On considère donc w dans $\{w \in V, (F_j'(u), w) \leq 0 \forall j \in I(u)\}$, et on forme, pour tout ε et pour tout δ positif fixé $u + \varepsilon(w + \delta w_0)$. Pour ε assez petit, par continuité de F_j pour $j \notin I(u)$, $F_j(u + \varepsilon(w + \delta w_0)) < 0$. D'autre part, pour $j \in I'(u)$, on a $F_j(u + \varepsilon(w + \delta w_0)) = F_j(u) + \varepsilon(F_j'(u), w + \delta w_0)) = \varepsilon(F_j'(u), w) \leq 0$. Enfin, pour $j \in I(u) - I'(u)$, il vient $F_j(u + \varepsilon(w + \delta w_0)) = F_j(u) + \varepsilon(F_j'(u), w + \delta w_0)) + o(\varepsilon)$. Comme $F_j(u) = 0$, $(F_i'(u), w_0) < 0$ et $(F_j'(u), w) \leq 0$, on trouve

$$F_j(u+\varepsilon(w+\delta w_0)) \leq \delta\varepsilon(F_j'(u),w_0)+o(\varepsilon).$$

Le second membre est strictement négatif lorsque ε tend vers 0, car $(F_j'(u),w_0)$ et $o(\varepsilon)/\varepsilon$ tend vers 0. Le lemme est démontré.

Théorème 2.6 *Sous l'hypothèse que J est dérivable, que les F_j sont dérivables, et que, en u, les contraintes sont qualifiables, pour que u soit une solution de (2.2.4), il* **faut** *qu'il existe $\lambda_1,...\lambda_m \geq 0$ tels que $\lambda_j = 0$ pour $j \in \{1,..,m\} - I(u)$ et*

$$J'(u) + \sum_{i=1}^{i=m} \lambda_i F_i'(u) = 0.$$

Remarquons que si on considère l'ensemble des contraintes égalités comme l'ensemble de toutes les contraintes inégalités ($F_j(u) = 0$, $1 \leq j \leq m$ équivaut à $F_j(u) \leq 0, -F_j(u) \leq 0$), toutes les contraintes sont actives, car si u est tel que $F_j(u) < 0$, alors $-F_j(u) > 0$ donc (bien sûr) u n'est pas dans l'ensemble!!. On écrit la condition sur les multiplicateurs de Lagrange $\lambda_j \geq 0, \mu_j \geq 0$, $J'(u) + \sum_{j=1}^{j=m} \lambda_j F_j'(u) + \sum_{j=1}^{j=m} \mu_j(-F_j'(u)) = 0$, $J'(u) + \sum_j p_j F_j'(u) = 0$ avec $p_j = \lambda_j - \mu_j \in \mathbb{R}$.

Le théorème 2.6 est une conséquence simple du lemme suivant, dit de Farkas, et de la représentation des directions admissibles du lemme 2.3. On applique alors le théorème 2.4 pour en déduire l'existence des multiplicateurs de Lagrange positifs.

Lemme 2.4 *(Farkas)*

Soit $\mathcal{K}$ l'intersection des demi hyperplans orthogonaux à $a_j, 1 \leq j \leq m$, $\mathcal{K} = \{(a_j,v) \leq 0 \forall j\}$.

$$\forall v \in \mathcal{K}, (p,v) \geq 0 \Rightarrow \exists(\lambda_1,...\lambda_m) \in (\mathbb{R}_+)^m, v = -\sum \lambda_i a_i.$$

On définit $\mathcal{B} = \{-\sum \lambda_i a_i, 1 \leq i \leq M\}$. Nous démontrerons que $\mathcal{B}$ est un convexe fermé. Admettons le pour l'instant. On peut alors appliquer la notion de projection sur un convexe fermé non vide. On suppose donc que p_0 vérifie les hypothèses du lemme de Farkas et que p_0 n'appartient pas à $\mathcal{B}$. On montre que la projection $\tilde{p}$ de p_0 sur $\mathcal{B}$ est égale à p_0, d'où contradiction. On trouve, de $||p_0-\tilde{p}||^2 \geq ||p_0-w||^2$, $w \in \mathcal{B}$, que $\forall w \in \mathcal{B}, (\tilde{p}-p_0, w-\tilde{p}) \leq 0$. Dans cette inégalité, on choisit alors $w = -\lambda a_i$ et on fait tendre λ vers $+\infty$. Il reste donc $(a_i, p_0-\tilde{p}) \geq 0$ pour tout i. Ceci implique que $\tilde{p}-p_0$ est dans $\mathcal{K}$. De l'inégalité $0 \leq (p_0, \tilde{p}-p_0) = -|p_0-\tilde{p}|^2 + (p_0-\tilde{p}, 0-\tilde{p}) \leq -|p_0-\tilde{p}|^2$ (car $0 \in \mathcal{B}$) on déduit que $p_0 = \tilde{p}$. On a montré que $p_0 \in \mathcal{B}$, contradiction.

Il reste à démontrer que $\mathcal{B}$ est fermé convexe. Il est convexe de manière évidente (pm considère $0 \leq \mu \leq 1$, alors $\mu\lambda_i^1 + (1-\mu)\lambda_i^2 \geq 0$, et donc il existe une représentation de $\mu v_1 + (1-\mu)v_2$ qui soit une combinaison linéaire à coefficients négatifs). En revanche le caractère fermé est plus difficile à obtenir.

Si la famille (a_i) est libre, la matrice $(a_i.a_j)$ est symétrique définie positive. On note $||a||$ le max des normes des a_i et α la plus petite valeur propre de la matrice. On obtient $\sum \lambda_i a_i.a_j = -v.a_j$, donc il vient $|\lambda_i| \leq \alpha^{-1}||v||||a||$. Si la suite v_n d'éléments de $\mathcal{B}$ converge vers v, on peut identifier les λ_i^n associés, et les suites λ_i^n sont bornées. Quitte à faire des extractions de suite en cascade, il existe une sous-suite convergente $\lambda_i^{\psi(n)}$, qui converge vers des valeurs positives λ_i, donc $v = -\sum \lambda_i a_i$. La limite est donc dans $\mathcal{B}$.

Deuxième cas, si la famille est linéairement dépendante, il existe $\mu_1, ..\mu_m$ tels que $\sum \mu_i a_i = 0$ (avec au moins un des coefficients qui est positif), et donc un élément de $\mathcal{B}$ s'écrit $v = -\sum(\lambda_i + t\mu_i)a_i$. Il faut montrer que pour une valeur de $t \leq 0$, cette somme est une combinaison à coefficients positifs de $m-1$ termes, et on se sera ramené à une famille avec moins d'éléments pour tout t. Pour $t = 0$, tous les coefficients sont positifs ou nuls, donc de deux choses l'une: ou bien $\mu_{i_1} \leq 0$, auquel cas $\mu_{i_1} t \geq 0$ et le coefficient correspondant ne s'annulera pas si $\lambda_{i_1} \neq 0$, ou bien $\mu_{i_1} > 0$, ce qui implique que $t = -\frac{\lambda_{i_1}}{\mu_{i_1}}$ est une valeur où le coefficient s'annule. On prend alors $t_0 = \min_{i,\mu_i>0} \frac{\lambda_i}{\mu_i}$ et la combinaison précédente a un coefficient qui s'annule pour $t = -t_0$. Cette construction est valable pour chaque élément de $\mathcal{B}$.

On considère alors une suite x^n d'éléments de $\mathcal{B}$, suite de Cauchy dans l'espace engendré par les a_i, espace vectoriel de dimension finie. Elle s'écrit $-\sum \lambda_i^n a_i$. Par la construction ci-dessus, pour chaque n, il existe $i(n)$ tel que $-\sum \lambda_i^n a_i = -\sum_{i\neq i(n)} \tilde{\lambda}_i^n a_i$. On a donc enlevé chaque fois un élément de la famille (a_i). On note $I_i = \{n, i(n) = i\}$. L'union des I_i est l'ensemble des entiers naturels, donc il existe au moins un i_0 tel que I_i est infini, soit $I_i = \{\phi(m), m = 0, 1..+\infty\}$. La suite extraite $x^{\phi(n)} = -\sum_{i\neq i_0} \tilde{\lambda}_i^{\phi(m)} a_i$ est une suite qui correspond à la famille $(a_i)_{i\neq i_0}$. Si cette famille est libre, on s'est ramené au cas précédent, et la suite extraite $x^{\phi(n)}$ converge vers un élément de $\mathcal{B}$. Comme la suite est de Cauchy, elle converge vers x et la limite de toute suite extraite est x.

Si cette famille est liée, on reprend l'argument avec la suite $x^{\phi(n)}$. Comme la famille n'est pas identiquement nulle (sinon $\mathcal{B}$ est réduit à $\{0\}$ et on n'a rien à démontrer), alors au bout d'un nombre fini d'itérations, on aboutit à une famille libre (a_j) et la démonstration est finie puisque la limite est dans $\mathcal{B}$ pour cette suite extraite.

On a donc montré que $\mathcal{B}$ est fermé, donc on peut utiliser le théorème de projection sur un convexe fermé.

Remarque: inégalités de Hardy. On peut obtenir en exercice l'inégalité

$$(\frac{1}{n}\sum_{i=1}^{i=n} |x_i|^p)^{\frac{1}{p}} \leq (\frac{1}{n}\sum_{i=1}^{i=n} |x_i|^q)^{\frac{1}{q}}, q \geq p$$

En effet, on suppose la contrainte $\sum |x_i|^q = 1$ et on cherche à minimiser $J(x) = \sum |x_i|^p$. On écrit, avec le multiplicateur de Lagrange λ, $y_i = |x_i|\ py_i^{p-1} + \lambda q y_i^{q-1} = 0$, sous la contrainte $\sum y_i^q = 1$. On trouve alors $y_i^{q-p} = -\frac{p}{\lambda q}$ ou $y_i = 0$. Soit k le nombre de valeurs de y_i non nulles. Alors elles sont égales, donc $y_i = (\frac{1}{k})^{\frac{1}{q}}$, ce qui donne $J(y) = k(\frac{1}{k})^{\frac{p}{q}} = k^{\frac{p-q}{q}}$. Lorsque $q < p$, la plus petite valeur est atteinte pour $k = 1$, et le minimum est atteint lorsque l'un seulement est non nul. Lorsque $q \geq p$, la plus petite valeur est atteinte lorsque tous les y_i sont égaux, et la plus petite valeur de J est $n^{\frac{p-q}{q}}$. On en déduit $\sum y_i^p \geq n^{\frac{q-p}{q}}, \sum y_i^p = 1$ ainsi, en notant $z_i = \frac{y_i}{(\sum y_i^p)^{\frac{1}{p}}}$, tel que $\sum z_i^p = 1$, on a le résultat.

Chapter 3

Calcul des variations, lagrangien, hamiltonien.

3.1 Introduction et un peu d'histoire

Dans cette section, qui est à l'origine des théories des extrema et de calcul des variations, on considère des fonctions d'un intervalle de $\mathbb{R}$ dans un espace de Hilbert H. Comme dans l'exemple 8 de l'introduction, il peut s'agir de la trajectoire d'une particule, le paramètre important variant dans un intervalle de $\mathbb{R}$ étant le temps. Il peut aussi s'agir de l'équation d'une courbe dans le plan Oxy, sous la forme $y = y(x)$. Les notations employées sont extrêmement variées, et nous les mettrons en relation. Alors on minimise un critère J, qui s'appelle une **intégrale d'action**, sous une contrainte, qui peut être les points origine et destination de la courbe, ou une contrainte de type commande sous la forme $\inf J(x, u)$ où x est solution de $\dot{x} = f(x, u, t)$. Il peut s'agir aussi d'une contrainte intégrale, comme une contrainte sur la longueur de la courbe $y = y(x)$: $l = \int_{x_1}^{x_2} (1 + (y')^2)^{\frac{1}{2}} dx$. Les résultats de ce chapitre sont très anciens; ils forment la base du calcul classique des variations. Les méthodes que nous verrons montrent en quel sens le mot "variations" doit être entendu.

En 1696, Leibniz a résolu le problème de la **brachistochrone**. Il faut trouver la courbe qui réalise le minimum du temps de parcours entre deux points (x_1, y_1) et (x_2, y_2) dans un même plan vertical lorsque le point matériel glissant est soumis à la force de pesanteur. Ce problème avait été posé par J. Bernoulli[1]. Ce problème peut être facilement résolu car les contraintes peuvent être intégrées à une intégrale première. Cependant, après sa publication, des problèmes plus généraux ont été énoncés sous le nom général de problèmes isopérimétriques, et on peut les résumer en "quelles sont les courbes de longueur donnée qui entoure la plus grande surface?". Le premier de ces problèmes est légendaire, comme nous l'avons rappelé dans l'exemple 11 (Problème de Didon). En effet, Didon, descendante des Troyens et fuyant sa cité après la chute de Troie, a demandé à Jarbas, roi des terres africaines, la terre que pouvait recouvrir une peau d'un bœuf. Ce roi, ne pensant pas à une quelconque astuce, accepta et Didon découpa la peau d'un bœuf en de fines lanières, qu'elle attacha entre elles (et si on suppose que la largeur de la lanière était d'un millimètre, la longueur obtenue était donc de $1000S$). Elle forma la plus grande surface enclose par cette lanière s'appuyant sur la côte méditerranéenne, et fonda Carthage, la grande rivale de

[1] Problema novum, ad cujus solitionem mathematici invitantur

Rome[2].

J. Bernoulli demanda à un de ses élèves, le mathématicien L. Euler, de résoudre ce problème, ce qu'il fit en 1744[3], par une méthode de série, suivi en 1755 par Lagrange, qui inventa la méthode classique de calcul des variations. Continuant ses travaux, Lagrange introduisit ses multiplicateurs en 1797.

3.2 Problèmes isopérimétriques

3.2.1 Egalité d'Euler-Lagrange

On considère ici $y(x) \in C^1([x_1, x_2])$, $y(x_1) = y_1, y(x_2) = y_2$ et on cherche à minimiser:

$$I(y) = \int_{x_1}^{x_2} f(x, y, y')dx$$

où f est une fonction de classe $C^2(\mathbb{R} \times H \times H)$.

On suppose connue une famille de fonctions $y(x, \varepsilon)$ telle que $y(x_1, \varepsilon) = y_1, y(x_2, \varepsilon) = y_2$ et $y(x, 0) = y_0(x)$, solution à trouver du problème de minimisation. On suppose $y \in C^2([x_1, x_2] \times [0, \varepsilon_0])$. On introduit la **première variation de** y:

$$\eta(x, \varepsilon) = \frac{\partial y}{\partial \varepsilon}(x, \varepsilon)$$

(ce qui explique le nom de calcul des variations). On se ramène donc à une fonction de ε:

$$J(\varepsilon) = I(y(., \varepsilon)).$$

Une condition nécessaire pour que y_0 soit une solution du problème de minimisation est la suivante:

$$J'(0) = 0.$$

Par application du théorème de dérivation sous le signe intégral, et en remarquant que comme y est de classe C^2, alors $\frac{\partial}{\partial \varepsilon}(y'(x, \varepsilon)) = \frac{\partial}{\partial x}(\frac{\partial y}{\partial \varepsilon}(x, \varepsilon)) = \eta'(x, \varepsilon)$, on obtient

$$\int_{x_1}^{x_2} (\partial_y f(x, y_0(x), y_0'(x)).\eta(x, 0) + \partial_{y'} f(x, y_0(x), y_0'(x)).\eta'(x, 0))dx = 0. \qquad (3.2.1)$$

Notons dans cette égalité comme dans l'écriture de f que l'on a considéré le terme y' comme une variable indépendante de y et non comme la dérivée de y par rapport à x.

On utilise alors la relation $y(x_1, \varepsilon) = y_1$, de sorte que, en dérivant par rapport à ε, $\eta(x_1, \varepsilon) = 0$. De même, $\eta(x_2, \varepsilon) = 0$. On peut alors utiliser ces conditions de bord pour effectuer une intégration par parties:

$$\int_{x_1}^{x_2} \partial_{y'} f(x, y_0(x), y_0'(x)).\eta'(x, 0)dx = -\int_{x_1}^{x_2} \frac{d}{dx}(\partial_{y'} f(x, y_0(x), y_0'(x))).\eta(x, 0)dx.$$

[2]Delenda Cartago est! (Caton)

[3]Methodus inveniendi lineas curvas maximi minimive proprietate gaudentes, sive solutio problematis isoperimetrici latissimo sensu accepti

En écrivant l'égalité (3.2.1) et en vérifiant qu'elle est vraie quelle que soit la fonction $\eta(x,0)$ nulle en x_1 et en x_2 (pour s'en convaincre, il suffit d'écrire $y(x,\varepsilon) = y_0(x) + \varepsilon g(x)$, où g est nulle aux deux bouts), on trouve l'équation d'Euler-Lagrange:

$$\frac{d}{dx}(\frac{\partial f}{\partial y'}(x, y_0(x), y_0'(x))) = \frac{\partial f}{\partial y}(x, y_0(x), y_0'(x)). \tag{3.2.2}$$

Bien sûr, cette équation s'obtient facilement en utilisant le théorème 2.4 démontré dans le chapitre 2. Nous allons l'établir de deux façons distinctes. Avant cela, cependant, donnons un résultat important lorsque f ne dépend que des variables de position y et y':

Lemme 3.1 *Lorsque f ne dépend pas de x, une solution des équations d'Euler vérifie l'égalité suivante:*

$$\frac{d}{dx}\left(y_0'\partial_{y'}f(y_0,y_0')) - f(y_0,y_0')\right) = 0.$$

Cette égalité donne une intégrale première.

La démonstration intuitive la plus facile est de voir comment varie l'action lorsque l'intégrale d'action est minimale, soit

$$\begin{aligned}\tfrac{d}{dx}(f(y_0,y_0')) &= \partial_y f(y_0,y_0')y_0' + \partial_{y'}f(y_0,y_0')y_0'' \\ &= \tfrac{d}{dx}(\partial_{y'}f(y_0,y_0'))y_0' + \partial_{y'}f(y_0,y_0')y_0'' \\ &= \tfrac{d}{dx}(y_0'\partial_{y'}f(y_0,y_0')).\end{aligned}$$

3.2.2 Dérivée de Fréchet et de Gâteaux, inégalité d'Euler-Lagrange

Dans un premier temps, en vue d'appliquer le théorème 2.4 , nous allons calculer la dérivée de Fréchet (qui existe puisque f est de classe C^2) de J. En fait, nous allons calculer deux objets:

- le produit scalaire $(J'(y_0), w)$ pour $w \in K(y_0)$,
- la distribution $J'(y_0)$.

Le cône des directions admissibles $K(y_0) \subset H^1(]x_1,x_2[)$ est l'ensemble des w tels qu'il existe w_n et $e_n > 0$ tels que $e_n \to 0$ et $w_n \to w$ et $(y_0 + e_n w_n)$ est dans l'espace des contraintes, soit $y_0(x_1) + e_n w_n(x_1) = y_1 = y_0(x_1)$ et $y_0(x_2) + e_n w_n(x_2) = y_2 = y_0(x_2)$. Comme $e_n > 0$, on trouve que $w_n(x_1) = w_n(x_2) = 0$. Comme les fonctions $H^1(]x_1,x_2[)$ sont continues aux bords x_1 et x_2, et que l'application trace est continue, on en déduit que $w(x_1) = w(x_2) = 0$. Réciproquement, si $w(x_1) = w(x_2) = 0$, on construit $y_0 + \frac{1}{n}w$ qui vérifie bien les contraintes.

$$K(y_0) = H_0^1([x_1,x_2]).$$

Alors le calcul de $(J'(y_0), w)$, qui est le calcul de la limite

$$\lim_{\varepsilon\to 0}\frac{J(y_0+\varepsilon w) - J(y_0)}{\varepsilon}$$

conduit exactement à

$$\forall w \in H^1([x_0,x_1]), \int_{x_1}^{x_2}(\frac{\partial f}{\partial y}(x,y_0,y_0') - \frac{d}{dx}(\frac{\partial f}{\partial y'}(x,y_0(x),y_0'(x))))w(x)dx \geq 0$$

Le cône des directions admissibles est un espace vectoriel, donc cette inégalité devient une égalité, et cette égalité entraine l'équation d'Euler-Lagrange.

D'autre part, on vérifie aisément que, pour $w \in H^1([x_1, x_2])$, après intégration par parties, on trouve

$$\begin{aligned}(J'(y_0), w) &= \int_{x_1}^{x_2} (\frac{\partial f}{\partial y}(x, y_0, y_0') - \frac{d}{dx}(\frac{\partial f}{\partial y'}(x, y_0(x), y_0'(x)))) w(x) dx \\ &+ \frac{\partial f}{\partial y'}(x_2, y_0(x_2), y_0'(x_2)) w(x_2) - \frac{\partial f}{\partial y'}(x_1, y_0(x_1), y_0'(x_1)) w(x_1).\end{aligned}$$

En utilisant la distribution de Dirac $(\delta_{x_1}, w) = w(x_1)$, on trouve

$$\begin{aligned}J'(y_0) &= \frac{\partial f}{\partial y}(x, y_0, y_0') - \frac{d}{dx}[\frac{\partial f}{\partial y'}(x, y_0(x), y_0'(x))] \\ &+ \frac{\partial f}{\partial y'}(x_2, y_0(x_2), y_0'(x_2)) \delta_{x_2} - \frac{\partial f}{\partial y'}(x_1, y_0(x_1), y_0'(x_1)) \delta_{x_1}.\end{aligned}$$

L'emploi des multiplicateurs de Lagrange pour des contraintes égalités, qui sont respectivement $F_1(y) = y(x_1) - y_1$ et $F_2(y) = y(x_2) - y_2$, ce qui donne $F_1'(y_0) = \delta_{x_1}$ et $F_2'(y_0) = \delta_{x_2}$, conduit à

$$J'(y_0) + \lambda_1 F_1'(y_0) + \lambda_2 F_2'(y_0) = 0$$

(notons ici le rétablissement des signes permettant d'avoir la même formulation pour les contraintes égalité et inégalité). On trouve alors l'équation d'Euler-Lagrange et les égalités, qui donnent les multiplicateurs de Lagrange:

$$\lambda_1 = \frac{\partial f}{\partial y'}(x_1, y_0(x_1), y_0'(x_1)), \lambda_2 = -\frac{\partial f}{\partial y'}(x_2, y_0(x_2), y_0'(x_2)). \tag{3.2.3}$$

Cette égalité aura une très jolie interprétation ci-dessous.

3.2.3 Egalité d'Euler-Lagrange pour une contrainte intégrale

Dans cette section, nous cherchons la solution de

$$\inf \int_{x_1}^{x_2} f(x, y, y') dx$$

sous les contraintes $\int_{x_1}^{x_2} g(x, y, y') dx = C$, $y(x_1) = y_1, y(x_2) = y_2$. Le cas modèle est le problème de Didon: $f(x, y, y') = y$ et $g(x, y, y') = (1 + (y')^2)^{\frac{1}{2}}$.

Une méthode usuelle classique consiste à employer une double variation, c'est-à-dire à tenir compte de la contrainte $\int_{x_1}^{x_2} g(x, y, y') dx = C$ en ajoutant à une première variation $y_0 + \varepsilon\eta_1$ une deuxième variation faite pour la contrebalancer:

$$y_0 + \varepsilon_1\eta_1 + \varepsilon_2\eta_2.$$

On introduit dans η_1 et η_2 les contraintes d'extrémité sous la forme $\eta_i(x_j) = 0$, $i, j = 1, 2$.On écrit alors que $I = \int_{x_1}^{x_2} f(x, y, y') dx$ et $C = \int_{x_1}^{x_2} g(x, y, y') dx$ sont deux fonctions de ε_1 et de ε_2, et on forme

$$\Delta(\varepsilon_1, \varepsilon_2) = \begin{pmatrix} \frac{\partial I}{\partial \varepsilon_1} & \frac{\partial I}{\partial \varepsilon_2} \\ \frac{\partial C}{\partial \varepsilon_1} & \frac{\partial C}{\partial \varepsilon_2} \end{pmatrix}.$$

Ce déterminant doit être nul pour y_0, solution, en $\varepsilon_1, \varepsilon_2$. En effet, si $\Delta \neq 0$, il est clair que le couple (I, C) ne stationne pas, alors que par hypotèse C est constant

donc stationne et I stationne (noter l'emploi du mot "stationne"). Par intégration par parties, on trouve

$$\Delta(\varepsilon_1,\varepsilon_2)=\begin{pmatrix} \int_{x_1}^{x_2}(\partial_y f-\frac{d}{dx}(\partial_{y'}f))\eta_1 dx & \int_{x_1}^{x_2}(\partial_y f-\frac{d}{dx}(\partial_{y'}f))\eta_2 dx \\ \int_{x_1}^{x_2}(\partial_y g-\frac{d}{dx}(\partial_{y'}g))\eta_1 dx & \int_{x_1}^{x_2}(\partial_y g-\frac{d}{dx}(\partial_{y'}g))\eta_2 dx \end{pmatrix}.$$

On note les deux réels $\lambda_1=\int_{x_1}^{x_2}(\partial_y f-\frac{d}{dx}(\partial_{y'}f))\eta_2 dx$ et $\lambda_2=\int_{x_1}^{x_2}(\partial_y g-\frac{d}{dx}(\partial_{y'}g))\eta_2 dx$. Si les deux réels sont nuls pour tous les choix de η_2, cela veut dire que f et g vérifient tous deux l'équation d'Euler. Nous verrons ce cas plus tard. Sinon, on note, pour un η_2 donné non nul, que, pour tout η_1:

$$\int_{x_1}^{x_2}[\lambda_2(\partial_y f-\frac{d}{dx}(\partial_{y'}f))-\lambda_1(\partial_y g-\frac{d}{dx}(\partial_{y'}g))]\eta_1 dx=0$$

ce qui donne l'existence d'un $h=f+\lambda g$ tel que h vérifie l'équation d'Euler. Lorsque f et g vérifient toutes deux l'équation d'Euler, alors cette équation est vérifiée quel que soit λ.

A l'évidence, cette méthode est celle que l'on emploie pour les multiplicateurs de Lagrange. On écrit ainsi l'existence de $\lambda,\lambda_1,\lambda_2$ tels que

$$J'(y_0)+\lambda C'(y_0)+\lambda_1 F_1'(y_0)+\lambda_2 F_2'(y_0)=0$$

(par application du théorème 2.5). Ainsi on trouve immédiatement, sans avoir besoin de considérer des variations qui se compensent:

$$\begin{array}{l}\partial_y f-\frac{d}{dx}(\partial_{y'}f)+\lambda(\partial_y g-\frac{d}{dx}(\partial_{y'}g))\\ +(\lambda_1-\partial_y f(x_1,y_1,y_0'(x_1))-\lambda\partial_y g(x_1,y_1,y_0'(x_1)))\delta_{x_1}\\ +(\lambda_2+\partial_y f(x_2,y_2,y_0'(x_2))+\lambda\partial_y g(x_2,y_2,y_0'(x_2)))\delta_{x_2}=0.\end{array}$$

L'écriture de l'équation d'Euler pour $-y+\lambda(1+(y')^2)^{\frac{1}{2}}$ donne

$$1=\frac{d}{dx}(\lambda\frac{y'}{(1+(y')^2)^{\frac{1}{2}}})$$

soit encore

$$\frac{y'}{(1+(y')^2)^{\frac{1}{2}}}=\frac{x}{\lambda}.$$

On obtient $y'=\pm\frac{x}{(\lambda^2-x^2)^{\frac{1}{2}}}$, dont la solution s'écrit

$$y(x)=y(x_1)\pm(\lambda^2-x^2)^{\frac{1}{2}}.$$

On suppose $y_1<y_2$, donc $y(x)=y_1+(\lambda^2-x_1^2)^{\frac{1}{2}}-(\lambda^2-x^2)^{\frac{1}{2}}$ car $y(x_1)=y_1$. On identifie λ en écrivant $y(x_2)=y_2$, soit $(\lambda-x_2^2)^{\frac{1}{2}}-(\lambda-x_1^2)^{\frac{1}{2}}=y_1-y_2$, ce qui permet de trouver les valeurs de $(\lambda^2-x_2^2)^{\frac{1}{2}}$ et $(\lambda^2-x_1^2)^{\frac{1}{2}}$. Lorsque $y_1=y_2=0$, on trouve un demi-cercle de rayon R et l'aire est πR^2, correspondant à $R=\frac{1000S}{2\pi}$.

3.2.4 Les problèmes de Bolza

On peut aussi vouloir inclure les contraintes dans la fonctionnelle à minimiser. La classe de problèmes correspondants s'écrit

$$\inf[\int_{x_1}^{x_2} f(x,y,y')dx + l(y(x_1),y(x_2))].$$

Il est clair que l'on obtient l'équation d'Euler:

$$\frac{d}{dx}(\frac{\partial f}{\partial y'}(x,y_0,y_0')) = \frac{\partial f}{\partial y}(x,y_0,y_0')$$

et les équations sur les contraintes

$$\partial_{u_1} l(y(x_1),y(x_2)) = \partial_{y'} f(x_1,y(x_1),y'(x_1))$$

$$\partial_{u_2} l(y(x_1),y(x_2)) = -\partial_{y'} f(x_2,y(x_2),y'(x_2)).$$

Prenons un exemple simple pour le problème de Bolza:

$$l_\varepsilon(u_1,u_2) = \frac{1}{\varepsilon}[(u_1-y_1)^2 + (u_2-y_2)^2].$$

Soit y_0 la solution du problème de minimisation de $J(y) = \int_{x_1}^{x_2} f(x,y,y')dx$ avec les contraintes $y(x_1) = y_1$, $y(x_2) = y_2$. Si $K = \{y, y(x_1) = y_1, y(x_2) = y_2\}$, alors, pour tout $y \in K$, $J(y) + l_\varepsilon(y(x_1)-y_1, y(x_2)-y_2) = J(y)$. On utilise alors

$$\inf_{y\in H^1} J(y) + l_\varepsilon(y(x_1)-y_1,y(x_2)-y_2) \leq \inf_{y\in K} J(y) = J(y_0).$$

On note la solution du problème de Bolza y_ε. Ainsi

$$J(y_\varepsilon) + l_\varepsilon(y_\varepsilon(x_1)-y_1, y_\varepsilon(x_2)-y_2) \leq J(y_0)$$

Ainsi $J(y_\varepsilon)$ est majoré. De plus, si on suppose f positive, $l_\varepsilon(y_\varepsilon(x_1)-y_1, y_\varepsilon(x_2)-y_2)$ est majorée par $J(y_0)$. On en déduit que la suite $(y_\varepsilon(x_j))$ converge vers y_j, $j = 1..2$. En revanche, on ne sait rien sur la convergence de la suite y_ε dans ce cadre là. Il faut se reporter au chapitre concernant le programme convexe pour comprendre et obtenir des résultats convaincants; cela s'appellera la pénalisation des contraintes.

3.3 Les équations d'Euler pour les problèmes de la mécanique

On considère un problème de la mécanique du point, ainsi on introduit les coordonnées (x,y,z) et on veut retrouver $m\ddot{X} = \vec{f}$ lorsque m est la masse de la particule, $X = (x,y,z)$ et $\vec{f} = -\nabla U$ est la force dérivant d'un potentiel. Analysons d'abord le phénomène. Il est classique de reconnaitre, en multipliant les équations par $\dot{X}$ et en inégrant sur $0,T$, que

$$\frac{1}{2}m(\dot{X}(T))^2 + U(X(T)) = \frac{1}{2}m(\dot{X}(0))^2 + U(X(0)).$$

Cette égalité s'écrit comme la conservation de l'énergie. Ce n'est pas celle ci que l'on souhaite obtenir, mais on cherche à interpréter le problème comme la solution d'une

équation d'Euler. Il faut donc que $m\ddot{X} = \vec{f}$ s'écrive $\frac{d}{dt}(\frac{\partial f}{\partial \dot{X}}) = \frac{\partial f}{\partial X}$. Pour cela, il serait simple d'avoir $\frac{\partial f}{\partial \dot{X}} = m\dot{X}$ et $\frac{\partial f}{\partial X} = -\nabla U$. Une solution à variables séparées est alors

$$f(X, \dot{X}) = \frac{1}{2}m(\dot{X})^2 - U(X).$$

On vérifie que l'équation d'Euler dans ce cas est bien l'équation dite loi de Newton. On en déduit que

La solution des équations du mouvement d'une particule dans un champ de forces conservatif, c'est-à-dire dérivant d'un potentiel, est la fonction qui minimise l'intégrale d'action

$$\mathcal{A}(X) = \int_{t_0}^{t_1} [\frac{1}{2}m(\dot{X}(t))^2 - U(X(t))]dt = \int_{t_0}^{t_1} (T - U)dt.$$

On a noté ici l'énergie cinétique $T = \frac{1}{2}m(\dot{X}(t))^2$.

Soit $L(q, \dot{q}) = T(\dot{q}) - U(q)$. Le changement de notation ici illustre la façon dont les mécaniciens notent ce problème. Si ξ est un élément de l'espace $H^1(]t_0, t_1[)$, le calcul de $\frac{1}{\varepsilon}[L(q_0 + \varepsilon\xi, \dot{q}_0 + \varepsilon\dot{\xi}) - L(q_0, \dot{q}_0)]$ conduit à l'expression

$$L'(q_0, \dot{q}_0) = \partial_q L(q_0, \dot{q}_0) - \frac{d}{dt}[\partial_{\dot{q}} L(q_0, \dot{q}_0)] + \partial_{\dot{q}} L(q_0, \dot{q}_0)(t_1)\delta_{t_1} - \partial_{\dot{q}} L(q_0, \dot{q}_0)(t_0)\delta_{t_0}.$$

La théorie des multiplicateurs de Lagrange avec $q(t_0) = q_0$, $q(t_1) = q_1$ donne alors immédiatement le système

$$\begin{cases} \partial_q L(q_0, \dot{q}_0) - \frac{d}{dt}[\partial_{\dot{q}} L(q_0, \dot{q}_0)] = 0(\text{ équation d'Euler}) \\ q_0(t_0) = q_0, q_0(t_1) = q_1(\text{ contraintes actives}) \\ \lambda_1 = -\partial_{\dot{q}} L(q_0, \dot{q}_0)(t_1) \\ \lambda_0 = \partial_{\dot{q}} L(q_0, \dot{q}_0)(t_0) \end{cases}$$

L'écriture des deux premières égalités permet d'avoir les conditions d'extrémité et l'équation de Newton. Les deux dernières donnent les multiplicateurs de Lagrange. On obtient

$$\lambda_1 = -m\dot{q}_0(t_1), \lambda_0 = m\dot{q}_0(t_0).$$

On interprète alors les multiplicateurs de Lagrange comme les quantités de mouvement aux extrémités de la courbe. On verra que la quantité de mouvement (ou l'impulsion) joue un rôle particulier ci-dessous.

3.4 Formulation hamiltonienne

On écrit dans ce cas l'action $L(q, \dot{q})$. On sait que la quantité $\dot{q}_0\partial_{\dot{q}} L(q_0, \dot{q}_0) - L(q_0, \dot{q}_0)$ se conserve. Généralisons en étudiant la quantité $\dot{q}(t)p(t) - L(q(t), \dot{q}(t))$. Cette quantité a pour dérivée

$$\ddot{q}(p - \partial_{\dot{q}} L) + \dot{q}(\dot{p} - \partial_q L).$$

On voit que cette quantité est nulle lorsque $p = \partial_{\dot{q}} L$ et $\dot{p} = \partial_q L$, ce qui implique que q est solution de l'équation d'Euler. D'autre part, la maximisation de $\tilde{q}p - L(q, \tilde{q})$

dans le cas L convexe en $\tilde{q}$ conduit à la première égalité $p = \partial_{\dot{q}} L(q, \tilde{q})$, ce qui porte un nom: transformation de Legendre. Revenant au cas où L dépend de t (car ceci n'est pas essentiel pour cette partie de l'analyse), soit

$$H(t, q, p) = \max_{\tilde{q}} (\tilde{q}p - L(t, q, \tilde{q})).$$

Par définition, H est la transformée de Legendre de L lorsqu'elle existe, et on a le résultat suivant:

"La transformée de Legendre de H est L."

Dans le cas de la mécanique du point $L(t, q, \tilde{q}) = \frac{1}{2} m(\tilde{q})^2 - U(q)$ ce qui donne $p = m\tilde{q}$ et ainsi $H(t, q, p) = \frac{1}{2}\frac{p^2}{m} + U(q)$. Apparait dans cette égalité l'énergie qui est l'hamiltonien, et la quantité de mouvement p qui est égale à $m\tilde{q}$.

On vérifie que si la matrice hessienne de L en $\tilde{q}$ au point $(q, \tilde{q})$ est définie positive (au voisinage de $(q_0, \dot{q}_0)$), l'équation $p = \partial_{\tilde{q}} L(t, q, \tilde{q})$ admet une solution unique par le théorème des fonctions implicites, que l'on note $Q(t, q, p)$. On vérifie alors

$$H(t, q, p) = pQ(q, p) - L(t, q, Q(q, p)).$$

On trouve alors les relations

$$\partial_q H(t, q, p) = (p - \partial_{\tilde{q}} L(t, q, Q(q, p))).\partial_q Q(t, q, p) - \partial_q L(t, q, Q(q, p)) = -\partial_q L(t, q, Q(q, p))$$
$$\partial_p H(t, q, p) = Q(t, q, p) + (p - \partial_{\tilde{q}} L(t, q, Q(q, p))).\partial_p Q(t, q, p) = Q(t, q, p).$$

On remarque alors, par unicité de la solution de l'équation $p = \partial_{\tilde{q}} L$, que pour $p(t) = \frac{\partial L}{\partial \tilde{q}}(t, q_0(t), \dot{q}_0(t))$, alors $Q(t, q_0(t), p(t)) = \dot{q}_0(t)$, soit

$$Q(t, q_0(t), \frac{\partial L}{\partial \tilde{q}}(t, q_0(t), \dot{q}_0(t))) = \dot{q}_0(t).$$

On en tire que, pour toute fonction $q_0(t)$, on a l'identité

$$\partial_p H(t, q_0(t), \frac{\partial L}{\partial \tilde{q}}(t, q_0(t), \dot{q}_0(t))) = \dot{q}_0(t).$$

Maintenant, si q_0 est solution de l'équation d'Euler, on trouve

$$\frac{d}{dt}(\frac{\partial L}{\partial \tilde{q}}(t, q_0(t), \dot{q}_0(t))) = \frac{\partial L}{\partial q}(t, q_0(t), \dot{q}_0(t)),$$

soit

$$\frac{d}{dt}(\frac{\partial L}{\partial \tilde{q}}(t, q_0(t), \dot{q}_0(t))) = -\partial_q H(t, q_0(t), \frac{\partial L}{\partial \tilde{q}}(t, q_0(t), \dot{q}_0(t))).$$

On en déduit le système, appelé système hamiltonien:

$$\begin{cases} \frac{dp}{dt} = -\frac{\partial H}{\partial q}(t, q_0(t), p(t)) \\ \frac{dq_0}{dt} = \frac{\partial H}{\partial p}(t, q_0(t), p(t)) \end{cases}$$

On a ainsi transformé l'équation d'Euler, du second ordre, en un système d'équation du premier ordre, appelé système hamiltonien.

Lorsque, de plus, L ne dépend pas de t, alors H ne dépend pas de t et on sait que $H(q_0(t), p(t)) = H(q_0(t_0), p(t_0))$. L'hamiltonien est une intégrale première du système hamiltonien.

Réciproquement, soit $H(t, q, p)$ l'hamiltonien associé à $L(t, q, p)$ lorsque $\partial^2_{\tilde{q}^2} L > 0$. La solution du système hamiltonien $(q(t), p(t))$ permet de construire $\dot{q}(t)$ par la première équation du système hamiltonien, qui est $\dot{q}(t) = \tilde{q}(t)$, où $\tilde{q}(t)$ est la solution de $p(t) = \partial_{\tilde{q}} L(t, q(t), \tilde{q}(t))$ et la dexième équation permet de vérifier que

$$\frac{d}{dt}(\partial_{\tilde{q}} L(t, q(t), \dot{q}(t)) = \partial_p L(t, q(t), \dot{q}(t)).$$

Soit L une action (un lagrangien) de la forme $L(t, q(t), \dot{q}(t))$. Lorsque $q(t)$ est une fonction donnée, L est une fonction de t uniquement. Lorsque on veut considérer les problèmes d'intégrale d'action, on se ramène à la fonctionnelle de $\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d$ dans $\mathbb{R}$ qui à $(t, q, \tilde{q})$ fait correspondre $L(t, q, \tilde{q})$.

On a démontré la proposition suivante, dans le cas où L est une fonction strictement convexe dans les variables $(q, \tilde{q})$:

Proposition 3.1 *On introduit le* **hamiltonien**, *fonctionnelle sur $\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d$, par*

$$H(t, q, p) = \max_{\tilde{q}} (p\tilde{q} - L(t, q, \tilde{q})).$$

Dire que le couple de fonctions de $\mathbb{R}$ dans $\mathbb{R}^d$ $(q_0(t), p_0(t))$ est solution du système hamiltonien

$$\begin{cases} \dot{q}_0(t) = \frac{\partial H}{\partial p}(t, q_0(t), p_0(t)) \\ \dot{p}_0(t) = -\frac{\partial H}{\partial q}(t, q_0(t), p_0(t)) \\ p_0(0) = p_0, q_0(0) = q_0 \end{cases}$$

équivaut à dire que
la fonction $q_0(t)$ est solution de l'équation d'Euler

$$\frac{d}{dt}(\frac{\partial L}{\partial \tilde{q}}(t, q_0(t), \dot{q}_0(t))) = \frac{\partial L}{\partial q}(t, q_0(t), \dot{q}_0(t))$$

avec les conditions initiales $q_0(0) = q_0$, $\dot{q}_0(0) = \tilde{q}_0$, où $\tilde{q}_0$ est la solution de $p_0 = \frac{\partial L}{\partial \tilde{q}}(t, q_0, \tilde{q}_0)$.

Ce système hamiltonien est très couramment utilisé en optique, mais il faut modifier pour cela la formulation de l'exemple 12 de l'introduction. En effet, l'équation d'Euler devient alors

$$\frac{d}{dx}(\frac{y'(x)}{c(x, y(x))(1 + (y'(x))^2)^{\frac{1}{2}}}) = -(1 + (y'(x))^2)^{\frac{1}{2}} \frac{\partial_y c}{c^2} \tag{3.4.4}$$

d'où on déduit

$$\frac{y"(x)}{c(x, y(x))(1 + (y'(x))^2)^{\frac{3}{2}}} + \frac{1}{c^2(1 + (y'(x))^2)^{\frac{1}{2}}} \partial_x c = \frac{y'(x)}{c^2(1 + (y'(x))^2)^{\frac{1}{2}}}.$$

On en déduit donc

$$\frac{d}{dx}(\frac{1}{c(x, y(x))(1 + (y'(x))^2)^{\frac{1}{2}}} = -(1 + (y'(x))^2)^{\frac{1}{2}} \frac{\partial_x c}{c^2}. \tag{3.4.5}$$

Les deux relations (3.4.5) et (3.4.4) expriment que $\frac{\vec{t}}{c}$ a sa dérivée qui suit le gradient de $\frac{1}{c}$, les rayons suivent le gradient de l'indice.

D'autre part, le hamiltonien équivalent au lagrangien $\frac{(1+(y')^2)^{\frac{1}{2}}}{c(x,y(x))}$ ne peut pas être calculé, car le lagrangien n'est pas strictement convexe.

Pour se ramener à un lagrangien strictement convexe, on considère que le terme $\frac{(1+(y')^2)^{\frac{1}{2}}}{c(x,y(x))}$ est un double produit, donc on a

$$\frac{(1+(y')^2)^{\frac{1}{2}}}{c(x,y(x))} = \frac{1}{2}[-(\frac{w}{c(x,y)} - \frac{(1+(y')^2)^{\frac{1}{2}}}{w})^2 + \frac{w^2}{c^2} + \frac{1+(y')^2}{w^2}].$$

Nous allons faire le raisonnement sur $L_w(q_1,q_2,\dot{q}_1,\dot{q}_2) = \frac{\dot{q}_1^2+\dot{q}_2^2}{w^2} + \frac{w^2}{c^2(q_1,q_2)}$. En effet, $L_w(q_1,q_2,\dot{q}_1,\dot{q}_2) \geq L_{w_0}(q_1,q_2,\dot{q}_1,\dot{q}_2)$ pour w_0 qui réalise le minimum en w, c'est à dire $w_0^2 = c(\dot{q}_1^2+\dot{q}_2^2)^{\frac{1}{2}}$. Dans ce cas on sait que d'une part

$$\inf \int_{t_1}^{t_2} L_w(q_1,q_2,\dot{q}_1,\dot{q}_2)dt = \inf \int_{t_1}^{t_2} \frac{(\dot{q}_1^2+\dot{q}_2^2)^{\frac{1}{2}}}{c(q_1,q_2)}dt$$

et d'autre part

$$\inf \int_{t_1}^{t_2} L_w(q_1,q_2,\dot{q}_1,\dot{q}_2)dt = \inf \int_{t_1}^{t_2} L_{w_0}(q_1,q_2,\dot{q}_1,\dot{q}_2)dt$$

Ceci est une forme abstraite pour dire, dans le cas qui nous intéresse que

$$\inf \int_{t_1}^{t_2} \frac{(\dot{q}_1^2+\dot{q}_2^2)^{\frac{1}{2}}}{c(q_1,q_2)}dt = \inf \frac{1}{2}\int_{t_1}^{t_2} (\frac{\dot{q}_1^2+\dot{q}_2^2}{c^2(q_1,q_2)} + 1)dt$$

Pour ce nouveau lagrangien

$$\mathcal{L}(x,y,\dot{x},\dot{y}) = \frac{1}{2}(\frac{\dot{x}^2+\dot{y}^2}{c^2} + 1)$$

le hamiltonien est $\mathcal{H}(x,y,p,q) = \frac{1}{2}((p^2+q^2)c^2 - 1)$. Ses courbes intégrales sont

$$\begin{cases} \frac{dx}{ds} = pc^2 \\ \frac{dy}{ds} = qc^2 \\ \frac{dp}{ds} = -c\partial_x c(p^2+q^2) \\ \frac{dq}{ds} = -c\partial_y c(p^2+q^2) \end{cases}$$

Il est constant sur les courbes bicaractéristiques. Si les données initiales sont telles que le hamiltonien soit nul, on trouve que $p^2+q^2 = \frac{1}{c^2}$. On choisit le changement d'abscisse curviligne donné par $du = c(x(s),y(s))ds$, alors

$$\begin{cases} \frac{dx}{du} = \frac{p}{(p^2+q^2)^{\frac{1}{2}}} \\ \frac{dy}{du} = \frac{q}{(p^2+q^2)^{\frac{1}{2}}} \\ \frac{dp}{du} = \partial_x \frac{1}{c} \\ \frac{dq}{du} = \partial_y \frac{1}{c}. \end{cases}$$

Le vecteur d'onde suit les courbes intégrales du gradient d'indice. Ceci correspond à une théorie d'optique géométrique, comme cela avait été vu ci-dessus .

Chapter 4

Programme convexe

4.1 Fonctions convexes

Nous voyons dans ce chapitre une application très importante des calculs précédents, dans la droite ligne des exemples 1, 2, 5, 13, 14, 15. Il s'agit du cas où J est convexe et où les contraintes sont convexes. Cette partie de l'analyse fonctionnelle est importante, car dans ce cas les conditions nécessaires et les conditions suffisantes d'optimalité deviennent des caractérisations des points d'extremum.

Nous avons déjà vu dans l'exemple que l'ensemble des points de minimum global d'une fonctionnelle convexe forment un ensemble convexe. Nous allons préciser les choses ici, par des définitions et par un résultat

Définition 4.1 *Soit K un ensemble convexe non vide (c'est-à-dire vérifiant, pour tout u, v dans K et tout réel β de $[0,1]$, $\beta u + (1-\beta)v \in K$.) On dit que la fonction J définie sur K est une fonction convexe si et seulement si on a*

$$\forall \beta \in [0,1], \forall (u,v) \in K^2, J(\beta u + (1-\beta)v) \leq \beta J(u) + (1-\beta)J(v).$$

La fonctionnelle J est strictement convexe si l'inégalité précédente est stricte pour $\beta \in]0,1[$ et $u \neq v$.

La fonctionnelle J est dite $\alpha-$convexe lorsque

$$J(\frac{u+v}{2}) \leq \frac{J(u)+J(v)}{2} - \frac{\alpha}{8}||u-v||^2$$

On peut définir un espace convexe simple à partir de J fonctionnelle convexe: il s'appelle l'épigraphe.

Définition 4.2 *On appelle épigraphe de J fonctionnelle convexe sur un convexe K l'espace $Epi(J)$ des $\{(\lambda, v), v \in K, \lambda \geq J(v)\}$. C'est un convexe.*

On vérifie que si (λ, v) et (μ, w) sont dans $Epi(J)$, alors pour $0 \leq \theta \leq 1$ on a $J(\theta v + (1-\theta)w) \leq \theta J(v) + (1-\theta)J(w) \leq \theta\lambda + (1-\theta)\mu$ donc $\theta(\lambda, v) + (1-\theta)(\mu, w)$ est dans $Epi(J)$.

Lemme 4.1 *Si J est $\alpha-$convexe et continue, elle est strictement convexe. De plus,*

$$J(\theta u + (1-\theta)v) \leq \theta J(u) + (1-\theta)J(v) - \frac{\alpha\theta(1-\theta)}{2}||u-v||^2.$$

Preuve On effectue d'abord un raisonnement par récurrence pour démontrer, pour tout $n \geq 1$, pour tout $p \leq 2^n$, l'inégalité pour $\theta = \frac{p}{2^n}$. Pour cela, on écrit, pour $p \geq 2^{n-1}$

$$\frac{pu + (2^n - p)v}{2^n} = \frac{u}{2} + \frac{\frac{p-2^{n-1}}{2^{n-1}}u + \frac{2^n-p}{2^{n-1}}v}{2}$$

et on fait l'hypothèse de récurrence sur l'indice $n-1$, pour tout p. Ainsi on a

$$J(\frac{pu + (2^n - p)v}{2^n}) \leq \frac{1}{2}(J(u)+J(\frac{p-2^{n-1}}{2^{n-1}}u+\frac{2^n-p}{2^{n-1}}v))-\frac{\alpha}{2}||\frac{p-2^{n-1}}{2^{n-1}}u+\frac{2^n-p}{2^{n-1}}v-u||^2.$$

Appliquant l'hypothèse de récurrence, il vient

$$\begin{array}{ll} J(\frac{pu+(2^n-p)v}{2^n}) & \leq \frac{1}{2}(J(u) + \frac{p-2^{n-1}}{2^{n-1}}J(u) + \frac{2^n-p}{2^{n-1}}J(v)) - \frac{1}{4}\alpha\frac{p-2^{n-1}}{2^{n-1}}\frac{2^n-p}{2^{n-1}}||v-u||^2 \\ & -\frac{\alpha}{8}||\frac{p-2^{n-1}}{2^{n-1}}u + \frac{2^n-p}{2^{n-1}}v - u||^2. \end{array}$$

Le premier terme est alors égal à $\frac{p}{2^n}J(u) + \frac{2^n-p}{2^n}J(v)$. Le second terme est ainsi $\frac{\alpha}{8}\frac{2^n-p}{2^{n-1}}\frac{p}{2^{n-1}}||u-v||^2$, et est donc égal à $\frac{\alpha}{2}\frac{p}{2^n}\frac{2^n-p}{2^n}||u-v||^2$. Le cas $p < 2^{n-1}$ se traite en échangeant les rôles de u et de v. L'inégalité est démontrée pour θ de la forme $\frac{p}{2^n}$, puisque pour $n-1$, on a $p = 0$ ou $p = 1$.

Pour la démontrer pour θ quelconque, on utilise le fait que, pour tout n, il existe θ_n égal à $\sum_{i=1}^{i=n}\frac{\alpha_i}{2^i}$ tel que $\alpha_i(\theta) \in \{0,1\}$ et tel que $|\theta - \theta_n| \leq \frac{1}{2^n}$ (développement binaire).

On a, pour tout n

$$J(\theta_n u + (1-\theta_n)v) \leq \theta_n J(u) + (1-\theta_n)J(v) - \frac{\alpha\theta_n(1-\theta_n)}{2}||u-v||^2.$$

La limite des deux membres existe, car J est continue, ainsi on a

$$J(\theta u + (1-\theta)v) \leq \theta J(u) + (1-\theta)J(v) - \frac{\alpha\theta(1-\theta)}{2}||v-u||^2.$$

Le lemme est démontré, et on vérifie la stricte convexité sans souci.

On a les résultats suivants:

Proposition 4.1 *Si J est convexe continue sur K convexe fermé non vide, il existe une forme linéaire continue L et une constante δ telles que $J(v) \geq L(v) + \delta$. Si J est $\alpha-$convexe, on a $J(v) \geq \frac{\alpha}{8}||v||^2 - C$*

Preuve Si J est convexe continu, son **épigraphe** est convexe fermé non vide. Démontrons qu'il est fermé. Soit (λ_n, v_n) une suite de points de l'épigraphe qui converge vers (λ, v) dans l'espace de Hilbert $\mathbb{R} \times V$ muni de la norme $(\lambda^2 + ||v||^2)^{\frac{1}{2}}$. On vérifie que

$$\lambda_n \geq J(v_n). \tag{4.1.1}$$

Soit, si $J(v_{\phi(n)})$ tend vers a, on en déduit que $\lambda \geq a$. Bien sûr, comme J est continue, $a = J(v)$.

On remarque aussi que si $J(v) \leq a$ pour tout a valeur d'adhérence de la suite $J(v_n)$, alors on a (λ, v) qui est dans l'épigraphe, et l'épigraphe est fermé.

On remarque alors que le Lemme suivant est vrai

Lemme 4.2 *Si, pour tout v, on a*

$$J(v) \leq inf\{a, a \text{ valeur d'adhérence de toute suite } J(v_n), v_n \to v\},$$

alors l'épigraphe de J est fermé.

La notion de continuité plus faible évoquée dans ce lemme porte le nom de semi continuité inférieure (et on note parfois J s.c.i.).

Reprenons la démonstration de la proposition.

Soit $v_0 \in K$ et $\lambda_0 < J(v_0)$.

On note ce point p_0, qui est à l'extérieur de l'épigraphe et on désigne sa projection sur l'épigraphe $Epi(J)$ par $p_* = (\lambda_*, w_0)$. On montre d'abord $\lambda_* = J(w_0)$.

Comme la projection réalise le minimum de la distance, on a $\forall(\lambda, v)$, $\lambda \geq J(v)$, l'inégalité $(\lambda - \lambda_0)^2 + (v - v_0)^2 \geq (\lambda_* - \lambda_0)^2 + (w_0 - v_0)^2$.

On suppose $v = w_0$, auquel cas pour $\lambda \geq J(w_0)$ on a $(\lambda - \lambda_0)^2 \geq (\lambda_* - \lambda_0)^2$. On sait que $\lambda_* \geq J(w_0)$. Si $J(w_0) \geq \lambda_0$, on trouve $\lambda \geq J(w_0) \Rightarrow \lambda \geq \lambda_0$, donc $\lambda \geq \lambda_*$ pour $\lambda \geq J(w_0)$ et on en déduit $J(w_0) \geq \lambda_*$ et comme (λ_*, w_0) est dans l'épigraphe, $\lambda_* = J(w_0)$.

Si $J(w_0) < \lambda_0$, le point (λ_0, w_0) est dans l'épigraphe, donc on trouve $(\lambda_* - \lambda_0)^2 \leq 0$, donc $\lambda_* = \lambda_0$.

Dans le cas où J est continue, il existe θ tel que $J(\theta v_0 + (1-\theta)w_0) = \lambda_0$, puisque $J(v_0) < \lambda_0 < J(w_0)$. Alors, pour ce θ, on trouve

$$(1-\theta)^2(v_0 - w_0)^2 \geq (v_0 - w_0)^2$$

ce qui est impossible puisque pour $\theta = 1$, la valeur est distincte de λ_0.

Dans le cas général, soit θ_0 tel que $\theta_0 J(v_0) + (1-\theta_0)J(w_0) = \lambda_0$. Alors $J(\theta_0 v_0 + (1-\theta_0)w_0) \leq \lambda_0$, et le point $(\lambda_0, \theta_0 v_0 + (1-\theta_0)w_0)$ est dans l'épigraphe. On en déduit

$$(1-\theta_0)^2(v_0 - w_0)^2 \geq (v_0 - w_0)^2$$

ce qui entraine $v_0 = w_0$, impossible car $J(v_0) < \lambda_0 < J(w_0)$.

On a donc montré que $\lambda_* = J(w_0)$.

On a alors l'inégalité fondamentale de la projection:

$$(p_0 - p_*, p_0 - p) \geq 0 \forall p \in Epi(J).$$

Cette inégalité s'écrit, pour $p = (J(v), v)$

$$(\lambda_0 - J(w_0))(\lambda_0 - J(v)) + (v_0 - w_0, v_0 - v) \geq 0$$

soit

$$(J(w_0) - \lambda_0)J(v) \geq (v_0 - w_0, v - v_0) + (J(w_0) - \lambda_0)\lambda_0. \tag{4.1.2}$$

La démonstration du premier alinéa est alors la conséquence de $J(w_0) - \lambda_0 > 0$, ce que nous allons démontrer.

Si on avait $J(w_0) - \lambda_0 \leq 0$, alors le point (λ_0, w_0) serait dans $Epi(J)$ donc on aurait

$$||(J(w_0), w_0) - (\lambda_0, v_0)|| \leq ||(\lambda_0, v_0) - (\lambda_0, w_0)||$$

soit $(J(w_0) - \lambda_0)^2 + ||w_0 - v_0||^2 \leq ||v_0 - w_0||^2$, ce qui donne $\lambda_0 = J(w_0)$.

Il faut alors éliminer l'égalité $\lambda_0 = J(w_0)$. Pour cela, introduisons $0 \leq \theta \leq 1$ et raisonnons par l'absurde, soit $J(w_0) = \lambda_0 < J(v_0)$. Le point $\theta v_0 + (1-\theta)w_0$ est dans le convexe K , donc $(\theta v_0 + (1-\theta)w_0, J(\theta v_0 + (1-\theta)w_0))$ est dans $Epi(J)$. On a donc, pour $\lambda \geq J(\theta v_0 + (1-\theta)w_0)$

$$(\lambda - J(w_0))^2 + (1-\theta)^2||v_0 - w_0||^2 \geq ||v_0 - w_0||^2.$$

Deux cas: ou il existe une suite θ_n tendant vers 0 telle que $J(\theta_n v_0 + (1-\theta_n)w_0) < J(w_0)$, et dans ce cas je prends $\lambda = \lambda_0 = J(w_0)$ ce qui donne $v_0 = w_0$ impossible, ou alors il existe θ_0 tel que pour $0 < \theta < \theta_0$ on ait $J(\theta v_0 + (1-\theta)w_0) \geq J(w_0)$. Dans ce cas, pour $0 < \theta < \theta_0$ on trouve, replaçant λ par $J(\theta v_0 + (1-\theta)w_0)$ et utilisant l'inégalité $J(\theta v_0 + (1-\theta)w_0) - J(w_0) \leq \theta(J(v_0) - J(w_0))$,on en déduit

$$\theta(J(v_0) - J(w_0))^2 \geq (2-\theta)||v_0 - w_0||^2.$$

La limite $\theta \to 0$ conduit à $v_0 = w_0$, impossible.

On a donc éliminé $J(w_0) = \lambda_0$ donc, par les deux raisonnements, $J(w_0) - \lambda_0 > 0$.

On divise par cette quantité l'inégalité (4.1.2). On trouve

$$J(v) \geq (\tfrac{v_0 - w_0}{(J(w_0) - \lambda_0)}, v - v_0) + (J(w_0) - \lambda_0)\lambda_0.$$

La première inégalité de la proposition est démontrée.

D'autre part, on trouve, pour v_0 fixé

$$\frac{J(v) + J(v_0)}{2} \geq J(\frac{v + v_0}{2}) + \frac{\alpha}{8}||v - v_0||^2 \geq L(\frac{v + v_0}{2}) + \delta + \frac{\alpha}{8}||v - v_0||^2$$

On utilise alors le fait que $\frac{\alpha}{8}||v - v_0||^2 + \frac{L(v)}{2}$ est quadratique en $+\infty$ pour voir que cette fonction est minorée par

$$\frac{\alpha}{8}||v||^2 - [||L|| + \frac{\alpha}{4}||v_0||]||v||$$

qui peut être minoré par $\frac{\alpha}{4}||v||^2 - C_1$, d'où le résultat.

La relation entre les fonctionnelles convexes et les problèmes de minimisation est la suivante:

Proposition 4.2 *Soit J une fonctionnelle convexe sur un ensemble convexe K. Tout point de minimum local est un point de minimum global, et les points de minimum forment un ensemble convexe. Cet ensemble convexe est réduit à un point lorsque J est strictement convexe*

Soit u un point de minimum local. Pour $v \in K$, et pour θ petit, $u + \theta(v - u)$ est dans un voisinage de u, et donc, pour $0 < \theta < \theta_0$, $J(u+\theta(v-u)) \geq J(u)$. De l'inégalité $J(u+\theta(v-u)) \leq (1-\theta)J(u) + \theta J(v)$, on déduit que $J(v) - J(u) \geq 0$, et donc u est un minimum global. On a déjà montré que si deux points étaient minimum global, alors tout le segment l'était, grâce à $J(u) \leq J(\theta u + (1-\theta)v) \leq \theta J(u) + (1-\theta)J(v) = J(u)$. Enfin, si u et v sont deux minima globaux distincts et si J est strictement convexe,

$$J(\frac{u + v}{2}) < \frac{1}{2}(J(u) + J(v)) = J(u)$$

ce qui est impossible.

On écrit ensuite des propriétés des fonctions convexes dérivables. On a la

Proposition 4.3 *Soit J une application différentiable. Il est équivalent de dire*
(i) la fonctionnelle J est convexe
(ii) Pour tous (u,v) dans V, $J(v) \geq J(u) + (J'(u), v-u)$
(iii) Pour tous (u,v) $(J'(u) - J'(v), u-v) \geq 0$.
De même on caractérise l'$\alpha-$convexité par

$$J(v) \geq J(u) + (J'(u), v-u) + \frac{\alpha}{2}||v-u||^2$$

ou par

$$(J'(u) - J'(v), u-v) \geq \alpha||u-v||^2.$$

Lorsque J est $\alpha-$convexe, on a

$$J(u+\theta(v-u)) \leq J(u) + \theta(J(v) - J(u)) - \frac{\alpha}{2}\theta(1-\theta)||u-v||^2.$$

Ainsi

$$\frac{J(u+\theta h) - J(u)}{\theta} \leq J(u+h) - J(u) - \frac{\alpha}{2}(1-\theta)||h||^2.$$

Passant à la limite en $\theta \to 0$, on trouve la première inégalité.

Ensuite, lorsque la première inégalité est vérifiée, on l'écrit pour u et pour v:

$$J(v) \geq J(u) + (J'(u), v-u) + \frac{\alpha}{2}||v-u||^2$$

$$J(u) \geq J(v) + (J'(v), u-v) + \frac{\alpha}{2}||v-u||^2$$

et on les additionne pour trouver la deuxième inégalité.

Enfin, considèrant u vérifiant la deuxième inégalité, on veut étudier $\phi(t) = J(tu + (1-t)v)$.

On voit que $\phi'(t) = J'(tu + (1-t)v), u-v)$. On en déduit $\phi'(t) - \phi'(s) = J'(tu + (1-t)v), u-v) - J'(su + (1-s)v), u-v) = \frac{1}{t-s}[J'(tu + (1-t)v - J'(su + (1-s)v), tu + (1-t)v - su - (1-s)v)]$. Lorsque $t \geq s$, on trouve bien $\phi'(t) - \phi'(s) \geq \alpha||v-u||^2(t-s)$. Intégrant de $s = 0$ à $s = \frac{1}{2}$ et de $t = \frac{1}{2}$ à $t = 1$, on trouve

$$\frac{1}{2}[\phi(1) - 2\phi(\frac{1}{2}) + \phi(0)] \geq \alpha||u-v||^2 \int_{\frac{1}{2}}^{1} [\frac{1}{2}t - \frac{1}{8}]dt = \frac{\alpha}{8}||u-v||^2.$$

On a donc l'inégalité d'$\alpha-$convexité. Les caractérisations d'$\alpha-$convexité sont obtenues.

D'autre part, on note que dans le cas $\alpha = 0$ on a $\phi'(t) - \phi'(s) \geq 0$ si $t \geq s$. Ainsi on trouve $\int_\theta^1 dt \int_0^\theta ds(\phi'(t) - \phi'(s))ds = \theta\phi(1) + (1-\theta)\phi(0) - \phi(\theta)$ et c'est un réel positif. On a la convexité. Le raisonnement précédent est valable pour $(i) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (i)$. On note finalement que la convexité et l'$\alpha-$convexité sont aussi caractérisées, pour le cas simple de J deux fois différentiable, par $(J''(u)w, w) \geq 0$ et par $(J''(u)w, w) \geq \alpha(w, w)$.

4.2 Minimisation de fonctionnelles convexes

Le résultat agréable dans le programme convexe est que, contrairement au cas de l'exemple 16, la condition J infinie à l'infini suffit.

Théorème 4.1 *Soit K un convexe fermé non vide dans un Hilbert V et soit J une fonctionnelle convexe continue sur K.*

- *Si J est infinie à l'infini, alors J admet un minimum.*
- *Si J est $\alpha-$convexe continue, le minimum u est unique, et on a*

$$\forall v \in K, ||v-u||^2 \leq \frac{4}{\alpha}[J(v)-J(u)].$$

Le premier résultat se base sur la convergence faible d'une suite minimisante u_n. Nous l'admettons ici.

Le deuxième résultat provient de l'écriture, pour u_n suite minimisante, de la relation, notant l l'inf de J

$$l \leq J(\frac{u_n+u_m}{2}) \leq \frac{J(u_n)+J(u_m)}{2} - \frac{\alpha}{8}||u_n-u_m||^2$$

qui implique

$$||u_n-u_m||^2 \leq \frac{4}{\alpha}[(J(u_m)-l)+(J(u_n)-l)]$$

Nous sommes exactement dans le cas d'application du critère de Cauchy, ainsi la suite u_m est de Cauchy, donc possède une limite u. On passe à la limite en m dans l'inégalité ci-dessus, ce qui implique que

$$||u_n-u||^2 \leq \frac{4}{\alpha}[J(u_n)-l] = \frac{4}{\alpha}[J(u_n)-J(u)].$$

Le résultat est démontré.

Dans le cas convexe, on a une condition **nécessaire et suffisante** d'optimalité, obtenue à partir de la condition nécessaire provenant de l'équation d'Euler, que je rappelle ci-dessous

Proposition 4.4 *Soit K convexe. On suppose que J est différentiable en u. Si u est un point de minimum local de J sur K, alors*

$$\forall v \in K, (J'(u), v-u) \geq 0$$

Cette proposition est une conséquence du fait que, pour $u \in K$, toutes les directions admissibles sont $v-u$ pour $v \in K$, car $u+\theta(v-u)$ est dans K pour $0<\theta<1$.

On a

Théorème 4.2 *Si K est convexe et si J est une fonctionnelle convexe,*

$$u \text{ minimum de } J \Leftrightarrow \forall v \in K, (J'(u), v-u) \geq 0.$$

On sait que, si $\forall v \in K, (J'(u), v-u) \geq 0$, alors, de (ii) de la proposition 4.3 implique que

$$\forall v \in K, J(v) \geq J(u).$$

Ainsi u est un minimum global.[1]

On note que, lorsque le K est un cône convexe fermé (c'est-à-dire $\lambda v \in K$ pour $v \in K$ et $\lambda > 0$), on a

Proposition 4.5 *Le minimum de J est* **caractérisé** *par*

$$(J'(u), u) = 0 \text{ et } (J'(u), w) \geq 0 \forall w \in K$$

La démonstration de cette proposition suit les idées utilisées dans la résolution de l'exemple 15, où on a choisi $v = cu$. On prend ainsi l'inégalité

$$(J'(u), v-u) \geq 0 \forall v \in K$$

et on prend $v = \lambda u$. Les deux cas $\lambda > 1$ et $0 < \lambda < 1$ donnent $(J'(u), u) = 0$, et le remplacer dans l'inégalité donne le résultat de la proposition.

4.3 Fonctionnelles quadratiques

Le cas particulier de ces résultats le plus important correspond à la **minimisation de fonctionnelles quadratiques**, c'est-à-dire, dans l'exemple le plus classique, si $(,)$ désigne le produit scalaire sur V Hilbert

$$J(v) = \frac{1}{2}a(v,v) - (b,v)$$

où a est une forme bilinéaire continue sur V et b est un élément de V.

Définition 4.3 *On dit que la forme bilinéaire a continue sur V est coercive si et seulement si il existe $\nu > 0$ tel que*

$$\forall u \in V a(u,u) \geq \nu ||u||^2.$$

On a alors le

Lemme 4.3 *Si a est coercive, et qu'une de ses constantes de coercivité est ν, alors a est $\nu-$convexe.*

ce qui entraine

Théorème 4.3 *Le minimum de J sur K convexe est unique et noté u. C'est l'unique solution du problème*

$$u \in K et \forall v \in K, a(u, v-u) \geq (b, v-u).$$

[1] La redémonstration rapide de l'inéquation d'Euler provient de $\frac{1}{\theta}(J(u+\theta(v-u)) - J(u)) \geq 0$ lorsque u est le minimum.

Preuve du Lemme On vérifie ainsi que

$$(J'(u),w)=\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}[J(u+\varepsilon w)-J(u)]=\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}[\varepsilon a(u,w)+\frac{\varepsilon^2}{2}a(w,w)-\varepsilon(b,w)]=a(u,w)-(b,w).$$

Alors $(J'(u)-J'(v),u-v)=a(u,u-v)-(b,u-v)-a(v,u-v)+(b,u-v)=a(u-v,u-v)$, donc

$$(J'(u)-J'(v),u-v)\geq\nu(u-v,u-v).$$

D'après la proposition 4.3, on a le lemme. L'identification de la dérivée donne l'inégalité caractérisant le minimum (obtenue au théorème 4.2):

$$a(u,v-u)-(b,v-u)\geq 0\forall v\in K$$

ce qui est le résultat du théorème.

4.4 Notion de point selle, et théorème de Kuhn et Tucker

4.4.1 Introduction à la notion de Lagrangien

Nous nous reportons à l'exemple $\inf\frac{1}{2}(y_1^2+y_2^2)-b.y$ sous la contrainte $a.y=0$ ou sons la contrainte $a.y\leq 0$. Nous avons vu que cela pouvait être simple (et que c'était certainement naturel) de considérer la projection du minimum absolu b sur l'ensemble des contraintes. Nous avons vu que si b est dans l'ensemble des contraintes, sa projection est lui même, et en revanche si b n'y est pas, le point où la fonctionnelle atteint son minimum est bien le point b_0 de projection de b sur l'ensemble des contraintes. Nous avons écrit le point $b_0=b-\lambda a$, c'est à dire nous avons résolu $y-b+\lambda a=0$.

Montrons d'abord que tous les arguments précédents s'appliquent. On vérifie que

$$J(\frac{x_1+y_1}{2},\frac{x_2+y_2}{2})-\frac{1}{2}J(x_1,y_1)-\frac{1}{2}J(x_2,y_2)=-\frac{1}{8}(x_1-y_1)^2-\frac{1}{8}(x_2-y_2)^2$$

ce qui fait que J est $1-$convexe! D'autre part, une contrainte linéaire est convexe, on est donc dans le cas du programme convexe. D'autre part, on trouve $J'(y_1,y_2)=y-b$. La condition nécessaire d'optimalité est alors

$$(y^0-b,y-y^0)\geq 0,\forall y,a.y=0$$

• cas égalité:

Si y^0 est intérieur à $a.y=0$ (c'est-à-dire $a.y^0\neq 0$) alors $y^0=b$ et si b vérifie $a.b=0$ cela convient.

Si y^0 est au bord de $a.y=0$ (c'est-à-dire $a.y^0=0$) on a $a.(y-y^0)=0$ donc $y-y^0$ est proportionnel à a^T, ainsi $(y^0-b,\mu a^T)\geq 0$ pour tout μ, donc $(y^0-b).a^T=0$, soit $y^0-b=-\lambda a$, et on identifie λ grâce à $y^0.a=0$.

• cas inégalité:

si y^0 est intérieur à $a.y\leq 0$, alors $a.y^0<0$ et donc toutes les directions sont admissibles et donc $y^0=b$. Si on n'est pas dans le cas $b.a<0$, le point b n'est pas le minimum sur l'espace des contraintes car il n'est pas intérieur à l'espace des contraintes.

On suppose donc maintenant que $a.b \geq 0$. On sait donc que y^0 est sur le bord $a.y^0 = 0$. On voit alors que pour tout $y \in \{a.y \leq 0\}$, alors $a.(y-y^0) \leq 0$. Les directions possibles pour $y - y^0$ sont donc a^T et a, le coefficient devant a étant **négatif**. On écrit $y - y^0 = \mu a^T - \mu_1 a$, et on en déduit que

$$\forall \mu \in \mathbb{R}, \forall \mu_1 \in \mathbb{R}_+, (y^0 - b, \mu a^T - \mu_1 a) \geq 0$$

Ceci implique que $y^0 - b$ est orthogonal à a^T et que $(y^0 - b, a) \leq 0$. On en déduit $y^0 - b = -\lambda a$ avec $\lambda \geq 0$ et de plus, comme y^0 est sur le bord, $y^0.a = 0$ donc $(b - \lambda a).a = 0$ donc $\lambda = \frac{b.a}{a^2}$, qui est négatif ou nul grâce à l'hypothèse $a.b \geq 0$.

Nous avons ici reconstruit les multiplicateurs de Lagrange, de manière plus directe puisque avec une seule contrainte dans $\mathbb{R}^2$ on n'a pas besoin d'un résultat aussi général que le lemme de Farkas.

Remarque Utilisons la forme du minimum obtenu pour écrire $y = b - \lambda a + z$. On trouve

$$J(y) = \frac{1}{2}z^2 - \frac{1}{2}b^2 + \frac{1}{2}\lambda^2 a^2 - \lambda a.b.$$

La contrainte s'écrit $a.b - \lambda a^2 + a.z \leq 0$.

Le minimum de la fonctionnelle en λ est donc obtenu pour $\lambda_0 = \frac{a.b}{a^2}$, la contrainte restante dans ce cas est alors $a.z \leq 0$ et il reste la minimisation de $\frac{1}{2}z^2$, minimum atteint pour $z = 0$.

Remarque Soit w une direction admissible pour la contrainte inégalité $F(y) \leq 0$ (ici c'est $a.y \leq 0$ et donc on a $(F'(y), w) \leq 0$ soit encore $a.w \leq 0$). On suppose qu'il existe un couple (y_0, λ_0) dans $\{F \leq 0\} \times \mathbb{R}_+$, tel que $J'(y_0) + \lambda_0 F'(y_0) = 0$ et $F(y_0) = 0$. Alors on introduit

$$\phi(t) = J(y_0 + tw)$$

On a $\phi'(t) = (J'(y_0 + tw), w)$ et $\phi'(0) = -\lambda_0(F'(y_0), w) \geq 0$. Comme w est une direction admissible, $y_0 + tw$ est dans l'espace des contraintes, donc on doit retrouver que $\phi'(t) \geq 0$. On a bien sûr $\phi'(0) \geq 0$ donc $\phi(t) \geq \phi(0)$ ce qu'il faut vérifier pour que y_0 soit un minimum.

D'autre part, on vérifie que $\frac{d}{dt}(F(y_0 + tw)) = (F'(y_0 + tw), w)$ donc il est trivial que

$$\frac{d}{dt}(\phi(t) + \lambda_0 F(y_0 + tw))|_{t=0} = 0.$$

On vérifie ainsi très directement que y_0 n'est pas seulement le minimum de J mais aussi le minimum de $J + \lambda_0 F$.

Ceci nous amène à introduire dans l'exemple canonique en dimension 2 cette nouvelle fonctionnelle. On pose

$$\mathcal{L}(y, \lambda) = J(y) + \lambda a.y$$

Le minimum sur $\mathbb{R}^2$ de cette fonctionnelle est obtenu en $y = b - \lambda a$, ce qui correspond à la remarque que nous avons déjà faite sur le fait que cette écriture est la bonne écriture pour trouver le minimum. Maintenant, lorsque y est dans l'intérieur de l'espace des contraintes $a.y < 0$ et que λ est assez petit, alors $y + \lambda a$ est aussi dans l'espace des contraintes, donc le minimum de $\mathcal{L}(y, \lambda)$ est atteint en un point y_λ de l'espace des contraintes,

et on vérifie que ce minimum vaut $-\frac{1}{2}(b-\lambda a)^2$. Cette fonction de λ admet un maximum en $\lambda = \frac{a.b}{a^2}$. et cette valeur du point où elle est maximum est celle cherchée pour obtenir le point critique de J sous les contraintes $a.y \leq 0$ lorsque b n'est pas dans l'espace des contraintes.

D'autre part, lorsque y n'est pas dans l'espace $F(y) = 0$, on voit que $\mathcal{L}(y, \lambda)$ n'a certainement pas d'extremum en λ (contrairement à ce que l'on a fait dans le paragraphe ci-dessus) et on a probablement identifié un problème équivalent.

4.4.2 Point selle, lagrangien, et minimisation de fonctionnelle convexe

On considère une fonctionnelle J à minimiser sur V, et on introduit, dans le cas de M contraintes inégalités ou de M contraintes égalités, une application de $V \times \mathbb{R}^M$ dans $\mathbb{R}$. Elle s'appellera Lagrangien, et on construit le Lagrangien associé à J et aux contraintes inégalités $F_j(v)$:

$$\mathcal{L}(v, q) = J(v) + \sum_j q_j F_j(v).$$

Dans le cas des contraintes inégalités, on désigne par $P = (\mathbb{R}_+)^M$, et dans le cas de contraintes égalités, on note $P = (\mathbb{R}^M)$. Soit $U \subset V$

Définition 4.4 *On dit que $(u, p) \in V \times P$ est un point selle de $\mathcal{L}$ sur $U \times P$ si on a les inégalités*

$$\forall q \in P, \mathcal{L}(u, q) \leq \mathcal{L}(u, p) \leq \mathcal{L}(v, p) \forall v \in U.$$

Notons que cette définition est la bonne définition pour les multiplicateurs de Lagrange, puisque les extrema sont caractérisés par la dérivée nulle.

On a

Proposition 4.6 *Si les fonctions J, $F_1, ...F_M$ sont continues sur V et si (u, p) est un point selle de $\mathcal{L}$ sur $U \times P$. Alors, K étant défini par les contraintes F_j (égalité si $P = \mathbb{R}^M$, inégalités si $P = (\mathbb{R}_+)^M$, et $K \subset U$, on a*

- *l'élément u est dans K*
- *c'est un minimum global de J sur K*
- *Dans le cas où K est inclus dans l'intérieur de U, et où les fonctionnelles sont dérivables, on a*

$$J'(u) + \sum_{j=1}^{M} p_j F_j'(u) = 0.$$

Preuve On suppose que (u, p) est un point selle. On se place tout d'abord dans le cas de contraintes d'égalité. Si on suppose que, pour tout q dans $\mathbb{R}^M$, alors $\mathcal{L}(q, u) \leq \mathcal{L}(p, u)$, comme $\mathcal{L}(q, u)$ est une fonction affine en q, cette inégalité ne peut être vérifiée que lorsque $F(u) = 0$. On a donc, écrivant la deuxième inégalité, $J(u) \leq J(v)$ pour tout $v \in U$, donc a fortiori pour tout $v \in K$, et donc u est un minimum global de J sur K.

On se place ensuite dans le cas de contraintes inégalités. Si on a, $\forall q \in (\mathbb{R}_+)^M$, l'inégalité, ceci veut dire que, en faisant tendre q vers $+\infty$ composante après composante, que $F(u) \leq 0$. On trouve alors $pF(u) \geq 0$ par l'inégalité, et comme $F_j(u) \leq 0$,

on trouve que $p_j F_j(u) = 0$ pour tout j. Ceci permet de conclure sur le fait que u est un minimum global de J car $pF(v) \leq 0$ ainsi $J(v) + pF(v) \leq J(v)$ et donc l'inégalité de droite de définition du point selle entraine $J(u) + 0 \leq J(v)$. Le point u est aussi minimum de la fonctionnelle $J(v) + pF(v)$, donc nécessairement la dérivée de cette fonctionnelle est nulle si K est intérieur à U.

Ce qui est extraordinaire est qu'il y a des conditions pour lesquelles cette proposition donne une condition **nécessaire et suffisante d'optimalité**

Théorème 4.4 *(Théorème de Kuhn et Tucker, 1951)*

On suppose que J, F sont convexes, continues, dérivables, et on suppose qu'il existe un élément de V tel que $\tilde{v}$ vérifie

"ou bien $F_i(\tilde{v}) < 0$, ou bien $F_i(\tilde{v}) = 0$ et F_i affine."

u est minimum global de J sur K si et seulement si il existe $p \in (\mathbb{R}_+)^M$ tel que (u, p) soit un point selle du Lagrangien $\mathcal{L}$ sur $V \times (\mathbb{R}_+)^M$.

Autrement dit, un minimum d'une fonctionnelle convexe avec contraintes est un minimum libre du Lagrangien lorsqu'on choisit les paramètres de Lagrange.

Preuve On considère un point de minimum global sur K. Soit $I(u)$ l'ensemble des indices où les contraintes sont actives, qui est, rappelons le, l'ensemble des indices tels que $F_i(u) = 0$. La convexité de F_i entraine que

$$F_i(\tilde{v}) - F_i(u) \geq (F_i'(u), \tilde{v} - u)$$

donc $(F_i'(u), \tilde{v} - u) < 0$ dans le cas où $F_i(\tilde{v}) < 0$ et

$(F_i'(u), \tilde{v} - u) = F_i(\tilde{v}) - F_i(u) = 0$ si F_i est affine et $F_i(\tilde{v}) = 0$.

On retrouve la notion de contraintes qualifiées au sens de la définition 2.6, le w_0 dans ce cas étant $\tilde{v} - u$. La condition nécessaire du théorème 2.6 donne donc l'égalité

$$\exists \lambda \in (\mathbb{R}_+)^M, J'(u) + \lambda F'(u) = 0.$$

Cette inégalité ne suffit pas pour montrer que le Lagrangien a un point selle. Pour cela, on considère l'ensemble $A \subset \mathbb{R}^{M+1}$ suivant

$$A = \{(\mu_0, \mu) \in \mathbb{R}^{M+1}, \exists v \in K, \mu_0 > J(v), \mu_j > F_j(v)\}.$$

A est un ouvert convexe, et si u est un minimum global pour la fonctionnelle sur l'espace des contraintes, alors $\forall v, F_j(v) \leq 0$ on a $J(v) \geq J(u)$.

Ceci veut dire que $(J(u), 0) \notin A$. La projection sur un convexe ouvert est aussi possible. Il existe donc $(p_0, p) \in \mathbb{R}^{M+1}$, $(p_0, p) \neq (0, 0)$ (ceci car on peut définir, si le point est dans l'adhérence du convexe ouvert, une direction normale au bord) tel que

$$p_0(\mu_0 - J(u)) + p\mu > 0 \forall (\mu_0, \mu) \in A.$$

En faisant tendre μ_0 et μ vers $+\infty$, on en déduit $p_0 \geq 0$, $p \geq 0$.

Le réel p_0 est non nul, car sinon en choisissant $(J(\tilde{v}) + 1, 0)$ qui est dans A pour les contraintes non affines (et on prend $\mu_j > 0$ tendant vers 0 pour les contraintes affines, et μ_j tendant vers $F_j(\tilde{v})$ pour les contraintes non affines) on trouverait $(p, F_j(\tilde{v})) \geq 0$ pour les contraintes non affines, et $p \geq 0$ contradictoire avec $F_j(\tilde{v}) < 0$. Ainsi $p_0 > 0$ donc on trouve

$$\forall(\mu_0,\mu)\in A, \mu_0 - J(u) + \frac{p}{p_0}\mu > 0$$

Comme $A = \cup_{v\in V}]J(v),+\infty[\times]F_j(v),+\infty[$, il vient

$$\forall v, J(v) - J(u) + \frac{p}{p_0}F(v) \geq 0.$$

Finalement, si $v = u$ on en déduit $\frac{p}{p_0}F(u) \geq 0$, donc comme $p_j \geq 0$ et $F_j(u) \leq 0$ on trouve $\frac{p}{p_0}F(u) = 0$ donc on trouve

$$\forall v \in V, J(v) + (\frac{p}{p_0}, F(v)) \geq J(u) + (\frac{p}{p_0}, F(u)) \geq J(u) + (q, F(u)) \forall q, q_j \geq 0.$$

Le point $(u, \frac{p}{p_0})$ est donc un point selle et on a montré l'implication

"minimum global $\Rightarrow$ il existe un point selle".

On s'intéresse maintenant à la condition avec multiplicateurs de Lagrange. On sait que si u est minimum global, alors il existe $(\lambda_1, .., \lambda_m)$ positifs tels que

$$J'(u) + \sum_{i=1}^{i=m} \lambda_i F_i'(u) = 0$$

(ce qui est équivalent à $J'(u) + \sum_{i\in I(u)} \lambda_i F_i'(u) = 0$ où $I(u)$ est l'ensemble des contraintes actives en u, et $\lambda_i = 0$ lorsque la contrainte est inactive).

il s'agit désormais de supposer qu'il existe $(\lambda_1, .., \lambda_m)$ tous positifs ou nuls tels que

$$J'(u) + \sum \lambda_i F_i'(u) = 0.$$

On veut montrer que (u, λ) est un point selle pour le Lagrangien, d'où on déduira que u est un minimum global donc que u est le minimum global.

La fonctionnelle $\mathcal{L}(v, \lambda)$ est convexe. De plus, on a la relation $\lambda_j F_j(u) = 0$, donc

$$\forall v \in K,$$

La condition nécessaire et suffisante est démontrée.

Remarque Dans ce cas ci, on ne peut pas transformer un ensemble de contraintes égalités en un ensemble de contraintes inégalités, sauf si elles sont affines, car si F est convexe, alors $-F$ est concave sauf si elle est affine.

4.4.3 Principe du Min-Max

De la définition d'un point selle (u, p), on déduit deux problèmes d'optimisation associés à $K = \{F_j(u) \leq 0\}$ et à la fonctionnelle $J(v)$. En effet, on a, pour $P = (\mathbb{R}_+)^m$ et $p \in P$:

$$\forall v \in V, \mathcal{L}(u, p) \leq \mathcal{L}(v, p)$$

ce qui implique que, utilisant $\mathcal{L}(v, p) \leq \sup_{q\in P} \mathcal{L}(v, q)$:

$$\forall v \in V, \mathcal{L}(u, p) \leq \sup_{q\in P} \mathcal{L}(v, q).$$

De même,

$$\forall q \in P, \mathcal{L}(u,q) \leq \mathcal{L}(u,p)$$

donc, utilisant cette fois $\mathcal{L}(u,q) \geq \inf_{v\in V} \mathcal{L}(v,q)$, on obtient

$$\forall q \in P, \inf_{v\in V} \mathcal{L}(v,q) \leq \mathcal{L}(u,p).$$

Ceci donne l'idée d'introduire deux fonctionnelles définies par ces inégalités, l'une sur V, l'autre sur P, par

$$\tilde{J}(v) = \sup_{q\in P} \mathcal{L}(v,q), \mathcal{G}(q) = \inf_{v\in V} \mathcal{L}(v,q).$$

Dans le cas étudié, on a $\mathcal{L}(v,q) = J(v) + qF(v)$, donc, si il existe j_0 tel que $F_{j_0}(v) > 0$, alors $\sup_{q\in P} \mathcal{L}(v,q) = +\infty$, et, si on a $\forall j \in \{1,...,m\}, F_j(v) \leq 0$ alors $\sup_{q\in P} \mathcal{L}(v,q) = \max_{q\in} \mathcal{L}(v,q) = \mathcal{L}(v,0) = J(v)$.

Ainsi

$$\tilde{J}(v) = \begin{cases} J(v), v \in K \\ +\infty, v \notin K \end{cases}$$

La minimisation de $\tilde{J}$ est équivalente à celle de J sur K. Ce problème s'apppelle le problème **primal**.

Le problème **dual** est le problème de maximisation de $\mathcal{G}$ sur P.

On remarque que $\forall q \in P, \mathcal{L}(u,q) \leq \mathcal{L}(u,p)$, donc $\sup_{q\in P} \mathcal{L}(u,q) = \mathcal{L}(u,p) = \tilde{J}(u)$. On sait que $\mathcal{L}(u,p) \leq \sup_{q\in P} \mathcal{L}(v,q)$, donc

$$\forall v \in V, \mathcal{L}(u,p) \leq \tilde{J}(v)$$

ce qui s'écrit

$$\forall v \in V, \tilde{J}(u) \leq \tilde{J}(v)$$

On en déduit que u est le minimum de $\tilde{J}$ sur V. De même

$$\forall v \in V, \mathcal{L}(u,p) \leq \mathcal{L}(v,p)$$

donc

$$\inf_{v\in V} \mathcal{L}(v,p) = \mathcal{L}(u,p) = \mathcal{G}(p).$$

Comme $\inf_{v\in V} \mathcal{L}(v,q) \leq \mathcal{L}(u,p)$, on a, $\forall q \in P, \mathcal{G}(q) \leq \mathcal{G}(p)$, donc p **est un maximum de** $\mathcal{G}$. On a ainsi démontré:

$$\min_{v\in V}(\max_{q\in P} \mathcal{L}(v,q)) = \max_{q\in P}(\min_{v\in V} \mathcal{L}(v,q))$$

et le point de min-max est atteint en $v = u, q = p$. **Le point selle est solution du problème de min-max, et la réciproque est vraie**.

Exemple minimisation de la fonctionnelle $J(v) = \frac{1}{2}(Av, v) - (b, v)$ sur l'ensemble convexe $K = \{bV - c \leq 0\}$. Pour être dans le cadre d'application du théorème de Kuhn et Tucker, on suppose la matrice A symétrique définie positive. La fonctionnelle du problème primal est calculée facilement. Celle du problème dual $\mathcal{G}$ est donnée par l'équation sur v

$$\frac{\partial \mathcal{L}}{\partial v}(v, q) = 0$$

qui admet une solution unique car $\mathcal{L}$ est $\alpha-$convexe, où α est la plus petite valeur propre de la matrice $\frac{1}{2}A$.

On trouve $Av - b + {}^tBq = 0$, soit $v = A^{-1}b - A^{-1t}Bq$, donc

$$\mathcal{G}(q) = -\frac{1}{2}({}^tBq, A^{-1t}Bq) + (BA^{-1}b - c, q) - \frac{1}{2}(b, A^{-1}b)$$

qui est strictement concave donc admet un maximum. Le gain dans cette formulation est que les contraintes s'écrivent vraiment simplement: en l'occurence elles sont sous la forme $q \geq 0$.

Chapter 5

Equation de Hamilton-Jacobi-Bellmann

On cherche à minimiser un critère dépendant de variables d'état $x(t), t \in [0,1]$, et d'une commande $u(t)$, sachant que x est solution d'une équation de commande:

$$\dot{x}(t) = f(x(t), u(t), t)$$

avec une valeur initiale $x(0) = x^0$.

Le critère étudié est $J(u) = \int_0^1 g(x(t), u(t), t)dt + C(x(1))$.

On forme le lagrangien du problème, sous les contraintes

$$\begin{array}{l}(i) x(0) - x^0 = 0 \\ (ii) \dot{x}(t) - f(x(t), u(t), t) = 0\end{array}$$

La contrainte (i) admet λ comme multiplicateur, la contrainte (ii) admet $p(t)$ comme multiplicateur (en effet, l'une est continue, l'autre est ponctuelle). Le lagrangien est

$$\mathcal{L}(x, u, \lambda, p) = \int_0^1 g(x(t), u(t), t)dt + C(x(1)) + \int_0^1 p(t)(\dot{x}(t) - f(x(t), u(t), t))dt + \lambda(x(0) - x^0).$$

Par intégrations par parties, on trouve

$$\begin{array}{rl}\mathcal{L}(x, u, \lambda, p) = & \int_0^1 g(x(t), u(t), t)dt + p(1)x(1) + C(x(1)) + \lambda(x(0) - x^0) - p(0)x^0 \\ & - \int_0^1 (\dot{p}(t)x(t) + p(t)f(x(t), u(t), t))dt\end{array}.$$

Les équations de point selle sont $\mathcal{L}_x = 0$, $\mathcal{L}_u = 0$, $\mathcal{L}_p = 0$. On obtient les équations formelles

$$\int_0^1 g_x(x(t), u(t), t)w(t)dt - \int_0^1 (\dot{p}(t) + p(t)f_x(x(t), u(t), t))w(t)dt = 0,$$

$$\int_0^1 g_u(x(t), u(t), t)\tilde{w}(t)dt - \int_0^1 p(t)f_u(x(t), u(t), t)\tilde{w}(t)dt = 0,$$

$$\int_0^1 (\dot{\pi}(t)x(t) + \pi(t)f_x(x(t), u(t), t))dt = 0.$$

De la deuxième, on déduit $g_u(x(t),u(t),t) = p(t)f_u(x(t),u(t),t)$. De la première, on déduit $\dot{p}(t) + f_x(x(t),u(t),t)p(t) = g_x(x(t),u(t),t)$. De la troisième, en effectuant une intégration par parties, on déduit l'équation (ii).

On note que le multiplicateur de Lagrange p est solution d'une équation que l'on appelle équation adjointe de $\dot{x} = f(x,u,t)$.

On remplace l'équation obtenue pour p dans le lagrangien. Alors

$$\begin{array}{rl}\mathcal{L}(x,u,p,t) = & \int_0^1 [g(x(t),u(t),t) - xg_x(t)]dt + p(1)x(1) + C(x(1)) - \int_0^1 p(t)(-x(t)f_x \\ & +f(x(t),u(t),t))dt + \lambda(x(0)-x^0) - p(0)x^0.\end{array}$$

Les expressions ci-dessus ressemblent de manière frappante aux expressions du hamiltonien (intégrale première de l'équation d'Euler). En effet, $g - xg_x$ ressemble a $L - xL_x$.

On introduit alors l'hamiltonien de Pontryaguine:

$$\mathcal{H}(x,u,p,t) = pf(x,u,t) - g(x,u,t).$$

On vérifie $\partial_x\mathcal{H} = pf_x - g_x$ et $\partial_u\mathcal{H} = pf_u - g_u$. L'égalité $g_u = pf_u$ obtenue à partir de la deuxième équation ci-dessus implique que $\partial_u\mathcal{H} = 0$.

L'équation adjointe s'écrit $\dot{p} = -\partial_x\mathcal{H}(x(t),u(t),p(t),t)$. D'autre part, l'équation sur x se réécrit $\dot{x} = \partial_p\mathcal{H}(x(t),u(t),p(t),t)$.

Ainsi les conditions nécessaires d'optimalité impliquent que $(x(t),u(t),p(t))$ est solution du système:

$$\left\{\begin{array}{l}\dot{x}(t) = \partial_p\mathcal{H}(x(t),u(t),p(t),t) \\ \dot{p}(t) = -\partial_x\mathcal{H}(x(t),u(t),p(t),t) \\ 0 = \partial_u\mathcal{H}(x(t),u(t),p(t),t)\end{array}\right.$$

Si on introduit le Lagrangien instantané $L(x,\dot{x},u,p,t) = g(x,u,t) + p(\dot{x} - f(x,u,t))$, alors l'équation de l'état adjoint est

$$\frac{d}{dt}(L_{\dot{x}}) = L_x$$

qui est l'équation d'Euler associée à ce lagrangien. D'autre part, de ce problème, on déduit **l'équation de Hamilton-Jacobi-Bellman**.

Pour écrire cette équation on considère le même problème:

$$\inf\left|\begin{array}{l}J(u) = \int_0^1 g(x(t),u(t),t)dt + C(x(1)) \\ \dot{x}(t) = f(x(t),u(t),t), x(0) = x^0\end{array}\right.$$

et on introduit, comme pour l'étude des problèmes primaux et duaux, la solution de $\inf B(x,u)$. Plus exactement, on considère $\tau \in [0,1]$, y dans l'espace d'arrivée, et x la solution de $\dot{x}(t) = f(x(t),u(t),t), x(\tau) = y$. On introduit

$$V(y,t) = \min\left|\begin{array}{l}\int_\tau^1 g(x(t),u(t),t)dt + C(x(1)) \\ \dot{x}(t) = f(x(t),u(t),t), x(\tau) = y\end{array}\right.$$

Il semble bien sûr que le problème est aussi compliqué que de trouver le minimum pour le problème précédent. Mais on va montrer que V est solution d'une équation aux dérivées partielles.

Pour cela, on cherche $V(y, \tau + \epsilon)$.

$$V(y, \tau + \epsilon) = \min_u [\int_{\tau+\epsilon}^1 g(x(t), u(t), t)dt + c(x(1)), \dot{x}(t) = f(x(t), u(t), t), x(\tau + \epsilon) = y].$$

D"autre part

$$\int_\tau^1 g(x(t), u(t), t)dt = \int_\tau^{\tau+\epsilon} g(x(t), u(t), t)dt + \int_{\tau+\epsilon}^1 g(x(t), u(t), t).$$

Soit u la solution du problème de minimisation pour $\int_\tau^1 g(x(t), u(t), t)dt$. On trouve

$$V(y, \tau) = \min_{v=u(\tau)} [g(y, v, \tau)\epsilon + o(\epsilon) + V(x(\tau + \epsilon), \tau + \epsilon)]$$

$$V(y, \tau) = \min_v [g(y, v, \tau) + V(y + \epsilon f(y, v, \tau) + o(\epsilon), \tau + \epsilon)].$$

Heuristiquement, l'équation s'en déduit aisément en soustrayant à $V(y+\epsilon f(y, v, \tau)+ o(\epsilon), \tau + \epsilon)$ le terme $V(y, \tau + \epsilon)$ et en divisant par ϵ. On a

$$-\partial_\tau V(y, \tau) = \min_v [g(y, v, \tau) + \partial_y V(y, \tau) f(y, v, \tau)].$$

Donc, même si V n'est pas connue, on peut accéder à l'équation différentielle sur V.

Ceci s'exprime dans le

Théorème 5.1 *Si l'équation de Hamilton-Jacobi-Bellman*

$$\frac{\partial V}{\partial t} + \min_v [g(y, v, t) + \frac{\partial V}{\partial y} f(y, v, t)] = 0$$

admet une solution de classe C^1 telle que $V(x, 1) = C(x)$, alors le problème

$$inf \left| \begin{array}{l} J(u) = \int_0^1 g(x(t), u(t), t)dt + C(x(1)) \\ \dot{x}(t) = f(x(t), u(t), t), x(0) = x^0 \end{array} \right.$$

admet une commande optimale $v(x, t)$, qui minimise en v à chaque instant

$$g(x, v, t) + \frac{\partial V}{\partial x}(x, t) f(x, v, t).$$

L'équation de HJB s'écrit $V_t = \max \mathcal{H}(x, -V_x^t, u, t)$.

On considère pour cela $G(x, u, t) = g(x, u, t) + \frac{\partial V}{\partial x}(x, t) f(x, u, t) + \frac{\partial V}{\partial t}(x, t)$. Elle vérifie

$$\forall t \in [0, 1], \min_u G(x, u, t) = 0.$$

On note u^* le point où ce minimum est atteint.

On remarque alors que $\int_0^1 G(x(u), u, t)dt \geq 0$ pour tout u et que

$$\int_0^1 [\frac{\partial V}{\partial x}(x(u), t) f(x(u), u, t) + \frac{\partial V}{\partial t}(x(u), t)]dt = V(x(1), 1) - V(x(0), 0)$$

d'où on déduit

$$0 = J(u^*) - V(x^0, 0) \leq J(u) - V(x^0, 0).$$

et donc bien sûr u^* réalise le minimum de J.

Exemple Dans le cadre de cette équation de Hamilton-Jacobi Bellman, étudions un exemple. C'est un problème de contrôle-commande (objet de la page de garde ...)

On considère un oscillateur, qui peut être excité, et on souhaite le faire passer d'un état donné à un autre état.

Cet oscillateur est régi par l'équation différentielle

$$\ddot{x} + \omega^2(1 - \varepsilon u(t))x = 0,$$

où $x(0)$ et $\dot{x}(0)$ sont connus, et on veut l'amener à l'état $(x(t_1), \dot{x}(t_1))$, où $(x(t_1))^2 + (\dot{x}(t_1))^2 > (x(0))^2 + (\dot{x}(0))^2$. On peut le faire en introduisant la commande $u(t)$ qui vérifie $0 \leq u(t) \leq 1$. Ainsi, on peut faire varier la fréquence d'oscillation du ressort entre ω^2 et $\omega^2(1 - \varepsilon)$.

On est dans la situation de ce chapitre lorsque on écrit cette équation différentielle sous la forme du système différentiel

$$\dot{x} = y, \dot{y} = -(1 - \varepsilon u(t))x.$$

Ainsi $f_1(x, y, u, t) = y$, $f_2(x, y, u, t) = -(1 - \varepsilon u(t))x$ et $\dot{X} = f$. D'autre part, on introduit le multiplicateur de Lagrange (p, q) associé à (x, y). Il n'y a pas d'équation de contrôle sur u.

Le Lagrangien est alors

$$\begin{array}{ll} \mathcal{L}(x, u, \lambda, \mu, k, p, q) & = \int_0^{t_1} (\dot{x}(t) - f_1(x, y, u, t))p(t) + (\dot{y}(t) - f_2(x, y, u, t))q(t))dt \\ & +\lambda(x(0) - x_0) + \mu(y(0) - y_0) + k((x(t_1))^2 + (y(t_1))^2 - 1). \end{array}$$

Après intégration par parties en temps, on trouve les équations adjointes pour p et q de sorte que ce Lagrangien ait un extremum (point selle). Il s'agit de

$$\begin{array}{ll} \mathcal{L}(x, u, \lambda, \mu, k, p, q) = & -\int_0^{t_1} [x\dot{p} + yp + y\dot{q} - (1 - \varepsilon u)xq]dt + x(t_1)p(t_1) + y(t_1)q(t_1) \\ & -x(0)p(0) - y(0)q(0) + \lambda(x(0) - x_0) + \mu(y(0) - y_0) \\ & +k((x(t_1))^2 + (y(t_1))^2 - 1) \end{array}$$

et on en déduit les relations $\dot{p} = (1 - \varepsilon u(t))q$ et $\dot{q} = -p$. En utilisant l'extremalité en t_1, on trouve aussi que $p(t_1) = -kx(t_1)$, $q(t_1) = -ky(t_1)$. De plus, en regardant en $t = 0$, on trouve $p(0) = \lambda$, $q(0) = \mu$, ce qui fait que les conditions initiales ne sont pas connues. Il faudra alors partir de la condition finale.

Le Hamiltonien de Pontriaguine est alors $H = pf_1 + qf_2 = py - q(1 - \varepsilon u)x = py - qx + \varepsilon uxq$. Le principe du maximum de Pontriaguine, énoncé ici sans démonstration (car on se trouve dans le cas discontinu) est de choisir (x, u, p) qui réalise l'extremum de H, et plus précisément on prend le maximum en u sur les contraintes. Lorsque $xq < 0$, ce maximum est atteint en $u = 0$, lorsaue $xq > 0$, il est atteint en $u = 1$. Le contrôle optimal prendra donc les valeurs 0 ou 1 selon le signe de qx.

Si $k = 0$, les conditions finales pour q et p sont 0, et l'équation différentielle de second ordre sur q a ses conditions de Cauchy nulles en $t = t_1$, donc p et q sont nulles, ce qui est impossible car on ne peut pas commander le système. Donc $k \neq 0$, et donc, en divisant q et p par cette constante, on se ramène à $k = 1$. Dans ce cas, pour $t = t_1$, $q(t_1)x(t_1) = -\frac{1}{2}\frac{d}{dt}[(x(t))^2](t_1)$. Si cette quantité est négative, elle le reste dans un intervalle $]t_1 - \varepsilon, t_1[$, donc le contrôle u est égal à 0 dans cet intervalle, et donc l'énergie en t_1 est égale à l'énergie en $t_1 - \varepsilon$, ce qui est contradictoire avec le fait que le contrôle est optimal. Ainsi le contrôle est égal à 1 dans ce voisinage,

donc $-x\dot{x}(t_1) < 0$. On peut positionner le point d'arriver dans le quatrième quadrant $(x > 0, y < 0)$.On écrit $x(t_1) = \cos\alpha, y(t_1) = \sin\alpha,\ \alpha \in]-\frac{\pi}{2}, 0[$. Ainsi on trouve $q(t_1) = \cos(\alpha + \frac{\pi}{2})$, $p(t_1) = \sin(\alpha + \frac{\pi}{2})$. Le point $(p(t), q(t))$ est, dans un voisinage de t_1, sur l'ellipse $q^2 + \frac{p^2}{1-\varepsilon} = a^2 = \sin^2\alpha + \frac{\cos^2\alpha}{1-\varepsilon}$, et le point $(x(t), y(t))$ est sur l'ellipse $x^2 + \frac{y^2}{1-\varepsilon} = b^2 = \cos^2\alpha + \frac{\sin^2\alpha}{1-\varepsilon}$. On contrôle que $a^2 = \frac{1-\varepsilon\sin^2\alpha}{1-\varepsilon}$ et $b^2 = \frac{1-\varepsilon\cos^2\alpha}{1-\varepsilon}$.

Dans ce qui suit, on va construire une trajectoire 'en remontant le sens du temps' à partir du point d'arrivée. PLus précisément, on adopte la démarche suivante:

1. on détermine $T > t_1$ tel que $x(t)$ ne s'annule pas sur $[t_1, T[$ et s'annule en $t = T$. Le contrôle reste $u = 1$.
2. on cherche le premier point $t_2 < t_1$ tel que q s'annule ($u = 1$ sur $]t_2, T[$)
3. on construit $t_3 < t_2$ tel que x s'annule en t_3 ($u = 0$ sur $]t_3, t_2[$)
4. on construit $t_4 < t_3$ tel que q s'annule en t_4 ($u = 1$ sur $]t_4, t_3[$)
5. on construit $\tilde{T} < t_4$ tel que x s'annule en $\tilde{T}$ ($u = 0$ sur $]\tilde{T}, t_4[$).

• Sur $]t_2, T[$:

On commence par donner la forme des fonctions x et q. On trouve $x(t) = b\cos((1-\varepsilon)^{\frac{1}{2}}(t-t_1)+\beta)$, $\dot{x}(t) = y = -b(1-\varepsilon)^{\frac{1}{2}}\sin((1-\varepsilon)^{\frac{1}{2}}(t-t_1)+\beta)$, d'où on déduit $\beta \in]0, \frac{\pi}{2}[$ et $\tan\beta = -\frac{\tan\alpha}{(1-\varepsilon)^{\frac{1}{2}}}$.

On suppose que le système reste dans l'état excité avec $u = 1$. On sait que $q(t) = a\cos((1-\varepsilon)^{\frac{1}{2}}(t-t_1)+\gamma)$ avec $\gamma \in]-\frac{\pi}{2}, 0[$, $a\cos\gamma = -\sin\alpha$, $a(1-\varepsilon)^{\frac{1}{2}}\sin\gamma = \cos\alpha$. On en déduit $\gamma \in]-\frac{\pi}{2}, 0[$ et $\tan\gamma = \frac{1}{(1-\varepsilon)^{\frac{1}{2}}\tan\alpha}$. On contrôle alors que $ab\cos(\gamma-\beta) = \frac{\varepsilon\sin\alpha\cos\alpha}{1-\varepsilon} < 0$, donc, ajoutant le fait que $\gamma - \beta \in]-\pi, 0[$, il vient $\gamma - \beta \in]-\pi, -\frac{\pi}{2}[$. On remarque que $ab\sin(\gamma-\beta) = -\frac{1}{(1-\varepsilon)^{\frac{1}{2}}}$.

Soit T tel que $(1-\varepsilon)^{\frac{1}{2}}(T-t_1)+\beta = \frac{\pi}{2}$. On en déduit que, pour $t \in]t_1, T]$, $\gamma + (1-\varepsilon)^{\frac{1}{2}}(t-t_1)$ décrit $]\gamma, \gamma+\frac{\pi}{2}-\beta] \subset]-\frac{\pi}{2}, 0]$, avec

$$q(T) = a\cos(\frac{\pi}{2}+\gamma-\beta), \dot{q}(T) = -a(1-\varepsilon)^{\frac{1}{2}}\sin(\frac{\pi}{2}+\gamma-\beta).$$

Lorsque l'on introduit $\rho(\alpha)$ et $\omega(\alpha)$ tels que $q(T) = \rho(\alpha)\cos\omega(\alpha)$ et $\dot{q}(T) = \rho(\alpha)\sin\omega(\alpha)$, on obtient $\tan\omega(\alpha) = -(1-\varepsilon)^{\frac{1}{2}}\tan(\frac{\pi}{2}+\gamma-\beta)$, ce qui donne $\tan\omega(\alpha) = -\varepsilon\cos\alpha\sin\alpha$. De plus, $(\rho(\alpha))^2 = a^2\sin^2(\gamma-\beta)+a^2(1-\varepsilon)\cos^2(\gamma-\beta) = \frac{1+\varepsilon^2\sin^2\alpha\cos^2\alpha}{1-\varepsilon\cos^2\alpha}$.

De plus $\dot{x}(T) = -b(1-\varepsilon)^{\frac{1}{2}} = -(1-\varepsilon\cos^2\alpha)^{\frac{1}{2}}$.

On commence à remonter le temps à partir de $t = T$. On écrit

$$x(t) = b\cos((1-\varepsilon)^{\frac{1}{2}}(t-T)+\frac{\pi}{2})$$
$$q(t) = a\cos((1-\varepsilon)^{\frac{1}{2}}(t-T)+\frac{\pi}{2}+\gamma-\beta).$$

Comme $\frac{\pi}{2}+\gamma-\beta \in]-\frac{\pi}{2}, 0[$, on voit qu'en remontant le sens du temps, le premier point òu le produit qx change de signe est atteint pour q au temps t_2 tel que

$$(1-\varepsilon)^{\frac{1}{2}}(t_2-T)+\frac{\pi}{2}+\gamma-\beta = -\frac{\pi}{2}.$$

Le contrôle est $u = 1$ pour $t \in]t_2, T[$, et $\dot{q}(t_2) = a(1-\varepsilon)^{\frac{1}{2}}$. On vérifie aussi que

$x(t_2) = b\cos(\beta-\gamma-\pi+\frac{\pi}{2}) = \rho(\alpha)\frac{b}{a}\cos\omega(\alpha), \dot{x}(t_2) = -b(1-\varepsilon)^{\frac{1}{2}}\sin(\beta-\gamma-\frac{\pi}{2}) = \rho(\alpha)\frac{b}{a}\sin\omega(\alpha).$

• Sur $]t_3, t_2[$:
Le contrôle est $u = 0$, et les trajectoires sont des cercles. On identifie directement

$$x(t) = \rho(\alpha)\tfrac{b}{a}\cos(t - t_2 - \omega(\alpha))$$
$$q(t) = a(1-\varepsilon)^{\frac{1}{2}}\cos(t - t_2 - \tfrac{\pi}{2}).$$

On voit que la première quantité qui s'annule est $x(t)$, au point $t_3 = t_2 + \omega(\alpha) - \frac{\pi}{2}$. On a alors

$$\dot{x}(t_3) = \rho(\alpha)\frac{b}{a}, q(t_3) = -a(1-\varepsilon)^{\frac{1}{2}}\cos\omega(\alpha), \dot{q}(t_3) = a(1-\varepsilon)^{\frac{1}{2}}\sin\omega(\alpha).$$

• Sur $]t_4, t_3[$:
Le contrôle est a nouveau $u = 1$. Les courbes décrites par les points sont

$$(x(t))^2 + \frac{(\dot{x}(t))^2}{1-\varepsilon} = \rho^2(\alpha)\frac{b^2}{a^2(1-\varepsilon)}, (q(t))^2 + \frac{(\dot{q}(t))^2}{1-\varepsilon} = a^2(1-\varepsilon\cos^2\omega(\alpha))$$

ce qui donne

$$x(t) = \rho(\alpha)\tfrac{b}{a}\tfrac{1}{(1-\varepsilon)^{\frac{1}{2}}}\cos((1-\varepsilon)^{\frac{1}{2}}(t-t_3) - \tfrac{\pi}{2})$$
$$q(t) = a(1-\varepsilon\cos^2\omega(\alpha))^{\frac{1}{2}}\cos((1-\varepsilon)^{\frac{1}{2}}(t-t_3) + \beta(\alpha))$$

avec les relations

$$\sin\beta(\alpha) = -\frac{\sin\omega(\alpha)}{(1-\varepsilon\cos^2\omega(\alpha))^{\frac{1}{2}}}, \cos\beta(\alpha) = -\frac{(1-\varepsilon)^{\frac{1}{2}}\cos\omega(\alpha)}{(1-\varepsilon\cos^2\omega(\alpha))^{\frac{1}{2}}}.$$

On trouve donc $\beta(\alpha) \in]-\pi, -\frac{\pi}{2}[$ et $\tan\beta(\alpha) = -\frac{\varepsilon\sin\alpha\cos\alpha}{(1-\varepsilon)^{\frac{1}{2}}}$.

Le point où $q(t)$ s'annule (qui est le premier point inférieur à t_3 où xq change de signe) est donné par

$$(1-\varepsilon)^{\frac{1}{2}}(t_4 - t_3) + \beta(\alpha) = -\frac{3\pi}{2}.$$

On a

$$x(t_4) = -\mu(\alpha)\cos\omega(\alpha), \dot{x}(t_4) = -\mu(\alpha)\sin\omega(\alpha),$$

avec

$$(\mu(\alpha))^2 = (\rho(\alpha)\frac{b}{a})^2(\frac{\cos^2\beta(\alpha)}{1-\varepsilon} + \sin^2\beta(\alpha)) = \frac{(1+\varepsilon^2\cos^2\alpha\sin^2\alpha)}{(1-\varepsilon+\varepsilon^2\cos^2\alpha\sin^2\alpha)(1-\varepsilon\sin^2\alpha)}.$$

• Pour $t \in]\tilde{T}, t_4[$:
le contrôle est alors $u = 0$, les points se déplacent sur des cercles, donc $x(t) = \mu(\alpha)\cos(t - t_4 - \pi + \omega(\alpha))$. Le point où $x(t)$ s'annule est alors $\tilde{T} = t_4 - \frac{\pi}{2} - \omega(\alpha)$, ce qui donne tout de suite $\dot{x}(\tilde{T}) = -\mu(\alpha)$.

Dans ce cas, on a fait un tour complet de l'espce des phases pour $x(t), y(t)$ de $t = \tilde{T}$ à $t = T$. Le gain d'orbite (rapport entre la valeur du point pour les deux temps) est alors

$$\frac{\dot{x}(T)}{\dot{x}(\tilde{T})} = \frac{b(1-\varepsilon)^{\frac{1}{2}}}{\mu(\alpha)} = \frac{1-\varepsilon+\varepsilon^2\cos^2\alpha\sin^2\alpha}{1+\varepsilon^2\cos^2\alpha\sin^2\alpha}$$

en ayant utilisé $1-\varepsilon+\varepsilon^2\cos^2\alpha\sin^2\alpha = (1-\varepsilon\cos^2\alpha)(1-\varepsilon\sin^2\alpha)$.

On vérifie alors que $\frac{\dot{x}(t_2)}{x(t_2)} = \tan\omega(\alpha)$, $\frac{\dot{x}(t_4)}{x(t_4)} = \tan\omega(\alpha)$ et $\lim_{t\to T, t<T} \frac{\dot{x}(t)}{x(t)} = +\infty$, $\lim_{t\to t_3, t>t_3} \frac{\dot{x}(t)}{x(t)} = -\infty$, $\lim_{t\to t_3, t<t_3} \frac{\dot{x}(t)}{x(t)} = +\infty$, $\lim_{t\to \tilde{T}, t>\tilde{T}} \frac{\dot{x}(t)}{x(t)} = -\infty$.

On a ainsi vu que le contrôle est donné par $u(t) = H(\frac{\dot{x}(t)}{x(t)} - \tan\omega(\alpha))$, où H désugne la fonction de Heaviside.

Chapter 6

Approximation de solutions de problèmes d'optimisation

Nous donnons dans cette section des algorithmes d'approximation de solutions de problèmes de minimisation, afin de pouvoir mettre en œuvre des méthodes numériques. Nous nous restreignons aux fonctionnelles convexes, car, si il est difficile de trouver la solution de minimisation de problèmes non convexes, il est encore moins évident de trouver des algorithmes qui convergent vers de telles solutions. Nous étudierons ici les algorithmes de **relaxation**, où on fait les calculs successifs sur chaque variable, les algorithmes de gradient, l'algorithme d'Uzawa, et, chose que je considère comme très importante, la méthode de pénalisation des contraintes, qui est celle que nous avons abordé dans l'étude du problème de Bolza.

6.0.4 Algorithme de relaxation

On suppose que l'on étudie un minimum sans contraintes pour $J(v) = J(v_1, ..., v_N)$, chaque v_j étant élément d'un espace de Hilbert V_j. On suppose J $\alpha-$**convexe différentiable**. Le minimum existe et est unique. On note ce minimum $(u_1, ..., u_N)$.

L'algorithme de relaxation utilise le fait que la restriction de J à V_j, toutes les autres composantes étant fixées, est aussi $\alpha-$convexe. On dit que c'est de la relaxation, car on 'ne traite pas' toutes les composantes en même temps, on en relaxe une sur laquelle on minimise.

Soit $u^0 = (u_1^0, .., u_N^0)$ donné. On écrit une suite $u^n = (u_1^n, .., u_N^n)$. Pour simplifier la compréhension, on suppose $N = 3$, mais le résultat s'étend, avec une petite surchage de notations, pour N quelconque.

On suppose le $n-$ième terme construit $u^n = (u_1^n, u_2^n, u_3^n)$. On résout

$$\inf_{v_1 \in V_1} J(v_1, u_2^n, u_3^n) = J(u_1^{n+1}, u_2^n, u_3^n)$$

puis

$$\inf_{v_2 \in V_2} J(u_1^{n+1}, v_2, u_3^n) = J(u_1^{n+1}, u_2^{n+1}, u_3^n)$$

enfin

$$\inf_{v_3 \in V_3} J(u_1^{n+1}, u_2^{n+1}, v_3) = J(u_1^{n+1}, u_2^{n+1}, u_3^{n+1}).$$

Exemple d'utilisation de la méthode de relaxation On considère la fonctionnelle $J(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2 + x_1 x_2) - \alpha x_1 - \beta x_2$.

Son minimum est atteint en un point (x_1^0, x_2^0) donné par

$$x_1 + \frac{1}{2}x_2 = \alpha, x_2 + \frac{1}{2}x_1 = \beta$$

soit

$$x_1^0 = \frac{4}{3}\alpha - \frac{2}{3}\beta, x_2^0 = \frac{4}{3}\beta - \frac{2}{3}\alpha.$$

L'algorithme de relaxation consiste à partir du point (x, y) quelconque, puis à déterminer le point où $J(x_1, y)$ est minimum (c'est donc $x_1^1 = \alpha - \frac{1}{2}y$), évaluer le point x_2 où $J(x_1^1, x_2)$ est minimum, soit $x_2^1 = \beta - \frac{1}{2}x_1^1$, et donc étudier la suite récurrente

$$x_1^{n+1} = \alpha - \frac{1}{2}x_2^n, x_2^{n+1} = \beta - \frac{1}{2}x_1^{n+1}.$$

On obtient ainsi une relation de récurrence qui est

$$x_1^{n+1} - (\frac{4}{3}\alpha - \frac{2}{3}\beta) = \frac{1}{4}(x_1^n - (\frac{4}{3}\alpha - \frac{2}{3}\beta))$$

qui conduit à

$$x_1^n - (\frac{4}{3}\alpha - \frac{2}{3}\beta) = \frac{1}{4^n}[x_1^1 - (\frac{4}{3}\alpha - \frac{2}{3}\beta)]$$

dont on a la convergence vers la valeur x_1^0.

Un résultat général est le suivant:

Théorème 6.1 *On suppose que J est $\alpha-$convexe différentiable et que, de plus J' est Lipschitzien sur tout borné:*

$$||J'(v) - J'(w)|| \leq C||v - w||.$$

Alors la suite u^n construite par le procédé décrit converge vers la solution de

$$\inf_{(v_1,...,v_n)\in V_1\times...\times V_N} J(v_1, ..., v_N).$$

Preuve On introduit, pour chaque i, la solution du $i-$ème problème intermédiaire. Ainsi

$$u^{n+1,1} = (u_1^{n+1}, u_2^n, u_3^n), u^{n+1,2} = (u_1^{n+1}, u_2^{n+1}, u_3^n), u^{n+1,3} = (u_1^{n+1}, u_2^{n+1}, u_3^{n+1}).$$

On note J_i' la dérivée de J par rapport à l'élément de V_j, tous les autres éléments étant fixes:

$$(J_i'(v_1, ..., v_N), w_i) = \lim_{\varepsilon\to 0} \frac{J(v_1, .., v_i + \varepsilon w_i, ..., v_N) - J(v)}{\varepsilon}.$$

Comme u_i^{n+1} est solution d'un problème de minimisation avec une fonctionnelle $\alpha-$convexe, il est unique et $J_i'(u^{n,i}) = 0$.

Revenons à $N = 3$ pour alléger les notations. En utilisant l'$\alpha-$convexité de J, on écrit

$$J(u^n) - J(u^{n,1}) \geq (J_1'(u^{n,1}), u^n - u^{n,1}) + \frac{\alpha}{2}||u^{n,1} - u^n||^2,$$

$$J(u^{n,1}) - J(u^{n,2}) \geq (J_2'(u^{n,2}), u^{n,1} - u^{n,2}) + \frac{\alpha}{2}||u^{n,2} - u^{n,1}||^2,$$

$$J(u^{n,2}) - J(u^{n,3}) \geq (J_3'(u^{n,3}), u^{n,2} - u^{n,3}) + \frac{\alpha}{2}||u^{n,3} - u^{n,2}||^2,$$

et en sommant les trois égalités et en utilisant les égalités d'Euler partielles

$$J(u^n) - J(u^{n+1}) \geq \frac{\alpha}{2}||u^{n+1} - u^n||^2.$$

• Comme la suite $J(u^n)$ est ainsi décroissante, minorée par $J(u)$, elle converge, donc la différence $J(u^{n+1}) - J(u^n)$ tend vers 0, donc $u^{n+1} - u^n$ tend vers 0. **Notons que cela ne permet pas de conclure sur la convergence de** u^n.

• La suite u^n est bornée. En effet, si elle ne l'était pas, il existerait une sous-suite telle que $||u_{n'}||$ tendrait vers l'infini. Ainsi, comme J est $\alpha-$convexe, $J(u_{n'})$ tendrait vers l'infini, ce qui est impossible car la suite $J(u_n)$ est décroissante. On peut alors appliquer l'inégalité Lipschitz.

• On utilise l'$\alpha-$convexité:

$$(J'(u^n) - J'(u), u^n - u) = (J'(u^n), u^n - u) \geq \alpha|u^n - u|^2$$

puis la définition des dérivées partielles:

$$(J'(u^n), u^n - u) = \sum_i (J_i'(u^n), u_i^n - u_i)$$

puis les N équations d'Euler partielles[1]

$$\begin{array}{ll} (J'(u^n), u^n - u) & = \sum_i (J_i'(u^n) - J_i'(u^{n,i}), u_i^n - u_i) \\ & \leq C \sum_{i \leq N-1} ||u^n - u^{n,i}||||u_i^n - u_i|| \\ & \leq C(N-1)^{\frac{1}{2}}||u^{n+1} - u^n||.||u^n - u||. \end{array}$$

Il vient alors, par l'inégalité d'$\alpha-$convexité:

$$\alpha||u^n - u||^2 \leq C(N-1)^{\frac{1}{2}}||u^{n+1} - u^n||.||u^n - u||.$$

Cela donne

$$||u^n - u|| \leq \frac{C(N-1)^{\frac{1}{2}}}{\alpha}||u^{n+1} - u^n||.$$

On a démontré la convergence de u^n vers u et la majoration entre les deux suites.

[1] noter la différence de notations entre u_i^n et $u^{n,i}$, on l'explicite pour $N = 3$ et on utilise $J_3'(u^{n,3}) = 0$:

$(J'(u^n), u^n - u) = (J_1'(u_1^n, u_2^n, u_3^n) - J_1'(u_1^n, u_2^{n-1}, u_3^{n-1}), u_1^n - u_1) + (J_2'(u_1^n, u_2^n, u_3^n) - J_2'(u_1^n, u_2^n, u_3^{n-1}), u_2^n - u_2)$

ce qui permet d'utiliser le caractère Lipschitz, pour avoir

$(J'(u^n), u^n - u) \leq C[(||u_2^{n-1} - u_2^n||^2 + ||u_3^{n-1} - u_3^n||^2)^{\frac{1}{2}}||u_1^n - u_1|| + ||u_3^n - u_3^{n-1}||.||u_2^n - u_2||] \leq C\sqrt{2}||u^{n+1} - u^n||.||u^n - u||$

grâce à $||u_1^n - u_1|| + ||u_2^n - u_2|| \leq \sqrt{2}(||u_1^n - u_1||^2 + ||u_2^n - u_2||^2)^{\frac{1}{2}}$ ce qui achève la preuve de l'inégalité.

6.1 Algorithmes de descente

On commence par la définition d'une direction de descente. Pour cela, on se place en un point u du domaine d'étude, pour une fonctionnelle J et on cherche des points v tels que $J(v) < J(u)$ et v aussi dans le domaine. On en déduit qu'il suffit que $v - u$ soit une direction admissible pour $\epsilon = 1$.

Ceci nous amène à la

Définition 6.1 *Soit J une fonctionnelle continue sur V, espace de Hilbert et soit K l'espace des contraintes. On dit que d est une direction de descente au point u de K si*

i) d est une direction admissible de $\dot{K}(u)$

ii) Il existe $\rho_0 > 0$ tel que

$$\forall \epsilon \in]0, \rho_0[, J(u + \epsilon d) < J(u).$$

On peut aussi écrire une définition plus générale, qui tienne compte des contraintes égalités:

Définition 6.2 *On suppose que $d \in K(u)$ et que, de plus, il existe $\epsilon_0 > 0$ et $d(\epsilon)$ tels que $d(\epsilon) \to d$ et $\forall \epsilon < \epsilon_0$, $u + \epsilon d(\epsilon) \in K$ (généralisation continue de la direction admissible au sens de Fréchet).*

On dit que d est une direction de descente limite au point u de K si il existe $\epsilon_1 \leq \epsilon_0$ tel que

pour $0 < \epsilon < \epsilon_1$, on a $J(u + \epsilon d(\epsilon)) < J(u)$.

Il est alors clair que

Lemme 6.1 *Si d est une direction de descente, c'est une direction de descente limite.*

Ceci est une conséquence du fait que si d est une direction de descente, $d \in \dot{K}(u)$ donc $d \in K(u)$ et la suite que l'on peut définir est $d(\epsilon) = d$.

On a alors le résultat suivant

Lemme 6.2 *Si J est différentiable en u et si $(J'(u), d) < 0$, d direction admissible continue, alors d est une direction de descente limite.*

Comme d est une direction admissible continue, il existe $d(\epsilon)$ et ϵ_0 tels que, pour $\epsilon < \epsilon_0$, $u + \epsilon d(\epsilon)$ soit dans K. Comme J est différentiable en u, on peut écrire l'égalité de Taylor définissant la dérivabilité au sens de Fréchet:

$$J(u + \epsilon d(\epsilon)) = J(u) + \epsilon[(J'(u), d) + (J'(u), d(\epsilon) - d) + o(1)].$$

On sait que $(J'(u), d) < 0$ et la forme linéaire représentée par $J'(u)$ est continue donc $(J'(u), d(\epsilon) - d) + o(1)$ tend vers 0. Il existe $\epsilon_1 < \epsilon_0$ tel que, pour $\epsilon < \epsilon_1$, $|(J'(u), d(\epsilon) - d) + o(1)| \leq -\frac{1}{2}(J'(u), d)$. Ainsi, pour de tels ϵ on trouve $[(J'(u), d) + (J'(u), d(\epsilon) - d) + o(1)] < 0$, donc $J(u + \epsilon d(\epsilon)) < J(u)$, ce qu'il fallait démontrer.

Remarque: la réciproque est fausse. Il suffit de prendre la fonction $J(x, y) = -(x^4 + y^4)$. Au point $(0, 0)$, toute direction est une direction de descente continue et pourtant la dérivée est la forme différentielle nulle. Si on prend $J(x, y) = x + y -$

$(x^4 + y^4)$, la forme linéaire dérivée est $(J'(0,0), h_1, h_2) = h_1 + h_2$, et toute direction telle que $h_1 + h_2 \leq 0$ est une direction de descente.

La définition où on étudie le point $u+\epsilon d$ n'est pas adaptée aux contraintes égalités, pour lesquelles la bonne notion (pour une direction admissible) est la notion de direction admissible continue. En fait, avoir à la fois le paramètre ϵ et la direction $d(\epsilon)$ qui varient n'est pas pratique dans l'écriture d'un algorithme. On écrit donc un résultat, qui permet de s'affranchir du cas des contraintes égalité:

Proposition 6.1 *Soit J une fonctionnelle différentiable sur un espace de Hilbert V et F une fonctionnelle différentiable. Le problème:*

$$\begin{cases} infJ(v) \\ v \in K, F(v) = 0 \end{cases}$$

est équivalent, pour tous les points u où $F(u) = 0$, $F'(u) \neq 0$, à un problème de minimisation sur $(F'(u))^\perp$ de la forme

$$\{v + tF'(u) \in K, t = g(v), v \in (F'(u))^\perp\}$$

pour la fonctionnelle $\tilde{J}(v) = J(v + g(v)F'(u))$.

Ceci est un résultat de réduction des variables. On en verra l'utilisation plus loin, lorsqu'on étudiera l'algorithme de gradient réduit.

Comme $F'(u)$ est non nul, il définit une droite vectorielle dans l'espace de Hilbert, qui est un fermé convexe. Ainsi tout point w de l'espace de Hilbert se projette en un point $\phi(w)F'(u)$, et on a $w - \phi(w)F'(u)$ dans l'espace orthogonal à $F'(u)$.

L'égalité $F(v+u+tF'(u)) = 0$ a pour solution $t = 0, v = 0$ car u vérifie $F(u) = 0$. Pour chaque v dans $(F'(u))^\perp$, on trouve, par le théorème des fonctions implicites (dû à $\partial_t(F(v+u+tF'(u))) = ||F'(u)||^2 > 0$) une unique solution de l'égalité ci-dessus, soit $t = g(v)$. Alors, au voisinage de u, on étudie pour tout v dans l'intersection I_u d'une boule de petit rayon et de $(F'(u))^\perp$, la fonctionnelle sous les contraintes. On voit alors que pour tout v dans I_u, le problème de minimisation s'écrit $u + v + tF'(u) \in K$ et $u+v+tF'(u) \in \{F(w) = 0\}$, soit $u+v+tF'(u) \in K$ et $t = g(v)$, soit $u+v+g(v)F'(u) \in K$. Ainsi on s'est ramené à la fonctionnelle $\tilde{J}(v) = J(u+v+g(v)F'(u))$ et au problème

$$\begin{cases} \inf\tilde{J}(v) \\ v \in I_u \\ v + g(v)F'(u) \in K \end{cases}$$

La contrainte égalité a ainsi été résolue. On note cependant que résoudre un problème numérique en utilisant le théorème des fonctions implicites est quasiment impossible, sauf si les contraintes sont affines.

6.2 Cas classiques d'algorithmes de descente

Un algorithme de descente est donné par la définition suivante:

Définition 6.3 *Un algorithme de descente est une suite de points de $V \times V \times \mathbb{R}_+$, qui s'écrit*

$$(u_n, d_n, l_n)$$

telle que

i) d_n *est une direction de descente en* x_n *pour* J*, associée à* ρ_n *tel que* $J(u_n+\epsilon d_n) < J(u_n)$ *pour* $0 < \epsilon < \rho_n$

ii) l_n *est un pas vérifiant* $0 < l_n < \rho_n$

iii) $u_{n+1} = u_n + l_n d_n$*.*

Les algorithmes les plus courants sont des algorithmes de recherche **linéaires**. En effet, ces algorithmes conduisent, une fois la direction de descente choisie, à la recherche d'une valeur réelle qui est la valeur du pas. On suppose ainsi que, à chaque étape, la direction de descente d_n soit choisie. Nous allons décrire dans ce qui suit un certain nombre d'algorithmes.

Dans tous les cas, on notera, par souci de simplicité

$$\phi(\epsilon) = J(u + \epsilon d). \tag{6.2.1}$$

6.2.1 Pas optimal

Définition 6.4 *Pour chaque couple* (u, d)*, on note,* **si elle existe***, la solution du problème*

$$Min_{\epsilon \geq 0} J(u + \epsilon d) = Min_{\epsilon \geq 0} \phi(\epsilon).$$

Il s'appelle le pas optimal.

L'algorithme dit du pas optimal conduit à associer, à chaque (u_n, d_n), le point ϵ_n construit par la définition 6.4. C'est l'algorithme le plus satisfaisant, en théorie, mais il conduit à déterminer la solution d'un problème de minimisation chaque fois.

6.2.2 Pas de Curry

Le pas de Curry est donné par:

Définition 6.5 *Le pas de Curry est le premier extremum local de* ϕ*, soit encore*

$$l_c = inf\{\epsilon > 0, \phi'(\epsilon) = 0\}.$$

Alors $\phi(l_c) < \phi(0)$*, et pour* $0 \leq \epsilon \leq l_c$*,* $\phi(\epsilon) \geq \phi(l_c)$*.*

Comme ϕ' ne s'annule pas sur $]0, l_c[$, ϕ' garde le même signe sur cet intervalle, soit $\phi' \geq 0$ ou $\phi' \leq 0$. Dans le cas $\phi' \geq 0$, on vérifie que $\phi(\epsilon) - \phi(0) \geq \int_0^\epsilon \phi'(t)dt$, ainsi $\phi(\epsilon) \geq \phi(0)$, contradiction avec le fait que d soit une direction de descente.

Ainsi $\phi'(\epsilon) \leq 0$ sur $[0, l_c]$. Pour $\epsilon \in [0, l_c]$, on vérifie

$$\phi(l_c) - \phi(\epsilon) = \int_\epsilon^{l_c} \phi'(t)dt$$

donc, pour $0 \leq \epsilon \leq l_c$, on trouve $\phi(l_c) \leq \phi(\epsilon)$.

Dans le cas où l_c est un point d'inflexion, on ne peut bien sûr pas conclure sur le fait que l_c soit un minimum local. En revanche, on sait que pour cette valeur, $\phi(l_c)$ est le minimum de ϕ sur $[0, l_c]$.

6.2.3 Pas de Goldstein

Définition 6.6 *On dit que l_g est un pas de Goldstein si il existe m_1, m_2 tels que $0 < m_1 < m_2 < 1$ tels que*

$$\begin{cases} \phi(l_g) \leq \phi(0) + m_1 l_g \phi'(0) \\ \phi(l_g) \geq \phi(0) + m_2 l_g \phi'(0) \end{cases}$$

C'est un pas pseudo optimal, qui vérifie

$$0 < m_1 \leq \frac{\phi(l_g) - \phi(0)}{l_g \phi'(0)} \leq m_2 < 1.$$

Exemples:
figure 1 figure 2

Dans la situation de la figure 2, il n'existe pas de pas de Goldstein, mais en revanche on a $\forall \epsilon \in [0, \rho_0], \phi(\epsilon) \leq \phi(0) + \epsilon\phi'(0)$, ce qui fait que l'on peut choisir pour ϵ la valeur ρ_0, même si cela a un inconvénient, comme on le verra ci-dessous.

La situation importante est la situation où il existe au moins ϵ_1, $0 < \epsilon_1 < \rho_0$ tel que

$$\phi(0) + \epsilon_1 \phi'(0) < \phi(\epsilon_1) < \phi(0).$$

Dans ce cas, on a la

Proposition 6.2 *i) Si $\phi(\epsilon) \leq \phi(0) + \epsilon\phi'(0)$ pour tout $\epsilon \in [0, \rho_0]$, il n'existe pas de pas de Goldstein.*

ii) Dans le cas contraire, il existe $m_1, m_2 \in]0, 1[$, $m_1 < m_2$ tel que l'ensemble des points l vérifiant les inégalités de la définition 6.6 soit non vide.

iii) Toujours dans le cas contraire, il existe $\epsilon_2 > 0$ et $M > 0$ (dans le cas où la fonctionnelle admet un minimum) tel que, pour tout l_g, $\epsilon_2 \leq l_g \leq M$.

Selon le point iii), il y a une borne supérieure pour l_g, et l_g n'est pas trop petit. Ces deux remarques sont importantes, et en particulier si on avait $\phi(\epsilon) \leq \phi(0) + \epsilon\phi'(0)$ on n'aurait pas de majorant a priori de ϵ.

Preuve:

On note $m = \frac{\phi(\epsilon_1) - \phi(0)}{\epsilon_1 \phi'(0)}$. On sait que $m \in]0, 1[$ et si on choisit $m_1 < m < m_2$, l'ensemble des pas de Goldstein associés à $[m_1, m_2]$ est non vide. En effet, définissons $h(\epsilon) = \frac{\phi(\epsilon) - \phi(0)}{\epsilon\phi'(0)}$ et, par continuité, $h(0) = 1$. La fonction h est une fonction continue.

Par le théorème des valeurs intermédiaires, comme $h(0) = 1$ et $h(\epsilon_1) = m$, l'image réciproque dans $[0, \epsilon_1]$ de $[m, m_2] \subset [m, 1]$ est non vide. Tout point de $[m, m_2]$ a au moins un antécédent par h, qui est un pas de Goldstein.

D'autre part, l'image réciproque de $]m_2, 1]$ contient un voisinage $[0, \epsilon_2]$ de $\epsilon = 0$ puisque $h(0) = 1$. Ainsi on a $\forall \epsilon \in h^{-1}(]m_2, 1]), \epsilon$ n'est pas un pas de Goldstein, donc si l_g est un pas de Goldstein, $l_g \geq \epsilon_2$.

Enfin, on ne peut pas avoir $\epsilon \to \infty$. En effet, cela impliquerait que pour tout ϵ, ou au moins pour une suite ϵ_n tendant vers $+\infty$, la relation

$$\frac{\phi(\epsilon_n) - \phi(0)}{\epsilon_n \phi'(0)} \geq m_1$$

soit $\phi(\epsilon_n) \leq \phi(0) + m_1\phi'(0)\epsilon_n$. Il existe donc une suite ϵ_n telle que $J(u+\epsilon_n d) \to -\infty$, et le minimum n'existe pas.

6.2.4 Pas de Wolfe

Définition 6.7 *l_w est un pas de Wolfe si il existe m_1, m_2, $0 < m_1 < m_2 < 1$ tels que*

$$\begin{cases} \phi(l_w) \leq \phi(0) + m_1 l_w \phi'(0) \\ \phi'(l_w) \geq m_2 \phi'(0) \end{cases}$$

Ce pas de Wolfe a les mêmes propriétés que celui de Goldstein; en effet on a

Proposition 6.3 *i) Si $\phi'(\epsilon) \leq \phi'(0)$ pour tout $\epsilon \in [0, \rho_0[$, il n'existe pas de pas de Wolfe. (On note que cela implique qu'il n'existe pas de pas de Goldstein).*

ii) Dans le cas contraire, il existe (m_1, m_2) tels que l'ensemble des points l vérifiant les inégalités de la définition 6.7 est non vide.

iii) Il existe $\epsilon_2' > 0$ et $M > 0$ tels que $l_w \geq \epsilon_2'$, $l_w \leq M$.

Preuve

Si ϵ_1 donné tel que $\phi'(\epsilon_1) > \phi'(0)$, alors $m = \frac{\phi'(\epsilon_1)}{\phi'(0)} < 1$ et donc on choisit $m_2 \in]m, 1[$. Comme $\frac{\phi'(0)}{\phi'(0)} = 1$ et que la fonction $\epsilon \to \frac{\phi'(\epsilon)}{\phi'(0)}$ est continue, par le théorème des valeurs intermédiaires, tout point de $]m, 1]$ a au moins un antécédent, et l'image réciproque de $]m_2, 1]$ contient un voisinage de 0. On prend un point l de $(\phi')^{-1}[m_2\phi'(0), m\phi'(0)]$, ainsi $l \geq \epsilon_2'$.

La fonction $\epsilon \to \frac{\phi(\epsilon)-\phi(0)}{\phi'(0)\epsilon}$ est continue sur le compact $[\epsilon_2', \rho_0]$ et ne s'annule pas sur cet intervalle, donc

$$\inf_{\epsilon \in [\epsilon_2', \rho_0]} \frac{\phi(\epsilon) - \phi(0)}{\phi'(0)\epsilon} = \alpha > 0.$$

Si on choisit $0 < m_1 < \alpha$, on trouve que pour tout $\epsilon \in [\epsilon_2', \rho_0]$, $\frac{\phi(\epsilon)-\phi(0)}{\phi'(0)\epsilon} \geq \alpha$, donc ϵ est un pas de Wolfe.

Enfin, si on était dans le cas $\rho_0 = +\infty$ et si il existait une suite de pas de Wolfe qui tendait vers $+\infty$, il existe donc ϵ_n telle que $\phi(\epsilon_n) \leq \phi(0) + m_1\epsilon_n\phi'(0)$, donc $J(u + \epsilon_n d) \to -\infty$ et le minimum n'existe pas.

6.3 Résultats de convergence

On a le

Théorème 6.2 *On suppose J continuement différentiable et on suppose que l'on a un algorithme de descente (u_n, d_n, l_n) vérifiant $||d_n|| = 1$. On suppose qu'il existe $\alpha > 0$ tel que*

$$(H) \quad (J'(u_n), d_n) \leq -\alpha |d_n| |J'(u_n)| = -\alpha |J'(u_n)|.$$

i) Si, à chaque étape n, l_n est un pas de Curry ou de Wolfe, et si la suite u_n converge, elle converge vers une solution de $J'(u) = 0$.

ii) si l_n est un pas de Goldstein ou de Wolfe, alors $J(u_n) \to -\infty$ ou lim inf$||J'(u_n)|| = 0$.

On démontre ce théorème.

Preuve de i)

On suppose que la suite u_n converge (dans le cas du pas de Curry). Ainsi, comme $u_{n+1} - u_n$ tend vers 0, l_n tend vers 0 puisque d_n est de norme 1. D'autre part, comme J est continuement différentiable, la dérivée de ϕ est

$$\phi'(\epsilon) = (J'(u_n + \epsilon d_n), d_n).$$

Dans le cas où l_n est le pas de Curry, on a $(J'(u_n + l_n d_n), d_n) = 0$. D'autre part

$$(J'(u_n + l_n d_n) - J'(u_n), d_n) = -(J'(u_n), d_n) \geq \alpha ||J'(u_n)||.$$

On a l'inégalité

$$|(J'(u_n + l_n d_n) - J'(u_n), d_n)| \leq ||J'(u_n + l_n d_n) - J'(u_n)||$$

On trouve alors

$$||J'(u_n)|| \leq \frac{1}{\alpha} ||J'(u_n + l_n d_n) - J'(u_n)||.$$

Comme J' est continue, on vérifie que $J'(u_{n+1}) - J'(u) - (J'(u_n) - J'(u))$ tend vers 0 dans l'espace des formes linéaires, donc on en déduit que $J'(u_n)$ tend vers 0.

D'autre part, la suite $J(u_n)$ est strictement décroissante (par construction) donc comme u_n converge vers u, la suite $J(u_n)$ converge vers $J(u)$ et la suite $J'(u_n)$ converge vers $J'(u)$. On en déduit $J'(u) = 0$. Le point i) est démontré pour le pas de Curry.

Démontrons le point i) pour la règle de Wolfe. On suppose que u_n converge. Par continuité $J(u_n)$ converge vers $J(u)$ et $J'(u_n)$ converge vers $J'(u)$. On a $(J'(u_n), d_n) \in [-\alpha||J'(u_n)||, 0]$ donc toute suite extraite convergente de $(J'(u_n), d_n)$ converge vers une limite l dans l'intervalle $[-\alpha||J'(u)||, 0]$.

On utilise la deuxième inégalité du pas de Wolfe. On a alors $(J'(u_{n+1}), d_n) \geq m_2(J'(u_n), d_n)$. On note que si on prend une suite extraite convergente de $(J'(u_n), d_n)$, notée $(J'(u_{\phi(n)}), d_{\phi(n)})$, la suite $(J'(u_{\phi(n)+1}), d_{\phi(n)})$ converge aussi vers l car la différence est majorée par un terme tendant vers 0 par continuité de J' et convergence de la suite u_n. Ainsi, l qui est négatif vérifie l'inégalité $l \geq m_2 l$, soit $(1 - m_2)l \geq 0$ donc $l = 0$.

On a démontré le point i) pour la règle de Wolfe.

Démontrons le point ii). Pour cela, suposons que $\liminf ||J'(u_n)|| = \alpha_0 > 0$. Alors il existe N assez grand tel que, pour tout $n \geq N$ on ait $||J'(u_n)|| > \frac{\alpha_0}{2}$. Si cela n'était pas le cas, il existerait un nombre infini de termes de cette suite de nombres positifs qui sont compris entre 0 et $\frac{\alpha_0}{2}$, donc il existerait une sous-suite extraite de cette suite qui convergerait vers une valeur comprise entre 0 et $\frac{\alpha_0}{2}$, contradictoire avec l'hypothèse que α_0 est la plus petite des limites des suites extraites.

On en déduit alors

$$\frac{\alpha \alpha_0}{2} ||u_{n+1} - u_n|| \leq J(u_n) - J(u_{n+1}).$$

Si $J(u_n)$, qui est une suite décroissante, ne tend pas vers $-\infty$, alors elle tend vers une limite l et la série de terme général $(J(u_n) - J(u_{n+1}))$ est une série convergente, donc la somme de la série $u_1 + \sum_n (-u_n + u_{n+1})$ existe, et on la note u, qui est la limite de la suite u_n. Deux cas se présentent: l'application de la règle de Wolfe et de celle de Goldstein.

i) Règle de Wolfe. D'après le i), comme u_n a une limite, notée u, on sait que la suite $J'(u_n)$ est convergente et que sa limite est $J'(u) = 0$, ce qui est contradictoire avec l'hypothèse que la limite inf de $||J'(u_n)||$ est nulle.

On a donc démontré que $\liminf||J'(u_n)|| = \alpha_0 > 0 \Rightarrow J(u_n) \to -\infty$. On en déduit que si $J(u_n)$ converge vers une limite finie, alors $\liminf||J'(u_n)|| = 0$. Notons qu'on ne peut pas conclure directement que la suite u_n converge.

ii) Règle de Goldstein

On suppose donc que la suite $J(u_n)$ converge vers une limite l. On suppose aussi que $\liminf||J'(u_n)|| = \alpha_0 > 0$. Ceci implique que la suite u_n est convergente, et sa limite est notée u. Par continuité de J et de J', $J(u_n)$ tend vers $J(u)$ et $J'(u_n)$ tend vers $J'(u)$. Contrairement à la règle de Wolfe, on n'a pas d'autre information sur la dérivée. En effet, l'information sur la limite inf nous apprend que $||J'(u_n)|| \geq \frac{\alpha_0}{2}$ pour $n \geq n_0$, mais on n'a pas le même résultat pour $(J'(u_n), d_n)$.

On sait, par la règle de Goldstein, que

$$\frac{J(u_n) - J(u_{n+1})}{(J'(u_n), u_n - u_{n+1})} \in [m_1, m_2].$$

Dans le cas où on suppose que J' est uniformément continue sur un borné contenant u, alors pour n assez grand comme la suite u_n converge vers u, les points u_n sont dans ce borné. Ainsi on aura

$$-J(u_n) + J(u_{n+1}) = \int_0^1 (J'(u_n + \theta(u_{n+1} - u_n)), u_{n+1} - u_n) d\theta$$

donc on en déduit que

$$|J(u_n) - J(u_{n+1}) - (J'(u_n), u_{n+1} - u_n)| \leq \epsilon||u_{n+1} - u_n||, n \geq n_\epsilon.$$

Ainsi, divisant les deux membres par $(J'(u_n), u_{n+1} - u_n)$ et utilisant l'inégalité $(J'(u_n), d_n) \leq -\alpha||J'(u_n)||$, dans le cas où $J'(u_n)$ ne tend pas vers 0, pour $n \geq n_\epsilon$,

$$|\frac{J(u_n) - J(u_{n+1})}{(J'(u_n), u_{n+1} - u_n)} - 1| \leq \frac{||u_{n+1} - u_n||}{|(J'(u_n), u_{n+1} - u_n)|}\epsilon = \frac{\epsilon}{|(J'(u_n), d_n)|} \leq \frac{\epsilon}{\alpha||J'(u_n)||} \leq \frac{2\epsilon}{\alpha\alpha_0}.$$

On en déduit que le quotient $\frac{J(u_n)-J(u_{n+1})}{(J'(u_n),u_{n+1}-u_n)}$ tend vers 1. Comme ce quotient appartient à $[m_1, m_2]$ et que $m_2 < 1$ il y a contradiction. Le résultat est démontré sous l'hypothèse d'uniforme continuité ou de continuité dans un borné en dimension finie.

Remarque 1 : le i) peut s'étendre à toute sous-suite convergente dans le cas où la suite l_n tend vers 0. On note que ceci n'implique pas que la suite u_n converge : exemple si $d_n = e_1$ pour tout n et si $l_n = \frac{1}{n}$ alors il n'y a pas convergence de u_n.

Remarque 2 :Pour la règle de Goldstein, il suffit, en dimension finie que J vérifie l'une des deux conditions suivantes :

(*) J' est uniformément Lipschitz sur tout borné

(**) la fonctionnnelle J est deux fois Fréchet dérivable à dérivée continue (qui implique la condition (*) et qui se retrouve le plus fréquemment)

6.4 Algorithmes de gradient

6.4.1 Définition

On commence par le résultat suivant, qui nous donnera l'algorithme de gradient

Proposition 6.4 *Soit $\phi(l,d) = J(u_n + ld)$. On suppose $J'(u_n) \neq 0$.*

$$\inf_{||d||=1} \phi'(0,d) = -||J'(u_n)||$$

et ce minimum est atteint pour $d = -\frac{J'(u_n)}{||J'(u_n)||}$.

On note que $\phi'(0,d) = -(J'(u_n), d)$. On a, par l'inégalité triangulaire

$$|\phi'(0,d)| \geq -||d||||J'(u_n)||$$

et l'égalité est atteinte dans le cas d'égalité pour Cauchy-Schwartz, soit pour d colinéaire à $J'(u_n)$, ce qui correspond au vecteur indiqué.

La direction du gradient est, parmi les directions de norme 1, la meilleure pour le taux de décroissance de la fonctionnelle. C'est par ce type d'algorithme que l'on recherche la solution de $f = 0$ par la méthode de Newton.

6.4.2 L'algorithme de gradient à pas optimal

On démontre le

Théorème 6.3 *Soit J une fonctionnelle $\alpha-$convexe sur un espace de Hilbert H, telle que J' est uniformément continue sur tout borné. La suite, définie par la relation*

$$u^{n+1} = u^n - \mu_n J'(u^n),$$

où μ_n est la solution unique de $J'(u^n - \mu J'(u^n)) = 0$ qui s'appelle l'algorithme de gradient à pas optimal, converge vers l'unique valeur qui rend minimum la fonctionnelle J.

L'agorithme de gradient à pas optimal est défini par la suite

$$u^{n+1} = u^n - \mu J'(u^n)$$

et on cherche $u^{n+1} = \inf_{\mu \in \mathbf{R}} J(u^n - \mu J'(u^n))$. Il est clair que la dérivée de $\phi(\mu) = J(u^n - \mu J'(u^n))$ est donnée par

$$\phi'(\mu) = -(J'(u^n - \mu J'(u^n)), J'(u^n)).$$

Comme J est $\alpha-$convexe, lorsque $J'(u^n) \neq 0$ (ce qui correspond au cas où on n'a pas atteint le point de minimum) on a ϕ $\alpha(||J'(u_n)||^2)-$convexe. En effet

$$\begin{aligned}(J'(u^n - \mu_1 J'(u^n)) - J'(u^n - \mu_2 J'(u^n)), u^n - \mu_1 J'(u^n) - u^n + \mu_2 J'(u^n)) \\ \geq \alpha||u^n - \mu_1 J'(u^n) - u^n + \mu_2 J'(u^n))||^2 \\ = \alpha(\mu_2 - \mu_1)^2 ||J'(u^n)||^2.\end{aligned}$$

En remplaçant la différence, on trouve

$$(\phi'(\mu_1) - \phi'(\mu_2), \mu_1 - \mu_2) \geq \alpha(\mu_2 - \mu_1)^2 ||J'(u^n)||^2$$

d'où l'$\alpha-$convexité de ϕ. Le problème de minimisation admet donc une solution unique μ_n. De plus, μ_n est solution de $\phi'(\mu_n) = (J'(u^n - \mu_n J'(u^n)), J'(u^n)) = 0$, on en déduit que $(J'(u^{n+1}), J'(u^n)) = 0$ et deux directions de descente successives sont orthogonales.

La démonstration du théorème 6.3 s'appuie sur l'inégalité de convexité

$$J(u^n) - J(u^{n+1}) \geq (J'(u^{n+1}), u^n - u^{n+1}) + \frac{\alpha}{2}||u^{n+1} - u^n||^2$$

et sur l'égalité $u^{n+1} - u^n = -\mu_n J'(u^n)$, ce qui annule le premier terme de l'inégalité ci-dessus car $(J'(u^{n+1}), J'(u^n)) = 0$.

On a donc démontré que $J(u^n) - J(u^{n+1}) \geq \frac{\alpha}{2}||u^n - u^{n+1}||^2$. La suite $J(u^n)$ est décroissante, bornée par le minimum de J, donc elle converge, donc on en déduit que $||u^n - u^{n+1}||$ tend vers 0.

D'autre part, on vérifie que

$$||J'(u^n)||^2 = (J'(u^n), J'(u^n) - J'(u^{n+1}))$$

car deux directions successives sont orthogonales. Ainsi

$$||J'(u^n)|| \leq ||J'(u^n) - J'(u^{n+1})||.$$

D'autre part, la suite u^n est bornée. En effet, si elle ne l'était pas, il existerait une sous suite $u^{\phi(n)}$ qui tendrait, en norme, vers $+\infty$, et comme la fonctionnelle J est $\alpha-$convexe, elle est infinie à l'infini et la suite $J(u^{\phi(n)})$ tendrait vers $+\infty$, contradiction. Dans ce cas, en utilisant l'uniforme continuité sur une boule fermée qui contient tous les termes de la suite u^n, on en déduit que $||J'(u^n) - J'(u^{n+1})|| \leq C||u^n - u^{n+1}||$. On a alors

$$||J'(u^n)|| \leq C||u^n - u^{n+1}|| \leq (\frac{2}{\alpha})^{\frac{1}{2}} C\sqrt{J(u^n) - J(u^{n+1})}.$$

On en déduit la convergence de la suite $J'(u^n)$ vers 0. On note u le point où J est minimale. Par la coercivité

$$(J'(u^n) - J'(u), u^n - u) \geq \alpha||u^n - u||^2.$$

Par l'inégalité de Cauchy-Schwarz, on trouve

$$\alpha||u^n - u||^2 \leq ||J'(u^n)||.||u^n - u||$$

ce qui implique

$$||u^n - u|| \leq \frac{1}{\alpha}||J'(u^n)||$$

donc

$$||u^n - u|| \leq \frac{1}{\alpha}(\frac{2}{\alpha})^{\frac{1}{2}} C\sqrt{J(u^n) - J(u^{n+1})}$$

et donc la suite u^n converge vers u.

Proposition 6.5 *Pour que les hypothèses du théorème 6.3 soient vérifiées, il suffit que J vérifie*

i) soit J fonctionnelle $\alpha-$convexe dérivable, J' continue en dimension finie

ii) soit J fonctionnelle $\alpha-$convexe dérivable, J' Lipschitzienne sur tout borné en dimension infinie

iii) soit J est une fonctionnelle deux fois Fréchet dérivable, telle que la dérivée seconde soit autoadjointe et vérifie

$$m||w||^2 \leq (J''(u)w, w) \leq M||w||^2$$

avec $m > 0$.

On remarque que ces conditions sont telles que iii) $\rightarrow$ ii) $\rightarrow$ i).

Ce résultat provient de l'uniforme continuité sur un compact d'une fonctionnelle continue en dimension finie.

6.4.3 Algorithme de gradient à pas constant

Théorème 6.4 *On a convergence de l'algorithme de gradient à pas fixe,* **seulement si J' est Lipschitzien sur V tout entier**, *lorsque $0 < \mu < \frac{2\alpha}{C}$.*

La preuve est plus simple. On écrit $u^{n+1} - u^n = -\mu J'(u^n)$. Ainsi, soit u la solution On trouve $u^{n+1} - u = u^n - u - \mu(J'(u^n) - J'(u))$. On utilise un argument de type "théorème du point fixe". Ainsi

$$\begin{aligned} ||u^{n+1} - u||^2 &= ||u^n - u||^2 - 2\mu(J'(u^n) - J'(u), u^n - u) + \mu^2||J'(u^n) - J'(u)||^2 \\ &\leq (1 - 2\mu\alpha + \mu^2 C^2)||u^n - u||^2 \end{aligned}$$

où C est la constante de Lipschitz de J' sur tout l'espace de Hilbert. La démonstration est terminée car la suite $||u^n - u||$ est alors majorée par une suite géométrique convergeant vers 0.

6.4.4 Taux de convergence de l'algorithme du gradient en dimension finie

Le but de cette section est de démontrer le résultat suivant:

Théorème 6.5 *On suppose J de classe C^2, $\alpha-$convexe et on suppose que le Hilbert V est de dimension finie d. Soit u la valeur du point où J atteint son minimum. On note λ_{max} et λ_{min} les plus grande et plus petite valeur propre de la matrice hessienne (définie positive) J"(u). On désigne par*

$$\gamma = \frac{\lambda_{max}}{\lambda_{min}}.$$

Cette valeur s'appelle le conditionnement de J"(u). On note $\beta = \frac{\gamma-1}{\gamma+1}$, et si β est proche de 1, l'algorithme peut converger très lentement. On dit dans ce cas que la matrice J"(u) est mal conditionnée.

i) Lorsque J est quadratique, l'algorithme de gradient vérifie l'inégalité:

$$||u^{n+1} - u||_{J''(u)} \leq \beta^n ||u^1 - u||_{J''(u)}.$$

ii) Lorsque J est quelconque, l'algorithme de gradient vérifie l'inégalité

$$\forall \beta > \frac{\gamma - 1}{\gamma + 1}, \exists n_0,$$

$$\forall n \geq n_0, ||u^{n+n_0} - u|| \leq D\beta^n ||u^{n_0} - u||.$$

Ce théorème est très important de manière théorique, mais la valeur du conditionnement est difficilement accessible donc il est difficile à utiliser en pratique. Sa démonstration se fait en deux temps:

i) on le démontre pour $J(x) = \frac{1}{2}(Ax, x)$

ii) on le démontre dans le cas général.

On se place d'abord dans le cas $J(x) = \frac{1}{2}(Ax, x)$.

Pour toute fonctionnelle quadratique, on peut se ramener à ce cas car si A est définie positive symétrique, on note x_0 la solution de $Ax = b$ et la forme quadratique (qui par définition a pour dérivée seconde A) vérifie $Q(x) - \frac{1}{2}(Ax, x)$ est affine continue, donc par le théorème de représentation de Riesz, $Q(x) - \frac{1}{2}(Ax, x) - Q(0)$ étant linéaire continue, il existe b telle que $Q(x) - \frac{1}{2}(Ax, x) - Q(0) = (b, x)$. On vérifie alors que $Q(x) - Q(0) = \frac{1}{2}(A(x - x_0), x - x_0) - \frac{1}{2}(Ax_0, x_0)$.

Une fois la représentation précédente obtenue, on introduit $\phi(l) = J(u - lJ'(u))$. On a

$$\phi(l) = J(u - lAu) = \frac{1}{2}(Au - lA^2u, u - lAu) = \frac{1}{2}(Au, u) - l(A^2u, u) + \frac{l^2}{2}(A^2u, Au).$$

On en déduit que la valeur du pas optimal est $l = \frac{(Au, Au)}{(A^2u, Au)}$ et que la valeur de ϕ est

$$\frac{1}{2}[(Au, u) - \frac{(Au, Au)^2}{(A^2u, Au)}] = J(u)(1 - \frac{(Au, Au)^2}{(A^2u, Au)(Au, u)}).$$

Le résultat dans ce cas s'appuie alors sur le lemme de Kantorovitch:

Lemme 6.3 *On a l'inégalité, pour A matrice symétrique définie positive:*

$$\forall y \in \mathbb{R}^m \backslash 0, \quad \frac{(y, y)^2}{(Ay, y)(A^{-1}y, y)} \geq \frac{4\lambda_{min}\lambda_{max}}{(\lambda_{min} + \lambda_{max})^2}.$$

On admet pour l'instant ce résultat.

On a alors, dans notre suite, la relation

$$J(u_{n+1}) = J(u_n)(1 - \frac{(Au_n, Au_n)^2}{(A^2u_n, Au_n)(Au_n, u_n)}).$$

Dans cette égalité, on prend $y_n = Au_n$ et on utilise le lemme de Kantorovitch. Alors on trouve

$$J(u_{n+1}) \leq J(u_n)(1 - \frac{4\lambda_{min}\lambda_{max}}{(\lambda_{max} + \lambda_{min})^2}) = (\frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}})^2.$$

Comme $||u_n||_A = \sqrt{2J(u_n)}$, on trouve l'inégalité

$$||u_{n+1} - 0||_A \leq \beta ||u_n - 0||_A$$

d'où la convergence géométrique de la suite u_n vers 0.

Nous passons à l'étude dans le cas général. Pour ce faire, on utilise la formule de Taylor avec reste intégral pour J et pour J'. Pour simplifier les notations, on effectue une translation sur l'inconnue u pour se ramener au minimum $u = 0$ et on change $J(u)$ en $J(u) - l$ où l est le minimum de J.

Les formules de Taylor s'écrivent

$$J(u) = \int_0^1 (1-\theta)(J''(0+\theta u)u, u)d\theta = \frac{1}{2}(J''(0)u, u) + ([\int_0^1 (1-\theta)(J''(\theta u) - J''(0))]u, u).$$

$$J'(u) = J''(0)u + (\int_0^1 J''(\theta u)d\theta - J''(0))u$$

que l'on écrira pour simplifier $J(u) = \frac{1}{2}(J''(0)u, u) + (Q(u)u, u)$ et $J'(u) = J''(0)u + R(u)u$, où Q et R, par la continuité de la dérivée seconde au sens de Fréchet, sont égales à $o(1)$ (c'est à dire tendent vers 0 lorsque u tend vers 0).

On sait déjà que l'algorithme du gradient converge, donc il existe n_0 tel que $||u_n|| \leq \delta_0$ pour $n \geq n_0$. On cherche donc, pour u donné l'unique solution de $(J'(u - \mu J'(u)), J'(u)) = 0$. On note, comme précédemment, $\phi(\mu) = J(u - \mu J'(u))$, $\phi'(\mu) = -(J'(u - \mu J'(u)), J'(u))$, $\phi''(\mu) = (J''(u - \mu J'(u))J'(u), J'(u))$.

On vérifie que

$$\begin{array}{rl} -\phi'(\mu) & = (J''(0)(u - \mu J'(u)) + R(u - \mu J'(u))(u - \mu J'(u)), J''(0)u + R(u)u) \\ & = (J''(0)u, J''(0)u) - \mu(J''(0)J'(u), J''(0)u) \\ & +R(u - \mu J'(u))(u - \mu J'(u)), J''(0)u + R(u)u) \\ & = (J''(0)u, J''(0)u) - \mu(J''(0)^2 u, J''(0)u) - \mu(J''(0)R(u)u, J''(0)u) \\ & +R(u - \mu J'(u))(u - \mu J'(u)), J''(0)u + R(u)u) \end{array} .$$

Ainsi si on étudie, pour u tendant vers 0, la solution de $\phi'(\mu) = 0$, on trouve que μ est proche de $\mu_0(u) = \frac{(J''(0)u, J''(0)u)}{(J''(0)^2 u, J''(0)u)}$, qui est homogène de degré 0 en u, non singulier car la matrice $J"(0)$ est symétrique définie positive. On écrit alors $\mu = \mu_0 + \beta$. On trouve

$$\begin{array}{rl} -\phi'(\mu) & = -\beta(J''(0)^2 u, J''(0)u) - (\mu_0 + \beta)(J''(0)R(u)u, J''(0)u) \\ & +R(u - (\mu_0 + \beta)J'(u))(u - (\mu_0 + \beta)J'(u)), J''(0)u + R(u)u). \end{array}$$

La relation $\phi'(\mu) = 0$ s'écrit alors aussi sous la forme

$$\beta + (\mu_0 + \beta)\frac{(J''(0)R(u)u, J''(0)u)}{(J''(0)^2 u, J''(0)u)} - \frac{R(u - (\mu_0 + \beta)J'(u))(u - (\mu_0 + \beta)J'(u)), J''(0)u + R(u)u)}{(J''(0)^2 u, J''(0)u)} = 0.$$

On vérifie alors que, par le théorème des fonctions implicites, il existe une fonction $\beta(u)$ telle que $\beta(u) = o(1)$ c'est-à-dire tend vers 0 avec $||u||$. Cette valeur de $\beta(u)$ détermine l'unique pas optimal.

On calcule alors

$$J(u-(\mu_0+\beta(u))J'(u)).$$

On s'intéresse au point de base. Il reste

$$\phi(u)=u-(\mu_0+\beta(u))J'(u)=u-\mu_0J"(0)u-\beta(u)J"(0)u-\mu_0R(u)u$$

et ce terme peut s'écrire

$$\phi(u)=u-\mu_0J"(0)u+S(u)u$$

où $S(u)=\beta(u)J"(0)+\mu_0R(u)$, et tend vers 0 dans l'espace des matrices comme $||u||$.

On a alors $J(\phi(u))=\frac{1}{2}(J"(0)(u-\mu_0J"(0)u+S(u)u),u-\mu_0J"(0)u+S(u)u)+(Q(\phi(u))\phi(u),\phi(u))$. On remarque alors que, comme $\phi(u)=u-\mu_0J"(0)u+S(u)u$, pour $||u||$ assez petit on trouve que $||\phi(u)||\leq C||u||$. Ainsi on trouve

$$J(\phi(u))=\frac{1}{2}(J"(0)(u-\mu_0J"(0)u),u-\mu_0J"(0)u)+\epsilon(u)||u||^2,$$

où le terme $\epsilon(u)$ tend vers 0 avec $||u||$.

On reconnait le calcul dans le cas de la forme quadratique $\frac{1}{2}(Au,u)$, ce qui donne tout de suite

$$J(\phi(u))=\frac{1}{2}(J"(0)u,u)(1-\frac{(J"(0)u,J"(0)u)^2}{(J"(0)u,u)((J"(0))^2u,J"(0)u)})+\epsilon(u)||u||^2.$$

Enfin, on reconnait que $J(u)=\frac{1}{2}(J"(0)u,u)(1+\eta(u))$ avec $\eta(u)$ tend vers 0 comme $||u||$ puisque $J"(0)$ est symétrique définie positive donc $(J"(0)u,u)\geq\lambda_{min}||u||^2$. Ainsi il vient

$$\begin{array}{rl} J(\phi(u)) & =\frac{J(u)}{1+\eta(u)}(1-\frac{(J"(0)u,J"(0)u)^2}{(J"(0)u,u)((J"(0))^2u,J"(0)u)})+\epsilon(u)||u||^2 \\ & =J(u)(1-\frac{(J"(0)u,J"(0)u)^2}{(J"(0)u,u)((J"(0))^2u,J"(0)u)})+\epsilon(u)||u||^2 \\ & -\frac{\eta(u)}{1+\eta(u)}(1-\frac{(J"(0)u,J"(0)u)^2}{(J"(0)u,u)((J"(0))^2u,J"(0)u)})J(u). \end{array}$$

Utilisant alors la plus petite valeur propre de $J"(0)$, on constate qu'il existe une fonction $g(u)$, tendant vers 0 si $||u||\to 0$, telle que

$$J(\phi(u))=J(u)(1-\frac{(J"(0)u,J"(0)u)^2}{(J"(0)u,u)((J"(0))^2u,J"(0)u)}+g(u)).$$

On se donne $\beta>\frac{\lambda_{max}-\lambda_{min}}{\lambda_{max}+\lambda_{min}}$. On remarque que $\beta^2+\frac{4\lambda_{max}\lambda_{min}}{(\lambda_{max}+\lambda_{min})^2}>1$. Alors, comme la suite u_n converge vers le minimum de la fonctionnelle 0, il existe n_0 tel que pour $n\geq n_0$ on ait

$$1+g(u)\leq\beta^2+\frac{4\lambda_{max}\lambda_{min}}{(\lambda_{max}+\lambda_{min})^2}.$$

On en déduit, par application du lemme de Kantorovitch

$$\begin{array}{rl} 1+g(u)-\frac{(J"(0)u,J"(0)u)^2}{(J"(0)u,u)((J"(0))^2u,J"(0)u)} & \leq\beta^2+\frac{4\lambda_{max}\lambda_{min}}{(\lambda_{max}+\lambda_{min})^2}-\frac{(J"(0)u,J"(0)u)^2}{(J"(0)u,u)((J"(0))^2u,J"(0)u)} \\ & \leq\beta^2. \end{array}$$

On a donc, pour $n \geq n_0$

$$J(u_{n+1}) \leq \beta^2 J(u_n)$$

ce qui donne

$$J(u_{n+n_0}) \leq \beta^{2n} J(u_{n_0}).$$

Il suffit de rappeler la relation que l'on a obtenue précédemment

$$||u_n - u|| \leq \frac{1}{\alpha}(\frac{2}{\alpha})^{\frac{1}{2}} C\sqrt{J(u_n) - J(u_{n+1})}.$$

On utilise $\alpha = \lambda_{min}$ et $C = \lambda_{max}$, et $J(u_n) - J(u_{n+1}) \leq \beta^2 J(u_n)$ pour obtenir

$$||u_{n+n_0} - u|| \leq \frac{\lambda_{max}}{\lambda_{min}^{\frac{3}{2}}} \beta^{n+1} \sqrt{2J(u_{n_0})}.$$

On a donc démontré une convergence géométrique de la suite u^n vers u, ayant un taux de convergence β arbitraire, strictement supérieur à $\frac{\gamma-1}{\gamma+1}$. Ce taux de convergence est moins bon au fur et à mesure que le conditionnement de la matrice γ tend vers $+\infty$. c'est par exemple ce qui se passe dans un espace de Hilbert lorsqu'on l'approxime par des espaces de dimension finie de plus en plus grand et que la matrice admet des valeurs propres formant une suite tendant vers $+\infty$. Le point ii) du théorème est démontré.

6.4.5 Démonstration du lemme de Kantorovich

On se place tout de suite dans le problème de maximisation sans contrainte de

$$\frac{(A^{-1}y, y)(Ay, y)}{(y, y)^2}.$$

Il est équivalent au problème de maximisation avec contrainte sur la fonctionnelle $(A^{-1}y, y)(Ay, y)$ sur $|y|$ de norme 1, puisque la fonctionnelle du lemme de Kantorovich est homogène d'ordre 0.

On doit donc calculer sur les vecteurs de norme 1

$$\sup(\sum \lambda_p y_p^2)(\sum \lambda_p^{-1} y_p^2).$$

On suppose pour simplifier que toutes les valeurs propres sont distinctes, $\lambda_1 < \lambda_2 < ... < \lambda_m$.

On voit que l'égalité du multiplicateur de Lagrange s'écrit

$$y_j[\lambda_j^{-1}(\sum \lambda_p y_p^2) + \lambda_j(\sum \lambda_p^{-1} y_p^2) + \mu] = 0 \forall j.$$

On remarque d'abord que l'égalité $x^{-1}a + xb = -\mu$ a au plus deux solutions x quand a et b sont non nuls. Donc il existe au plus deux valeurs distinctes de j telles que $y_j \neq 0$ (en notant $a = \sum \lambda_p y_p^2$ et $b = \sum \lambda_p^{-1} y_p^2$).

Dans le cas où $y = (\delta_{ii_0})$, on voit que la fonctionnelle vaut 1. On comparera cette valeur à celle obtenue dans le cas où il y a deux valeurs possibles pour i, pour lequel on a à étudier

$$(\lambda_p y_p^2 + \lambda_q y_q^2)(\lambda_p^{-1} y_p^2 + \lambda_q^{-1} y_q^2) = y_p^4 + y_q^4 + (\frac{\lambda_q}{\lambda_p} + \frac{\lambda_p}{\lambda_q}) y_p^2 y_q^2.$$

C'est une fonctionnelle concave, donc en prenant $x = y_p^2$, on se ramène à $f(x) = x^2 + (1-x)^2 + (\frac{\lambda_q}{\lambda_p} + \frac{\lambda_p}{\lambda_q})x(1-x)$, concave, qui est maximum pour $x = 0.5$. La valeur du maximum est alors $\frac{1}{2} + \frac{1}{4}(\frac{\lambda_q}{\lambda_p} + \frac{\lambda_p}{\lambda_q})$ et comme la fonction $\frac{1}{2} + \frac{1}{4}(t + \frac{1}{t})$ est strictement croissante pour $t \geq 1$, sa plus grande valeur est obtenue, dans l'hypothèse $\lambda_p > \lambda_q$, pour $t = max\frac{\lambda_p}{\lambda q} = \frac{\lambda_{max}}{\lambda_{min}}$.

On remarque alors que cette valeur est plus grande que la valeur en $t = 1$, qui est exactement 1, lorsque $\gamma \neq 0$.

Les deux seuls cas possibles sont alors

• un seul des y_i est non nul, auquel cas on trouve 1 pour la valeur de la fonctionnelle

• deux valeurs de y_i sont non nulles, et on trouve le résultat précédent. On remarque alors que la valeur obtenue dans le paragraphe précédent est maximum si $p = 1$ et $q = n$, soit $y_j = 0$ pour $j \neq 0$ et $j \neq n$, et $y_1 = \pm\frac{1}{\sqrt{2}}$, $y_n = \pm\frac{1}{\sqrt{2}}$.

On vérifie que la valeur de la dérivée seconde de $f(x)$ est

$$f"(x) = 2(2 - \frac{\lambda_q}{\lambda_p} + \frac{\lambda_p}{\lambda_q}) = 2(\frac{\lambda_q}{\lambda_p} - 1)(\frac{\lambda_p}{\lambda_q} - 1)$$

et comme si λ_p/λ_q est plus grand que 1, λ_q/λ_p est plus petit que 1 donc le produit est négatif.

Ce calcul est aussi celui qui prouve que la valeur 1 est plus petite que $\frac{1}{2} + \frac{1}{4}(\frac{\lambda_q}{\lambda_p} + \frac{\lambda_p}{\lambda_q})$.

6.4.6 Algorithme de gradient réduit

On cherche dans cette section à minimiser une fonctionnelle $J(x)$ sous la contrainte $x \in K = \{Ax = b\}$, A matrice $m \times n$ de rang $m < n$.

On suppose pour simplifier l'expression que les inconnues sont ordonnées de sorte que

$$A = (A_0, A_1)$$

où A_0 est une matrice $m \times m$ inversible et A_1 est une matrice $m \times (n - m)$.

Proposition 6.6 *L'algorithme de gradient réduit est une suite (u_n, d_n, μ_n) donnée par*

$$u^0 = (A_0^{-1}(b - A_1 y^0), y^0), d_0 = J'_y(u^0) - (A_0^{-1}A_1)^t J'_x(u^0)$$

et, tant que d_n non nul, on construit la suite par

$$y^1 = y^0 - \mu_0 d_0, u^1 = (A_0^{-1}(b - A_1 y^1), y^1), d_1 = J'_y(u^1) - (A_0^{-1}A_1)^t J'_x(u^1),$$

$$y^{n+1} = y^n - \mu_n d_n, u^{n+1} = (A_0^{-1}(b - A_1 y^{n+1}), y^{n+1}), d_{n+1} = J'_y(u^{n+1}) - (A_0^{-1}A_1)^t J'_x(u^{n+1}).$$

Cet algorithme de gradient réduit est un algorithme de descente pour le problème avec contrainte. Si le pas est choisi convenablement, il converge. Dans le cas où la fonctionnelle est $\alpha-$convexe et Lipschitzienne sur tout borné, il converge (pas optimal ou pas fixe).

On vérifie tout d'abord que $\mathbb{R}^n = \{(x,y), x \in \mathbb{R}^m, y \in \mathbb{R}^{n-m}\}$, et que $A(x,y) = A_0x + A_1y$. On en déduit que $(x,y) \in K \Leftrightarrow A_0x = b - A_1y$, soit $x = A_0^{-1}(b - A_1y)$.

On utilise la procédure décrite dans la proposition 6.1. On en déduit que

$$J(u) = J(A_0^{-1}(b - A_1y), y) = J_r(y).$$

Pour calculer la dérivée, on emploie la différentielle de Gâteaux. On trouve alors, pour $w \in \mathbb{R}^{n-m}$

$$\begin{aligned} J_r(y + \epsilon w) - J_r(y) &= J(A_0^{-1}(b - A_1(y + \epsilon w)), y + \epsilon w) - J(A_0^{-1}(b - A_1y), y) \\ &= J(A_0^{-1}(b - A_1y) - \epsilon A_0^{-1}A_1w, y + \epsilon w) - J(A_0^{-1}(b - A_1y), y) \\ &= (J'(A_0^{-1}(b - A_1y), y), (-A_0^{-1}A_1w, w)) + o(\epsilon) \end{aligned}$$

Si on écrit la dérivée J' en (J'_x, J'_y), on trouve que

$$(J'_r(y), w) = (J'_x(A_0^{-1}(b - A_1y), y), -A_0^{-1}A_1w) + (J'_y(A_0^{-1}(b - A_1y), y), w)$$

Utilisant la transposée, il vient

$$(J'_r(y), w) = (J'_y - (A_0^{-1}A_1)^t J'_x, w).$$

On en déduit la relation

$$J'_r = (J'_y - (A_0^{-1}A_1)^t J'_x.$$

L'algorithme de gradient usuel construit une suite (y^n, d_n) caracérisée par

$$u^n = (A_0^{-1}(b - A_1y^n), y^n), d_n = J'_y(u^n) - (A_0^{-1}A_1)^t J'_x(u^n).$$

On se place dans le cas où $d_n \neq 0$ (car sinon on aurait atteint le point de minimum). Dans ce cas, on introduit

$$D_x^n = -A_0^{-1}A_1d_n.$$

On a, par définition, $A_0D_x^n + A_1d_n = 0$. Soit $J'(u^n) = (d_x^n, d_y^n)$. Le vecteur $D^n = (D_x^n, d_n)$ vérifie

$$(D^n, J'(u^n)) = (-A_0^{-1}A_1d_n, d_x^n) + (d_n, d_y^n) = (d_n, d_y^n - (A_0^{-1}A_1)^t d_x^n) = (d_n, d_n) > 0$$

donc la direction $-D_n$ est à la fois une direction admissible (continue) et une direction de descente pour la fonctionnelle J. C'est donc une direction de descente pour le problème avec contrainte.

D'autre part, si on a $J'_r(y^n) = 0$, alors on a $d_y^n = (A_0^{-1}A_1)^t d_x^n$, ce qui s'écrit

$$\begin{cases} d_y^n = A_1^t((A_0^{-1})^t d_x^n) \\ d_x^n = A_0^t((A_0^{-1})^t d_x^n) \end{cases}$$

dont on déduit le multiplicateur de Lagrange, égal à $-(A_0^{-1})^t d_x^n$, car on a

$$J'(u^n) + \lambda A^t = 0.$$

L'algorithme ainsi construit est un algorithe de gradient pour J_r. Ainsi, pour la suite y_n, d_n, il suffit de choisir le pas convenablement pour se placer dans la catégorie des algorithmes de gradient convergents.

En particulier, si la fonctionnelle est $\alpha-$convexe Lipschitz alors J_r est aussi $\alpha-$convexe Lipschitz puisque les contraintes forment un espace convexe. La proposition est démontrée.

6.5 Algorithmes de gradient conjugué

Dans cette section, nous construisons un des algorithmes les plus utilisés: le gradient conjugué.

6.5.1 Exemple en dimension 2

Nous commençons par un exemple en dimension 2, qui prouve que même si localement la direction de gradient est la meilleure direction, ce n'est pas la meilleure globalement.

En effet, on considère $f(x,y) = a^2x^2 + b^2y^2$. Les isovaleurs de f sont des ellipses et le minimum est trivialement 0.

Lorsqu'on utilise l'algorithme du gradient à pas optimal, on sait que la suite vérifie, pour tout n:

$$(f'(x^{n+1}, y^{n+1}), f'(x^n, y^n)) = 0.$$

Comme on est en dimension 2, cela veut dire qu'il existe λ_n tel que

$$f'(x^{n+1}, y^{n+1}) = \lambda_n (f'(x^n, y^n))^\perp$$

On en déduit, utilisant

$$f'(x^{n+2}, y^{n+2}) = \lambda_{n+1} (f'(x^{n+1}, y^{n+1}))^\perp$$

$$f'(x^{n+2}, y^{n+2}) = -\lambda_n \lambda_{n+1} f'(x^n, y^n)$$

Dans le cas où $a \neq b$, la suite est donc infinie et converge par itérations successives vers le minimum. Si $a = b$, bien sûr une direction de gradient pointe vers le centre du cercle et on converge en une itération.

Mais il est clair que $(x^0, y^0) - (x^0, y^0) = (0,0)$, donc la direction optimale n'est pas celle du gradient mais celle du vecteur pointant vers le centre!

Nous cherchons à exploiter cette idée. En effet, en dimension 2, il n'y a que deux directions possibles, donc même si au premier pas on n'a pas trouvé la bonne direction, on le trouvera au deuxième pas. Pour cela, on considère la direction du gradient comme direction de départ. On trouve que

$$(x_1, y_1) = (x_0, y_0) - \lambda_0 (2a^2 x_0, 2b^2 y_0), \lambda_0 = \frac{a^4 x_0^2 + b^4 y_0^2}{2(a^6 x_0^2 + b^6 y_0^2)}.$$

La bonne direction est (x_1, y_1), car elle conduit tout de suite au minimum. On vérifie que

$$\begin{gathered}((2a^2x_0, 2b^2y_0), A(x_1, y_1)) \\ = ((2a^2x_0, 2b^2y_0), (2a^2x_1, 2b^2y_1)) \\ = ((2a^2x_0, 2b^2y_0), (2a^2x_0, 2b^2y_0)) - \lambda_0((2a^2x_0, 2b^2y_0), (4a^4x_0, 4b^4y_0)) \\ = 4a^4x_0^2 + 4b^4y_0^2 - (8a^6x_0^2 + 8b^6y_0^2)\lambda_0 \\ = 0.\end{gathered}$$

La direction $d_1 = (x_1, y_1)$ vérifie alors $(d_0, Ad_1) = 0$ et grâce à elle, l'algorithme s'arrête immédiatement.

6.5.2 Algorithme de directions conjuguées

Dans le cas de la minimisation d'une fonctionnelle quadratique en dimension finie ou infinie, par exemple $J(x) = \frac{1}{2}(Ax, x) - (b, x)$, où on sait que $Ax = b$ admet une solution x_0, on vérifie que

$$J(x) = \frac{1}{2}(Ax, x) - (Ax_0, x) = \frac{1}{2}(A(x - x_0), x - x_0) - \frac{1}{2}(b, x_0).$$

Ainsi minimiser J revient à minimiser la norme $||x - x_0||_A$.

On se place en dimension finie N. La matrice A est symétrique définie positive, donc elle est diagonalisable dans une base orthogonale notée $(p_1, .., p_N)$. On a alors, comme $(Ap_i, p_j) = 0$ pour $i \neq j$

$$||x - x_0||_A^2 = \sum_i (x_i - x_{0,i})^2 (Ap_i, p_i).$$

On part du point x_1. On cherche le minimum de la fonction sur $\mathbb{R}$ égale à $\lambda \to J(x_1 + \lambda p_1)$. On trouve que la relation donnant le minimum en λ est

$$(A(x_1 + \lambda p_1) - b, p_1) = 0$$

soit

$$\lambda = \lambda_1 = \frac{(b - Ax_1, p_1)}{(Ap_1, p_1)}.$$

On regarde alors le deuxième point $x_2 = x_1 + \lambda p_2$. On trouve que la valeur de λ est $\lambda_2 = \frac{(b - Ax_2, p_2)}{(Ap_2, p_2)}$.

D'autre part, on considère $\phi(\lambda, \mu) = J(x_1 + \lambda p_1 + \mu p_2)$. C'est une fonction de deux variables, qui est minimale pour

$$\partial_\lambda \phi = \partial_\mu \phi = 0.$$

On obtient les relations

$$\begin{cases} (J'(x_1 + \lambda p_1 + \mu p_2), p_1) = 0 \\ (J'(x_1 + \lambda p_1 + \mu p_2), p_2) = 0 \end{cases}$$

soit

$$\begin{cases} (Ax_1 - b + \lambda Ap_1 + \mu Ap_2, p_1) = 0 \\ (Ax_1 - b + \lambda Ap_1 + \mu Ap_2, p_2) = 0 \end{cases}$$

$$\begin{cases} (Ax_1 - b, p_1) + \lambda(Ap_1, p_1) = 0 \\ (Ax_1 - b, p_2 + \mu(Ap_2, p_2) = 0 \end{cases}$$

ce qui conduit à $\lambda = \lambda_1$ et $\mu = \lambda_2$.

On voit donc que le point $x_3 = x_1 + \lambda_1 p_1 + \lambda_2 p_2$ est le point qui réalise le minimum de J sur l'espace affine $x_1 + \text{Vect}(p_1, p_2)$.

On définit alors la suite de récurrence par

$$x_{n+1} = x_n + \lambda_n p_n$$

avec

$$\lambda_n = \frac{(b - Ax_n, p_n)}{(Ap_n, p_n)}$$

Alors x_{n+1} est le point où J est minimum sur $E_n = x_1 + \text{Vect}(p_1, p_2, ..., p_n)$.

Cet algorithme est un algorithme de directions conjuguées. On écrit alors la

Proposition 6.7 *Soit (p_n) une suite dans V Hilbert de directions conjuguées au sens où $(p_i, Ap_j) = (Ap_i, p_j) = 0$ pour $i \neq j$ tel que l'espace vectoriel fermé engendré par la suite des p_j est l'espace de Hilbert tout entier (c'est à dire que tout élément de l'espace de Hilbert est limite d'une suite de combinaisons linéaires finies des p_j).*

La suite définie par

$$\begin{cases} x_{n+1} = x_n + \lambda_n p_n \\ \lambda_n = \frac{(p_n, b - Ax_n)}{(p_n, Ap_n)} \end{cases}$$

vérifie les relations

$$(b_A p_n, p_k) = 0 \text{ pour } k \leq n - 1$$

et x_n converge vers x_0 la solution unique de $Ax = b$.

Pour démontrer cette proposition, on écrit effectivement la norme. On voit alors que

$$x_1 = \sum x_1^i p_i, x_0 = \sum X_i p_i, b = \sum X_i A p_i$$

$$\lambda_1 = \frac{(p_1, b - Ax_1)}{(p_1, Ap_1)} = -\frac{(p_1, \sum (x_1^i - X_i) Ap_i)}{(p_1, Ap_1)} = -(x_1^1 - X_1)$$

donc $x_2 = X_1 p_1 + \sum_{i \geq 2} x_1^i p_i$.

On voit alors que $b - Ax_2 = \sum_{i \geq 2} (X_i - x_1^i) Ap_i$, donc $(b - Ax_2, p_2) = (X_2 - x_1^2)(Ap_2, p_2)$ donc $\lambda_2 = X_2 - x_1^2$ et $x_3 = X_1 p_1 + X_2 p_2 + \sum_{i \geq 3} x_1^i p_i$.

On continue le raisonnement pour obtenir

$$x_n = \sum_{1 \leq i \leq n-1} X_i p_i + \sum_{i \geq n} x_1^i p_i.$$

On voit alors que

$$||x_n - x_0||_A^2 = \sum_{i \geq n} (X_i - x_1^i)^2 (Ap_i, p_i)$$

et la suite $||x_n - x_0||_A^2$ est une suite décroissante positive. Elle a donc une limite. Cette limite est 0 car la famille (p_j) est une famille complète. On en déduit que la suite x_n

tend vers la solution du problème. La proposition est démontrée. On remarque aussi que x_n identifie déjà les $n-1$ premiers termes de x_0.

Ce raisonnement n'est réellement applicable que lorsqu'on connait A donc la forme quadratique. Dans le cas général, on va combiner cette méthode avec une méthode de gradient afin de construire une suite par un procédé d'orthogonalisation de Gram-Schmidt.

Application aux polynômes de Hermite On définit les polynômes de Hermite par la relation

$$H_n(x) = (-1)^n \frac{d^n}{dx^n}(e^{-\frac{x^2}{2}})e^{\frac{x^2}{2}}.$$

On vérifie par récurrence que H_n est un polynôme de degré n dont le monôme de plus haut degré est x^n. En effet,

$$H_{n+1}(x) = -\frac{d}{dx}(H_n(x)e^{-\frac{x^2}{2}})e^{\frac{x^2}{2}} = xH_n(x) - H_n'(x).$$

Comme, par hypothèse, H_n est de degré n dont le monôme de plus haut degré est x^n (dans le raisonnement par récurrence), on sait que H_n' est de degré $n-1$ donc $xH_n - H_n'$ est de degré $n+1$ et son terme de plus haut degré est x^{n+1}. D'autre part, $H_1(x) = 1$ donc l'hypothèse de récurrence est vraie pour $n = 1$.

On contrôle que

$$\int_{\mathbf{R}} H_n(x)H_p(x)e^{-\frac{x^2}{2}}dx = \int_{\mathbf{R}} H_n(x)(-1)^p \frac{d^p}{dx^p}(e^{-\frac{x^2}{2}})dx.$$

Sans restreindre la généralit; on peut supposer soit $p = n$ soit $p > n$. Dans le cas $p > n$, en faisant p intégrations par parties, on trouve que

$$\int_{\mathbf{R}} H_n(x)H_p(x)e^{-\frac{x^2}{2}}dx = \int_{\mathbf{R}} \frac{d^p}{dx^p}(H_n(x))e^{-\frac{x^2}{2}}dx = 0$$

car H_n est un polynôme de degré $n < p$.

D'autre part, pour $p = n$ on trouve que

$$\int_{\mathbf{R}} H_n(x)H_n(x)e^{-\frac{x^2}{2}}dx = n! \int_{\mathbf{R}} e^{-\frac{x^2}{2}}dx = n!\sqrt{2\pi}$$

La famille de polynômes H_n est donc une famille orthogonale pour le produit scalaire

$$\int f(x)g(x)e^{-\frac{x^2}{2}}dx$$

et c'est donc une famille conjuguée pour l'application $Af = fe^{-\frac{x^2}{2}}$.

6.5.3 Algorithme du gradient conjugué

Théorème 6.6 *On considère une fonctionnelle quadratique $J(x)$. On construit la suite de directions d_j par*

$$d_0 = -J'(x_0)$$

$$x_{n+1} = x_n + \rho_n d_n, \rho_n = arginf J(x_n + \rho d_n)$$

$$d_{n+1} = -J'(x_{n+1}) + \beta_{n+1} d_n.$$

$$\beta_{n+1} = \frac{|J'(x_{n+1})|^2}{|J'(x_n)|^2}, \rho_n = -\frac{|J'(x_n)|^2}{(Ad_n, J'(x_n))}.$$

La famille (d_j) définit une famille de directions conjuguées associées à A telle que $J'(x) - J'(y) = A(x-y)$.

La famille des directions $J'(x_p)$ est une famille orthogonale pour le produit scalaire usuel.

L'espace vectoriel engendré par la famille $(J'(x_p))$, $0 \leq p \leq j$ est égal à l'espace vectoriel engendré par la famille (d_p), $0 \leq p \leq j$.

En dimension finie N la famille de directions conjuguées est complète et l'algorithme donné dans la partie précédente converge **en au plus N itérations**.

Pour faire la démonstration correctement, il s'agit de construire les éléments de la suite successivement. On suppose que l'on minimise la fonctionelle quadratique $\frac{1}{2}(Ax, x) - (b, x)$. On utilisera la relation

$$J'(x) - J'(y) = A(x-y). \tag{6.5.2}$$

Etape 1. On commence avec un point x_0 et on introduit

$$\begin{cases} x_1 = x_0 + \rho_0 d_0 \\ d_0 = -J'(x_0) \end{cases}$$

La condition d'optimalité s'écrit

$$(J'(x_1), d_0) = 0.$$

On en déduit

$$(J'(x_1) - J'(x_0), d_0) + (J'(x_0), d_0) = 0.$$

$$(A(x_1 - x_0), d_0) = |J'(x_0)|^2$$

soit $\rho_0(Ad_0, d_0) = |J'(x_0)|^2$

$$\rho_0 = \frac{|J'(x_0)|^2}{(Ad_0, d_0)} = -\frac{|J'(x_0)|^2}{(Ad_0, J'(x_0))}.$$

On note alors que $(J'(x_1), J'(x_0)) = 0$.

Etape 2. On construit une direction conjuguée. Alors d_1 vérifie $(Ad_1, d_0) = 0$. On impose de plus que cette direction conjuguée soit une direction de descente reliée au gradient, par

$$d_1 = -J'(x_1) + \beta_1 d_0.$$

Ceci implique que l'on veuille trouver une direction conjuguée dans l'espace vectoriel engendré par les gradients successifs $(J'(x_0), J'(x_1))$. On a simplement imposé

que cette direction conjuguée soit telle que $d_1 + J'(x_1) = 0$. On verra plus loin que cela ne restreint pas la généralité de faire ainsi.

Comme c'est une direction conjuguée, on trouve

$$(d_1, Ad_0) = 0$$

soit

$$(J'(x_1), Ad_0) = \beta_1(Ad_0, d_0).$$

On multiplie les deux membres de l'égalité par ρ_0, et on remarque que $\rho_0 d_0 = x_1 - x_0$, ce qui donne

$$(J'(x_1), A(x_1 - x_0)) = \beta_1(A(x_1 - x_0), -J'(x_0))$$

ou encore en utilisant la relation (6.5.2)

$$(J'(x_1), J'(x_1) - J'(x_0)) = \beta_1(J'(x_1) - J'(x_0), -J'(x_0)).$$

On utilise l'orthogonalité de $J'(x_0)$ et de $J'(x_1)$ pour obtenir

$$\beta_1 = \frac{|J'(x_1)|^2}{|J'(x_0)|^2}.$$

La condition d'optimalité pour ρ_1 s'écrit $(J'(x_2), d_1) = 0$. Comme de plus

$$(J'(x_2), d_0) = (J'(x_2) - J'(x_1), d_0) + (J'(x_1), d_0) = \rho_1(Ad_1, d_0) + 0 = 0$$

on en déduit que $J'(x_2)$ est orthogonal à d_0 et à d_1, donc est orthogonal à $J'(x_0)$ et à $J'(x_1)$.

La condition d'optimalité donne alors la valeur de ρ_1 par

$$(J'(x_2) - J'(x_1), d_1) + (J'(x_1), d_1) = 0$$

$$\rho_1(Ad_1, d_1) = |J'(x_1)|^2$$

puisque $d_1 = -J'(x_1) + \beta_1 d_0$, et que $(J'(x_1), d_0) = -(J'(x_1), J'(x_0)) = 0$. D'autre part, $d_1 = -J'(x_1) + \beta_1 d_0$ et $(Ad_1, d_0) = 0$ donc $(Ad_1, d_1) = -(Ad_1, J'(x_1))$. Il vient

$$\rho_1 = -\frac{|J'(x_1)|^2}{(Ad_1, J'(x_1))} = \frac{|J'(x_1)|^2}{(Ad_1, d_1)}.$$

Pour bien comprendre la procédure, nous étudions l'étape 2.

On construit donc une direction conjuguée d_2. Elle est conjuguée donc

$$(Ad_2, d_1) = (Ad_2, d_0) = 0.$$

On suppose que cette direction conjuguée appartient à l'espace vectoriel engendré par la famille $(J'(x_0), J'(x_1), J'(x_2))$. Comme l'espace vectoriel engendré par $(J'(x_0), J'(x_1))$ est l'espace vectoriel engendré par (d_0, d_1), on écrit $d_2 = -J'(x_2) + \beta_2^0 d_0 + \beta_2^1 d_1$.

Pour justifier cette forme, prenons une direction quelconque de $Vect(J'(x_0), J'(x_1), J'(x_2))$. Comme l'espace vectoriel engendré par $J'(x_0), J'(x_1)$ est le même que l'espace vectoriel engendré par d_0, d_1, une direction quelconque est donc sous la forme

$$\tilde{d_2} = \alpha J'(x_2) + \beta d_0 + \gamma d_1.$$

Cette direction est une direction de descente, donc nécessairement $(\tilde{d_2}, J'(x_2)) \leq 0$. Comme $J'(x_2)$ est orthogonal à d_0 et à d_1, on en déduit que $\alpha \leq 0$. On veut éviter le cas où $\alpha = 0$ car on est dans l'espace vectoriel engendré par d_0 et d_1 qui sont deux directions de descente que l'on a utilisé, ainsi $\alpha < 0$.

D'autre part, si on considère un point dans cette direction de descente, il s'écrit

$$x_2 + r\tilde{d_2} = x_2 + (-\alpha r)(-J'(x_2) + \frac{-\beta}{\alpha} d_0 + \frac{-\gamma}{\alpha} d_1).$$

On a retrouvé le pas $\rho = -\alpha r \geq 0$ et l'écriture de la direction de descente d_2.

Pour identifier les coefficients, on n'a besoin que des conditions de conjugaison. On trouve alors

$$\begin{array}{c}(-J'(x_2) + \beta_2^0 d_0 + \beta_2^1 d_1, Ad_0) = 0 \\ (-J'(x_2) + \beta_2^0 d_0 + \beta_2^1 d_1, Ad_1) = 0\end{array}$$

En utilisant le fait que les directions d_0 et d_1 sont conjuguées, on trouve

$$\beta_2^0(d_0, Ad_0) = (J'(x_2), Ad_0), \beta_2^1(d_1, Ad_1) = (J'(x_2), Ad_1).$$

On multiplie respectivement chacune de ces égalités par ρ_0 et par ρ_1 et on utilise $\rho_1 d_1 = x_2 - x_1$, $\rho_0 d_0 = x_1 - x_0$. Alors il vient

$$\beta_2^0(d_0, A\rho_0 d_0) = (J'(x_2), A(x_1 - x_0)), \beta_2^1(d_1, A\rho_1 d_1) = (J'(x_2), A(x_2 - x_1))$$

On utilise la remarque (6.5.2) pour obtenir

$$\beta_2^0(d_0, A\rho_0 d_0) = (J'(x_2), J'(x_1) - J'(x_0)), \quad \beta_2^1(d_1, J'(x_1) - J'(x_0)) = (J'(x_2), J'(x_2) - J'(x_1))$$

et on utilise l'orthogonalité des vecteurs dérivées. Ainsi il reste $\beta_2^0 = 0$ et

$$\beta_2^1(d_1, J'(x_1) - J'(x_0)) = (J'(x_2), J'(x_2))$$

Comme $d_1 = -J'(x_1) + \beta_1 d_0 = -J'(x_1) - \beta_1 J'(x_0)$, il vient

$$-\beta_2^1 |J'(x_1)|^2 = |J'(x_2)|^2.$$

D'autre part la condition d'optimalité est $(J'(x_3), d_2) = 0$, $x_3 = x_2 + \rho_2 d_2$. On sait d'autre part que

$$\begin{array}{rl}(J'(x_3), d_1) = & (J'(x_3) - J'(x_2), d_1) + (J'(x_2), d_1) \\ & = (J'(x_3) - J'(x_2), d_1) \text{ optimalité pour } x_2 \\ & = (A(x_3 - x_2), d_1) = \rho_2(Ad_2, d_1) = 0 \text{ conjuguées}\end{array}$$

$$(J'(x_3), d_0) = (J'(x_2), d_0) + \rho_2(Ad_2, d_0) = (J'(x_2), d_0) = -(J'(x_2), J'(x_0)) = 0.$$

On sait donc que $J'(x_3)$ est orthogonal à l'espace vectoriel engendré par d_0, d_1, d_2 donc est orthogonal à $J'(x_0)$, $J'(x_1)$, $J'(x_2)$.

Finalement le coefficient ρ_2 est donné par

$$\rho_2(Ad_2, d_2) + (J'(x_2), d_2) = 0$$

soit, utilisant $d_2 = -J'(x_2) + \beta_2^1 d_1$ et l'orthogonalité de d_1 et de $J'(x_2)$

$$\rho_2(Ad_2, d_2) = |J'(x_2)|^2$$

et on en déduit, utilisant le fait que les directions sont conjuguées

$$\rho_2 = -\frac{|J'(x_2)|^2}{(J'(x_2), Ad_2)} = \frac{|J'(x_2)|^2}{d_2, Ad_2)}.$$

Raisonnement par récurrence On suppose donc que l'on a construit une suite (x_p, ρ_p, d_p), $p \leq n$, et x_{n+1} ayant les propriétés suivantes:

- la suite (d_p) est une suite de directions conjuguées
- $d_{p+1} = -J'(x_{p+1}) + \beta_{p+1} d_p$ pour $p \leq n-1$ avec

$$\beta_{p+1} = \frac{|J'(x_{p+1})|^2}{|J'(x_p)|^2}.$$

- les vecteurs $(J'(x_p))$ forment une famille orthogonale pour le produit scalaire usuel pour $0 \leq p \leq n+1$
- $x_{p+1} = x_p + \rho_p d_p$ pour $p \leq n$, les ρ_p étant donnés par la relation

$$\rho_p = -\frac{|J'(x_p)|^2}{(J'(x_p), Ad_p)}.$$

On construit x_{n+2}, d_{n+1} et ρ_{n+1} suivant les conditions suivantes. On veut que l'espace vectoriel engendré par $(J'(x_0), .., J'(x_{p+1})$ soit aussi l'espace vectoriel engendré par les directions $(d_0, .., d_{p+1})$. On impose de plus que $d_{p+1} = -J'(x_{p+1}) + l_p$, où l_p est dans l'espace vectoriel engendré par $(d_0, .., d_p)$ qui est égal, par l'hypothèse de récurrence, à l'espace vectoriel engendré par $(J'(x_0), .., J'(x_p))$. On écrit donc

On sait déjà que

$$d_{n+1} = -J'(x_{n+1}) + \sum_{j=0}^{n} \beta_{n+1}^j d_j$$

Les directions sont conjuguées, donc $(d_{n+1}, Ad_p) = 0 \forall p$.

On en déduit donc que

$$\sum_{j=0}^{n} \beta_{n+1}^j (d_j, Ad_p) = (J'(x_{n+1}), Ad_p).$$

Utilisant le fait que la famille de directions d_j est conjuguée, il vient

$$\beta_{n+1}^p (d_p, Ad_p) = (J'(x_{n+1}), Ad_p).$$

On multiplie les deux membres de l'égalité par ρ_p et on utilise $\rho_p Ad_p = J'(x_{p+1}) - J'(x_p)$. Ensuite, comme la famille $(J'(x_k))$, $0 \leq k \leq n+1$ est une famille orthogonale, on en déduit que $J'(x_{n+1})$ est orthogonal à tous les $J'(x_{p+1})$ pour $p+1 \leq n$ et à tous les $J'(x_p)$ pour $p \leq n$. On en déduit que $\beta_{n+1}^p = 0$ pour $p \neq n$. Il reste alors seulement un terme

$$\beta_{n+1}^n(d_n, J'(x_{n+1} - J'(x_n)) = (J'(x_{n+1}), J'(x_{n+1}) - J'(x_n)) = |J'(x_{n+1})|^2$$

Comme d'autre part $d_n = -J'(x_n) + \beta_{n-1}d_{n-1}$, utilisant le fait que d_{n-1} est dans l'espace vectoriel engendré par $J'(x_0), .., J'(x_{n-1})$ donc est orthogonal à $J'(x_n)$ et à $J'(x_{n+1})$, il reste

$$\beta_{n+1}^n(-J'(x_n), J'(x_{n+1}) - J'(x_n)) = |J'(x_{n+1})|^2$$

soit

$$\beta_n = \beta_{n+1}^n = \frac{|J'(x_{n+1})|^2}{|J'(x_n)|^2}.$$

On a donc construit une direction $d_{n+1} = -J'(x_{n+1}) + \beta_n d_n$ telle que les directions (d_p), $0 \leq p \leq n+1$ soient conjuguées.

La condition d'optimalité pour x_{n+2} s'écrit

$$(J'(x_{n+2}), d_{n+1}) = 0$$

On sait en outre que

$$(J'(x_{n+2}), d_k) = (J'(x_{n+2}) - J'(x_{k+1}), d_k) + (J'(x_{k+1}), d_k).$$

On utilise la condition d'optimalité pour x_{k+1} pour annuler $(J'(x_{k+1}), d_k)$. D'autre part, on utilise la remarque (6.5.2) pour obtenir, A étant symétrique

$$(J'(x_{n+2}), d_k) = (x_{n+2} - x_{k+1}, Ad_k).$$

Comme $x_{n+2} - x_{k+1} = \rho_{n+1}d_{n+1} + ... + \rho_{k+1}d_{k+1}$ et que la famille de directions est conjuguée, on trouve 0 pour $k \leq n$. Le vecteur $J'(x_{n+2})$ est orthogonal à toutes les directions d_k pour $0 \leq k \leq n+1$. Comme l'espace vectoriel engendré par les $J'(x_p)$, $0 \leq p \leq n+1$ est égal, **dans le cas où le minimum n'a pas été atteint** à celui engendré par les d_p, on vérifie que $J'(x_{n+2})$ est orthogonal à tous les $J'(x_p)$ pour $0 \leq p \leq n+1$.

Enfin, écrivons la condition d'optimalité. On a donc,

$$(A(x_{n+2} - x_{n+1}), d_{n+1}) + (J'(x_{n+1}), d_{n+1}) = 0$$

soit utilisant $d_{n+1} = -J'(x_{n+1}) + \beta_n d_n$, $\rho_{n+1}(Ad_{n+1}, d_{n+1}) = |J'(x_{n+1})|^2$.

On en tire la relation

$$\rho_{n+1} = -\frac{|J'(x_{n+1})|^2}{(Ad_{n+1}, J'(x_{n+1}))}.$$

Toutes les hypothèses du raisonnement par récurrence ont été vérifiées, ainsi l'algorithme continue jusquà obtenir $J'(x_N) = 0$. En dimension finie d, on aura nécessairement cette condition puisque la famille $(J'(x_0), .., J'(x_{d-1}))$ est une famille orthogonale. Si c'est une famille libre, c'est une base et $J'(x_d)$ orthogonal à tous les éléments implique que $J'(x_d) = 0$. Si c'est une famille liée, comme le vecteur $J'(x_{d-1})$ est orthogonal à tous les autres, si il est combinaison linéaire de tous les autres, cette combinaison linéaire est nulle si tous sont non nuls, donc il en existe au moins un qui est nul.

Corollaire 6.1 *Le coefficient de d_p dans la suite de directions conjuguées de l'algorithme de gradient conjugué est la valeur qui maximise le facteur de réduction de l'erreur, erreur définie par $E(x) = (r(x), A^{-1}(r(x)))$ où $r(x) = -J'(x)$.*

On remarque que dans le cas de la forme quadratique $J(x) = \frac{1}{2}(Ax, x)$, on trouve $J'(x) = Ax$ donc $E(x) = 2J(x)$. On a alors immédiatement

$$x_{n+1} = x_n + \rho_n d_n, d_n = -J'(x_n) + \beta_{n-1} d_{n-1}.$$

Le terme ρ_n est calculé par $0 = (Ax_n + \rho_n Ad_n, d_n)$, soit $\rho_n = -\frac{(Ax_n, d_n)}{(Ad_n, d_n)}$. Dans ce cas, on applique le résultat donné précédemment et on trouve

$$E(x_{n+1}) = E(x_n)[1 - \frac{(Ax_n, d_n)^2}{(Ad_n, d_n)(x_n, Ax_n)}],$$

On voit alors que $(Ax_n, d_n) = (Ax_n, -Ax_n + \beta_{n-1} d_{n-1}) = -(Ax_n, Ax_n)$ car Ax_n est orthogonal à d_{n-1}. Maximiser le facteur de réduction de l'erreur revient alors à maximiser $\frac{(Ax_n, d_n)^2}{(Ad_n, d_n)(x_n, Ax_n)}$, donc à minimiser (Ad_n, d_n). Comme

$$\begin{aligned}(Ad_n, d_n) &= (-A^2 x_n + \beta_{n-1} Ad_{n-1}, -Ax_n + \beta_{n-1} d_{n-1}) \\ &= (A^2 x_n, Ax_n) - 2\beta_{n-1}(Ad_{n-1}, Ax_n) + \beta_{n-1}^2 (Ad_{n-1}, d_{n-1})\end{aligned}$$

le minimum de cette fonction quadratique est obtenu pour $\beta_{n-1} = \frac{(Ad_{n-1}, Ax_n)}{(Ad_{n-1}, d_{n-1})}$, ce qui correspond à la formule obtenue précédemment en utilisant $\alpha_{n-1} d_{n-1} = x_n - x_{n-1}$. En effet, $\alpha_{n_1} d_{n-1} = x_n - x_{n-1}$ donc $\beta_{n-1} = \frac{(A(x_n - x_{n-1}), Ax_n)}{(A(x_n - x_{n-1}), d_{n-1})}$. En utilisant $d_{n-1} = -Ax_{n-1} + \beta_{n-2} d_{n-2}$ si $n \geq 2$, $d_0 = -Ax_0$, d_{n-2} est orthogonal à Ax_n et à Ax_{n-2} si $n \geq 2$, donc $(d_{n-1}, Ax_n - Ax_{n-1}) = (-Ax_{n-1}, Ax_n - Ax_{n-1}) = ||J'(x_{n-1})||^2 = ||r(x_{n-1})||^2$, et il reste $\beta_{n-1} = \frac{||Ax_n||^2}{||Ax_{n-1}||^2}$. Le Corollaire est démontré.

6.5.4 Un exemple en dimension 3

En dimension 3, on sait que pour une fonctionnelle quadratique l'algorithme du gradient conjugué converge en trois itérations au plus, c'est à dire on construit au mieux d_0, d_1, d_2. Nous donnons dans le cas de cet exemple les cas oú l'algorithme converge en une itération et en deux itérations.

La fonctionnelle étudiée ici est une fonctionnelle dont les lignes de niveau sont des ellipsoides. On prend

$$J(x, y, z) = \frac{1}{2}(a^2 x^2 + b^2 y^2 + c^2 z^2).$$

Le point de départ est le point (x_0, y_0, z_0). Le gradient en ce point est

$$(a^2 x_0, b^2 y_0, c^2 z_0).$$

Les points de la droite de descente sont

$$(x_0(1 - a^2 t), y_0(1 - b^2 t), z_0(1 - c^2 t)).$$

L'slgorithme converge en une itération lorsque le point d'arrivée est le point $(0, 0, 0)$. On trouve donc

$$\begin{cases} x_0(1-a^2t)=0 \\ y_0(1-b^2t)=0 \\ z_0(1-c^2t)=0 \end{cases}$$

et donc, si $x_0 \neq 0$, alors $t=\frac{1}{a^2}$ donc $y_0=z_0=0$, et si c'est y_0 qui est non nul on trouve $x_0=z_0=0$ et si $z_0 \neq 0$ alors $x_0=y_0=0$.

On en déduit que **l'algorithme converge en une itération lorsque le point est sur un des axes principaux de l'ellipsoide**

Dans le cas contraire, on calcule la valeur de la fonctionnelle.

On trouve, notant

$$\phi(t)=J(x_0(1-a^2t),y_0(1-b^2t),z_0(1-c^2t))$$

$$\phi(t)=\frac{1}{2}(x_0^2(1-a^2t)^2a^2+y_0^2(1-b^2t)^2b^2+z_0^2(1-c^2t)^2c^2)$$

qui atteint son minimum en t_0 que l'on ne calculera pas.

Le gradient en ce point est alors

$$J'(x^{(1)})=(a^2x_0(1-a^2t_0),b^2y_0(1-b^2t_0),c^2z_0(1-c^2t_0))$$

On trouve alors que la direction d_1, qui vaut $d_1=-J'(x^{(1)})+\beta_0 d_0$, est de la forme

$$d_1=(\alpha x_0,\beta y_0,\gamma z_0)=(a^2x_0(-1+a^2t_0+\beta_0),b^2y_0(-1+b^2t_0+\beta_0),c^2z_0(-1+c^2t_0+\beta_0))$$

et donc $x^{(2)}=x^{(1)}+\rho d_1$, soit

$$x^{(2)}=$$
$$(a^2x_0[(1-a^2t_0)+\rho(-1+a^2t_0+\beta_0)],b^2y_0[(1-b^2t_0)+\rho(-1+b^2t_0+\beta_0)],c^2z_0[(1-c^2t_0)+\rho(-1+c^2t_0$$

On suppose que l'algorithme a convergé en deux itérations. Alors les coordonnées dans l'expression ci-dessus sont nulles. On élimine le cas où une seulement des valeurs de (x_0,y_0,z_0) est non nulle car c'est le cas précédent. Si $x_0y_0z_0 \neq 0$, on en déduit que les coefficients sont nuls, c'est à dire on obtient un système sur t_0,β_0,ρ. On vérifie que ce système n'a pas de solutions. En effet, on trouve les relations $(1-a^2t_0)(1-\rho)+\rho\beta_0=(1-b^2t_0)(1-\rho)+\beta_0\rho=0$, d'où $(a^2-b^2)t_0(1-\rho)=0$. Le cas t_0 est impossible (il suffit de vérifier que $t_0(a^6x_0^2+b^6y_0^2+c^6z_0^2)=a^4x_0^2+b^4y_0^2+c^4z_0^2$). Il reste donc $\rho=1$, ce qui donne $\beta_0=0$. Comme β_0 est le quotient des normes de $J'(x^{(1)})$ et de $J'(x^{(0)})$, on trouve que c'est impossible. Ainsi, seulement deux valeurs sur les trois sont non nulles.

Dans ce cas, on considère par exemple $z_0=0$. Alors le point de départ est dans le plan $z=0$, ainsi que le vecteur gradient. Le point d'arrivée $x^{(1)}$ est alors dans ce plan, et on s'est ramené au minimum de la fonctionnelle $J(x,y,0)$ qui est atteint en deux itérations, la première direction $d_0=-J'(x^{(0)})$ et la deuxième direction $d_1=-J'(x^{(1)})+\beta_0 d_0$ comme dans le cas de l'ellipse.

On vérifie alors que **l'algorithme du gradient conjugué converge en deux itérations seulement si le point de départ appartient à un des espaces de dimension 2 invariants par la matrice** $J"(0)$**.**

Remarque On considére la forme quadratique associée à la matrice $A = \begin{pmatrix} a^2 & 1 & 0 \\ 1 & b^2 & 0 \\ 0 & 0 & c^2 \end{pmatrix}$.
On voit que les valeurs propres de cette matrice sont c^2 et λ solution de $\lambda^2 - (a^2 + b^2)\lambda + a^2b^2 - 1 = 0$, soit

$$(\lambda - \frac{a^2+b^2}{2})^2 = 1 + (\frac{a^2-b^2}{2})^2$$

Pour pouvoir écrire la matrice comme précédemment, il faut diagonaliser la matrice donc rechercher les vecteurs propres $(e_\pm, f_\pm, 0)$ pour les deux valeurs propres $\lambda_\pm = \frac{a^2+b^2}{2} \pm \sqrt{1 + (\frac{a^2-b^2}{2})^2}$.

L'algorithme du gradient conjugué converge en deux itérations dans les trois cas suivants:

point de départ de la forme $A(e_+, f_+, 0) + B(e_-, f_-, 0) = (x, y, 0)$,
point de départ de la forme $A(e_+, f_+, 0) + C(0, 0, 1)$,
point de départ de la forme $B(e_-, f_-, 0) + C(0, 0, 1)$.

6.6 Algorithme de descente pseudo-conjugué pour une forme non quadratique

On peut construire, en s'inspirant de l'algorithme ci-dessus, des algorithmes de descente déduits de l'algorithme du gradient conjugué. En fait, l'idée consiste à conserver la relation $d_{n+1} = -J'(x_n) + \beta_n d_n$ et $d_0 = -J'(x_0)$ en construisant la suite β_n et la suite de pas ρ_n.

On l'écrit dans la

Définition 6.8 *Les algorithmes de descente suivants sont la généralisation de l'algorithme du gradient conjugué pour une fonctionnelle quelconque:*

- *algorithme de Fletcher-Reeves:*

$$\begin{cases} d_0 = -J'(x_0) \\ x_{n+1} = x_n + \rho_n d_n \\ d_{n+1} = -J'(x_n) + \beta_n d_n \\ \beta_n = \frac{|J'(x_{n+1}|^2}{|J'(x_n)|^2} \end{cases}$$

- *algorithme de Polak-Ribiere*

$$\begin{cases} d_0 = -J'(x_0) \\ x_{n+1} = x_n + \rho_n d_n \\ d_{n+1} = -J'(x_n) + \beta_n d_n \\ \beta_n = \frac{(J'(x_{n+1}, J'(x_{n+1}) - J'(x_n))}{|J'(x_n)|^2} \end{cases}$$

On a le résultat suivant (admis)

Proposition 6.8 *L'algorithme de Fletcher-Reeves avec le choix du pas optimal pour ρ_n est un algorithme de descente.*

L'algorithme de Polak-Ribiere avec ρ_n pas de Wolfe pas trop grand est un algorithme de descente.

Si J est strictement convexe et de classe C^2 alors l'algorithme de Polak-Ribière avec pas optimal converge.

6.7 Méthode de Newton

On se place sur un espace de Hilbert V, et on considère une fonctionnelle J qui admet un gradient $G(u)$ et une matrice hessienne $H(u)$. On suppose que J admet son minimum absolu en u. Il est alors nécessaire que $G(u)$ soit nul.

Rappelons tout d'abord la formule de Taylor au voisinage de v: il existe $\theta \in]0,1[$ tel que

$$(G(u),\phi) = (G(v),\phi) + (H(v+\theta(u-v))(u-v), u-v).$$

La méthode de Newton-Rophson usuelle construit la solution comme limite de la suite u_k, définie par récurrence: on calcule u_{k+1} à partir de u_k en résolvant $G(u_k) + H(u_k)(u_{k+1} - u_k) = 0$. Cette méthode est efficace si la valeur initiale de la suite est proche de la solution cherchée.

Dans cette partie, on se restreint à des fonctionnelles assez régulières:

(H1) la fonctionnelle J est infinie à l'infini

(H2) la fonctionnelle J a un gradient et un hessien réguliers (au moins continus uniformément sur tout compact)

(H3) H est uniformément V coercive sur tout borné K:

$$(H(v),\phi,\phi) \geq \alpha_K ||\phi||^2, \forall \phi \in V, \forall v \in K$$

(H4) H vérifie une condition de Lipschitz sur les bornés:

$$||H(u) - H(v)|| \leq \beta_K ||u - v||, \forall (u,v) \in K^2$$

De plus, ce qui fait la différence avec la méthode de Newton habituelle, c'est l'introduction d'une forme bilinéaire supplémentaire b_k pour chaque élément de la suite u_k qui sera définie ultérieurement. Plus précisément, on définit $b(u)$ qui vérifie soit les hypothèses (H5) ou (H6) ci dessous (sur un borné, par exemple)

(H5a) coercivité faible

$$b(u)(\phi,\phi) \geq \lambda_0 (G(u),\phi)^2 \forall \phi \in V$$

(H5b) continuité: $|b(u)(\phi,\psi)| \leq \mu_0 ||G(u)||\phi||||\psi|| \forall \phi,\psi \in V$

(H6a) $(1+\epsilon)$-coercivité forte

$$b(u)(\phi,\phi) \geq \lambda_1 ||G(u)||^{1+\epsilon} ||\phi||^2 \forall \phi \in V$$

(H6b) $(1+\epsilon)$-continuité $|b(u)(\phi,\psi)| \leq \mu_1 ||G(u)||^{1+\epsilon} ||\phi||||\psi|| \forall \phi,\psi \in V$.

On a le

Théorème 6.7 *Sous les hypothèses (H1), (H2), (H3), (H4), et (H5) ou (H6) on a:*

• *Le problème de minimisation admet une solution unique u.*

On considère u_0 donné. Soit u_k un élément de la suite. L'élément u_{k+1} est construit comme $u_k + \Delta_k$, où Δ_k est la solution du problème variationnel

$$\forall \phi \in V, (H(u_k)\Delta_k,\phi) + b_k(\Delta_k,\phi) = -(G(u_k),\phi). \tag{6.7.3}$$

- *La suite u_k est bien définie, et elle converge vers u*
- *Il existe deux constantes γ_1 et γ_2 telles que*

$$\gamma_1||u_{k+1}-u_k|| \leq ||u-u_k|| \leq \gamma_2||u_{k+1}-u_k||$$

- *Il existe une constante γ_3 telle que*

$$||u_{k+1}-u|| \leq \gamma_3||u_k-u||^2.$$

On commence par donner des exemples de formes de la fonctionnelle $b(u)$. On notera b_k la fonctionnelle $b(u_k)$ pour simplifier les notations.

Pour $b_k(\phi,\psi) = \lambda^k(G(u_k),\phi)(G(u_k),\psi)$, les hypothèses (H5a) et (H5b) sont vérifiées. En revanche, on n'a pas l'hypothèse (H6a).

Pour $b_k(\phi,\psi) = \lambda^k||G(u_k)||^{1+\epsilon}(\phi,\psi)$, les hypothèses (H5a), (H5b), (H6a), (H6b) sont toutes vérifiées.

Preuve Etapes de la démonstration.

On commence par démontrer que la suite $J(u_k)$ est décroissante si μ_0 (resp. μ_1) est choisi de manière adéquate dans l'hypothèse (H5a) (resp. (H6a)). On en déduit que les termes de la suite restent dans un fermé borné fixe.

Dans un deuxième temps, en choisissant dans la formulation variationnelle et dans l'égalité de développement de Taylor des valeurs astucieuses de ϕ, on montre des estimations sur la différence de deux termes et sur la différence d'un terme de la suite avec la limite. Pour cela, on utilise de manière cruciale l'inégalité de coercivité sur le fermé borné.

On définit

$$U = \{v \in V, J(v) \leq J(u_0)\}.$$

• Si J admet un minimum, il est dans U. Comme J est infinie en l'∞, U est borné. Il est fermé. En effet, si $v_j \in U, v_j \to v$, alors $J(u_0) \geq J(v_j) = J(v) + (G(v), v_j - v) + \frac{1}{2}(H(v+\theta(v_j-v))(v_j-v), v_j-v)$. Comme H est coercive, on a $J(u_0) \geq J(v_j) \geq J(v) + (G(v), v_j - v)$. Comme v ne dépend pas de j, on passe à la limite et $J(u_0) \geq J(v)$. Il vient $v \in U$.

• Le problème variationnel linéaire (6.7.3) admet une seule solution Δ_k. Prenant $\phi = \Delta_k$ dans l'égalité variationnelle (6.7.3), on en déduit que

$$(H(u_k)\Delta_k, \Delta_k) + b_k(\Delta_k, \Delta_k) = -(G(u_k), \Delta_k). \tag{6.7.4}$$

Utilisant la coercivité de H et la positivité de b_k, on en déduit

$$(H(u_k)\Delta_k, \Delta_k) + b_k(\Delta_k, \Delta_k) \geq \alpha_U||\Delta_k||^2.$$

On utilise l'inégalité

$$|-(G(u_k), \Delta_k)| \leq ||\Delta_k||||G(u_k)||.$$

On divise, si $\Delta_k \neq 0$, par la norme et on obtient

$$\alpha||\Delta_k|| \leq ||G(u_k)||. \tag{6.7.5}$$

Désignant par $||G||$ le maximum de G sur le fermé U, on en déduit

$$||\Delta_k|| \leq \alpha^{-1}||G||.$$

Soit

$$U_1 = \{v \in V, ||v - w|| \leq \alpha^{-1}||G||, w \in U\}$$

Il vient $u_{k+1} = u_k + \Delta_k \in U_1$.

• Il s'agit maintenant de contrôler le terme $J(u_{k+1})$ par rapport au terme $J(u_k)$; On effectue un développement de Taylor pour J au voisinage de u_k. Ainsi

$$J(u_{k+1}) - J(u_k) = (G(u_k), \Delta_k) + \frac{1}{2}(H(u_k + \theta\Delta_k)\Delta_k, \Delta_k)$$

d'où, en utilisant l'égalité (6.7.4) pour remplacer le terme $(G(u_k), \Delta_k)$:

$$J(u_{k+1}) - J(u_k) = -\frac{1}{2}(H(u_k)\Delta_k, \Delta_k) - b_k(\Delta_k, \Delta_k) + \frac{1}{2}([H(u_k + \theta\Delta_k) - H(u_k)]\Delta_k, \Delta_k).$$

On note β_1 la constante de Lipschitz pour H sur U_1. Si on suppose $u_k \in U_0$, on trouve $u_k + \theta\Delta_k \in U_1$. Ceci permet de minorer le terme $-\frac{1}{2}([H(u_k + \theta\Delta_k) - H(u_k)]\Delta_k, \Delta_k)$. En utilisant la coercivité de H, on trouve l'inégalité

$$J(u_k) - J(u_{k+1}) \geq \frac{\alpha}{2}||\Delta_k||^2(1 - \frac{\beta_1}{\alpha}||\Delta_k||) + b_k(\Delta_k, \Delta_k) \geq \frac{\alpha}{2}||\Delta_k||^2(1 - \frac{\beta_1}{\alpha}||\Delta_k||).$$

Deux cas se présentent. Dans cette inégalité, on doit contrôler le signe du second membre.

- Si $||\Delta_k||$ est petit, c'est-à-dire $||\Delta_k|| \leq (1 - C)\frac{\alpha}{\beta_1}$, alors $J(u_k) - J(u_{k+1}) \geq \frac{\alpha C}{2}||\Delta_k||^2$ en utilisant uniquement la positivité de b_k.

  Dans le cas contraire, on utilise la forme de b.

- On suppose vérifiées les hypothèses (H5).

  Dans ce cas, le terme $b_k(\Delta_k, \Delta_k)$ vérifie

  $$b_k((\Delta_k, \Delta_k) \geq \lambda_0((G(u_k), \Delta_k))^2.$$

  On contrôle alors que par emploi de la relation (6.7.4), on trouve

  $$-(G(u_k), \Delta_k) \geq \alpha||\Delta_k||^2$$

  donc on tire

  $$((G(u_k), \Delta_k))^2 \geq \alpha^2||\Delta_k||^4$$

  Alors

  $$J(u_k) - J(u_{k+1}) \geq \frac{\alpha}{2}||\Delta_k||^2 1 + \lambda_0\alpha^2||\Delta_k||^4 - \frac{1}{2}\beta_1||\Delta_k||^3.$$

  La somme des deux derniers termes est du signe de $\lambda_0\alpha^2||\Delta_k|| - \frac{\beta_1}{2}$ donc est positive dès que $||\Delta_k|| \geq \frac{\beta_1}{2\lambda_0\alpha^2}$

  Si on choisit λ_0 de sorte que $\frac{\beta_1}{2\lambda_0\alpha^2} \leq (1 - C)\frac{\alpha}{\beta_1}$, soit

  $$\lambda_0 > \frac{\beta_1^2}{2\alpha^3}$$

il existe C telle que $\frac{\beta_1}{2\lambda_0\alpha^2} \leq (1-C)\frac{\alpha}{\beta_1}$. Dans ce cas, on voit que si $||\Delta_k|| \geq (1-C)\frac{\alpha}{\beta_1}$, on obtient

$$||\Delta_k|| \geq \frac{\beta_1}{2\lambda_0\alpha^2}$$

et donc

$$J(u_k) - J(u_{k+1}) \geq \frac{\alpha}{2}||\Delta_k||^2.$$

En résumé, sous cette hypothèse sur λ_0, on trouve, pour tout Δ_k

$$J(u_k) - J(u_{k+1}) \geq \frac{\alpha C}{2}||\Delta_k||^2. \tag{6.7.6}$$

- Dans le cas où J vérifie les hypothèses (H6) pour $\epsilon = 1$, et si la constante $lambda_1$ (que l'on suppose assez grande) vérifie $\lambda_1 > \frac{\beta_1^2}{8\alpha^3}$, on vérifie que $\lambda_1\alpha^2||\Delta_k||^2 + \frac{\alpha}{2} - \frac{\beta_1}{2}||\Delta_k|| \geq \frac{8\mu_0\alpha^3-\beta_1^2}{16\mu_0\alpha^2} = \delta_0 > \frac{\alpha}{2}$, et donc $J(u_k) - J(u_{k+1}) \geq \delta_0||\Delta_k||^2$ (la condition sur λ_1 est plus faible).

  • Le raisonnement est le même si l'hypothèse (H6) est vérifiée. En effet, on obtient

$$J(u_k) - J(u_{k+1}) \geq \frac{\alpha}{2}||\Delta_k||^2(1 - \frac{\beta_1}{\alpha}||\Delta_k||) + \mu_0||G(u_k)||^{1+\epsilon}||\Delta_k||^2,$$

et, utilisant (6.7.5), on obtient

$$J(u_k) - J(u_{k+1}) \geq \frac{\alpha}{2}||\Delta_k||^2[\frac{\alpha}{2} - \frac{\beta_1}{2}||\Delta_k||) + \mu_0\alpha^{1+\epsilon}||\Delta_k||^{1+\epsilon}],$$

Lorsque μ_0 grand, le minimum de cette fonction est strictement positif pour tout $\epsilon > 0$ (il s'écrit $\frac{\alpha}{m2} - \epsilon\psi(\epsilon)\mu_0^{-\epsilon}$), donc l'inégalité obtenue est toujours valable.

On a démontré que la suite $J(u_{k+1}) < J(u_k)$ lorsque $u_k \in U$. De $u_0 \in U$, on déduit alors $J(u_1) < J(u_0)$ donc $u_1 \in U$. Ainsi, par récurrence, $J(u_{k+1} < J(u_k)$ donc $u_{k+1} \in U$. La suite $J(u_k)$, décroissante et minorée, converge. Ainsi la suite $J(u_k) - J(u_{k+1})$ tend vers 0, donc Δ_k tend vers 0 grâce à l'inégalité (6.7.6)6.7.6).

Il faut montrer désormais que la suite u_k converge. On écrit la formule de Taylor $(G(u_k), \phi) = (G(u), \phi) + (H(u + \theta'(u_k - u))(u_k - u), \phi)$, ce qui donne

$$(H(u_k)\Delta_k, \phi) + b_k(\Delta_k, \phi) = -(H(u + \theta'(u_k - u))(u_k - u), \phi), \tag{6.7.7}$$

par l'application de l'égalité variationnelle définissant Δ_k. Comme U est convexe (la fonctionnelle est convexe car son Hessien est coercif), $u + \theta'(u_k - u)$ est dans U. Ainsi, prenant $\phi = u_k - u$ et appliquant les inégalités de Cauchy-Schwartz à $(H(u_k)\Delta_k, u_k - u) + b_k(\Delta_k, u_k - u) = -(H(u + \theta'(u_k - u))(u_k - u), u_k - u)$, on trouve, notant γ la constante majorant les normes de $H(u_k)$ et de b_k (ce qui est possible puisque $u_k \in U$ donc $G(u_k)$ est borné par $||G||$):

$$\gamma||\Delta_k||||u_k - u|| \geq \alpha||u_k - u||^2.$$

La convergence de Δ_k vers 0 et l'inégalité $||u_k - u|| \leq \alpha^{-1}\gamma||\Delta_k||$ entrainent la convergence de u_k vers u. De plus, on vérifie facilement que si on considère $\phi = \Delta_k$ dans l'égalité (6.7.7), alors on trouve $||u_k - u|| \geq \frac{\alpha}{\gamma}||\Delta_k||$.

On montre enfin la convergence quadratique. L'égalité (6.7.7) donne alors, écrivant $\delta_k = u_k - u$ et $\Delta_k = \delta_{k+1} - \delta_k$, l'égalité

$$(H(u_k)\delta_{k+1}, \phi) + b_k(\delta_{k+1}, \phi) = (H(u_k)\delta_k, \phi) + b_k(\delta_k, \phi) - (H(u + \theta'(u_k - u))\delta_k, \phi)$$

puis utilisant pour le terme de gauche la coercivité de H, pour le terme de droite le caractère Lipschitz de H, et la positivité de b_k pour le terme de gauche, il reste, pour $\phi = \delta_{k+1}$,

$$\alpha||\delta_{k+1}||^2 \leq \mu_1||G(u_k)||^{1+\epsilon}||\delta_{k+1}||||\delta_k|| + \beta_1||\delta_k||^2||\delta_{k+1}||$$

d'où on déduit

$$\alpha||\delta_{k+1}|| \leq \mu_1||G(u_k)||^{1+\epsilon}||\delta_k|| + \beta_1||\delta_k||^2$$

Comme G est Lipschitz (puisque H est continue) et que $G(u) = 0$, on en déduit $||G(u_k)|| = ||G(u_k) - G(u)|| \leq \Gamma||\delta_k||$. Comme cette quantité est bornée par D constante, on en déduit l'inégalité

$$\alpha||\delta_{k+1}|| \leq (\mu_1\Gamma^{1+\epsilon}D^\epsilon + \beta_1)||\delta_k||^2,$$

qui est la convergence quadratique.

Cette démonstration, bien que longue et fastidieuse, est importante et intéressante, car elle permet de manipuler les formulations variationnelles, de voir l'importance de la coercivité, de voir les choix de fonctions test. Notons que les deux hypothèses possibles (H5) ou (H6) conduisent au résultat, et sont utilisées de manière cruciale dans la preuve de la décroissance de $J(u_k)$, preuve suffisante pour la convergence. C'est pour cela que cette méthode conduit toujours à une solution. D'autre part, dire que μ_0 est assez grand est possible car on est libre du choix de b pour le problème d'optimisation. On peut rapprocher cette méthode des méthodes de pénalisation.

6.8 Algorithmes d'optimisation avec contraintes

Les trois algorithmes que je compte présenter correspondent aux algorithmes de minimisation sous contraintes.

6.8.1 Le gradient avec projection

On suppose dans ce premier cas que l'espace des contraintes K est convexe. On rappelle dans ce cas qu'il existe une projection sur K, définie par

$$||x - p_K(x)|| = \inf_{y \in K} ||x - y||$$

et caractérisé par l'inégalité

$$(y - p_K(x), x - p_K(x)) \leq 0 \forall y \in K.$$

Un des problèmes essentiels d'un algorithe de gradient, lorsqu'on n'est pas dans le cas du gradient réduit, est qu'il ne donne pas à l'itération $n+1$ un élément de l'espace des contraintes car on ne sait pas si la direction $-J'(x_n)$ est une direction admissible pour l'espace des contraintes si x_n est dans K. D'autre part, la projection est une application contractante, donc $||p_K(x)-p_K(y)|| \leq ||x-y||$, ce qui implique que $||p_K(x-\alpha J'(x))-p_K(y)|| \leq ||x-\alpha J'(x)-y||$ donc en projetant le résultat d'un algorithme de gradient, on se rapproche plus de y solution du problème de minimisation. L'algorithme de gradient avec projection est un algorithme de la forme

$$x_{n+1} = p_K(x_n - \rho_n J'(x_n)).$$

Proposition 6.9 *Si J est convexe et que K est convexe, un point solution du problème de minimisation de J sur K est un point stationnaire de l'égalité $x_0 = p_K(x_0 - \alpha J'(x_0))$.*

Preuve On suppose que x_0 est une solution du problème de minimisation. Comme J est convexe, la condition d'Euler est **équivalente** à

$$\forall y \in K, (J'(x_0), y - x_0) \geq 0.$$

On en déduit, pour tout $\alpha > 0$, que

$$(y - x_0, -\alpha J'(x_0)) \leq 0$$

donc

$$\forall y \in K, (y - x_0, x_0 - \alpha J'(x_0) - x_0) \leq 0$$

ce qui est la caractérisation de la projection de $x_0 - \alpha J'(x_0)$ en x_0. On en déduit que

$$\forall \alpha > 0, x_0 = p_K(x_0 - \alpha J'(x_0)).$$

Réciproquement, soit $\alpha_0 > 0$ tel que $x_0 = p_K(x_0 - \alpha_0 J'(x_0))$. On a alors

$$\forall y \in K, (y - x_0, x_0 - \alpha_0 J'(x_0) - x_0) \leq 0$$

soit

$$\forall y \in K, (y - x_0, J'(x_0)) \geq 0$$

ce qui, par la caractérisation dans le cas convexe, implique que x_0 est solution du problème de minimisation.

On a même un résultat lorsque le pas de l'algorithme de gradient avec projection est bien choisi:

Théorème 6.8 *On suppose K convexe fermé non vide, J bornée inférieurement sur K, de classe C^1, Lipschitz uniformément sur K dont une constante de Lipschitz est L:*

$$||J'(x) - J'(y)|| \leq L||x - y||.$$

Si il existe $\epsilon > 0$ tel que, pour tout n, $\rho_n \in [\epsilon, \frac{2}{L}(1-\epsilon)]$, la suite x_n donnée par l'algorithme de gradient avec projection vérifie

$$||x_{n+1} - x_n|| \to 0$$

Tous les points d'adhérence de cette suite sont des points stationnaires.

Preuve On vérifie que, par caractérisation de la projection

$$\forall y \in K, (y - p_K(x_n - \rho_n J'(x_n)), x_n - \rho_n J'(x_n) - p_K(x_n - \rho_n J'(x_n))) \leq 0,$$

donc

$$\forall y \in K, (y - x_{n+1}, x_n - \rho_n J'(x_n) - x_{n+1}) \leq 0.$$

On commence l'algorithme avec un point x_0, pas forcément dans K. En revanche, pour $n \geq 1$, tous les termes de la suite sont dans K donc on peut prendre $y = x_n$. On en déduit l'inégalité:

$$(x_n - x_{n+1}, x_n - x_{n+1}) - \rho_n(x_n - x_{n+1}, J'(x_n)) \leq 0$$

soit

$$(J'(x_n), x_{n+1} - x_n) \leq -\frac{1}{\rho_n}||x_n - x_{n+1}||^2.$$

On utilise

$$J(x_{n+1}) - J(x_n) - (J'(x_n), x_{n+1} - x_n) = \int_0^1 (J'(x_n + t(x_{n+1} - x_n)) - J'(x_n), x_{n+1} - x_n) dt.$$

Comme on a $L-$Lipschitz, on trouve

$$\begin{array}{ll} |J(x_{n+1}) - J(x_n) - (J'(x_n), x_{n+1} - x_n)| & \leq \int_0^1 ||J'(x_n + t(x_{n+1} - x_n)) - J'(x_n)||||x_{n+1} - x_n||dt \\ & \leq L(\int_0^1 t dt||x_{n+1} - x_n||)||x_{n+1} - x_n|| \\ & \leq \frac{L}{2}||x_{n+1} - x_n||^2 \end{array}$$

On utilise alors la convexité de J pour obtenir

$$J(x_{n+1}) \geq J(x_n) + (J'(x_n), x_{n+1} - x_n).$$

On en déduit l'inégalité

$$J(x_{n+1}) - J(x_n) - (J'(x_n), x_{n+1} - x_n) \leq \frac{L}{2}||x_{n+1} - x_n||^2$$

et de l'inégalité de caractérisation de la projection on déduit

$$(J'(x_n), x_{n+1} - x_n) \leq -\frac{1}{\rho_n}||x_{n+1} - x_n||^2$$

donc

$$J(x_{n+1}) - J(x_n) \leq (\frac{L}{2} - \frac{1}{\rho_n})||x_{n+1} - x_n||^2.$$

On utilise alors $\frac{1}{\rho_n} \in [\frac{L}{2}\frac{1}{1-\epsilon}, \frac{1}{\epsilon}]$ soit $\frac{L}{2} - \frac{1}{\rho_n} \in [\frac{L}{2} - \frac{1}{\epsilon}, -\frac{L}{2}\frac{\epsilon}{1-\epsilon}]$, donc finalement la suite $J(x_n)$ est décroissante et on a

$$\frac{L}{2}\frac{\epsilon}{1-\epsilon}||x_{n+1} - x_n||^2 \leq J(x_n) - J(x_{n+1}).$$

La suite $J(x_n)$ est minorée et décroissante, donc elle converge. La décroissance de la suite vient uniquement de l'hypothèse sur le pas... On en déduit que $J(x_{n+1})-J(x_n)$ tend vers 0, donc il en est de même de $x_{n+1}-x_n$.

Enfin, si y est une valeur d'adhérence de la suite, $x_{\phi(n)}$ tend vers y, dont on déduit que $x_{\phi(n)+1}$ tend aussi vers y. De l'égalité $x_{\phi(n)+1} = p_K(x_{\phi(n)} - \rho_{\phi(n)}J'((x_{\phi(n)}))$, on ne peut rien déduire car on ne sait pas si la suite $\rho_{\phi(n)}$ converge. Il s'agit alors de remarquer que cette suite est bornée, donc on peut extraire une sous-suite convergente, que l'on note $\rho_{\phi(\psi(n))}$. Elle converge vers $\alpha > 0$, et de la continuité de J', de la continuité de la projection sur un convexe fermé, on déduit l'égalité $y = p_K(y-\alpha J'(y))$.

6.8.2 Pénalisation des contraintes

Le premier concerne la pénalisation des contraintes;on cherche à minimiser $J(u)$ sous les contraintes $F_j(u) \leq 0$. On introduit

$$J_\varepsilon(v) = J(v) + \frac{1}{\varepsilon}\sum_{j=1}^{j=M}[\max(F_j(v),0)]^2$$

On a

Théorème 6.9 On suppose $V = \mathbb{R}^N$.

On suppose que J est continue, $\alpha-$convexe, que les F_j sont convexes et que l'ensemble des contraintes K est non vide. Si u_ε est l'unique solution de $\inf J_\varepsilon$ *et u l'unique solution de* $\inf_{v\in K} J$*, alors*

$$\lim_{\varepsilon\to 0} u_\varepsilon = u.$$

De plus, sous l'hypothèse $J, F_1, .., F_M$ continuement différentiables, les contraintes sont qualifiées en u, et la famille des contraintes actives est régulière en u, les multiplicateurs de Lagrange λ_j du problème non pénalisé vérifient

$$\lambda_i = \lim_{\varepsilon\to 0}\frac{2}{\varepsilon}\max(F_i(u_\varepsilon),0).$$

Preuve L'existence et l'unicité de u et de u_ε sont claires car $u \to \frac{1}{\varepsilon}\sum_{j=1}^{j=M}[\max(F_j(v),0)]^2 = \frac{G(u)}{\varepsilon}$ est une fonctionnelle convexe.

On sait d'autre part que

$$J_\epsilon(u_\epsilon) \leq \inf_K J_\epsilon,$$

et comme, pour $y \in K$, $J_\epsilon(y) = J(y)$, on vérifie que $J_\epsilon(u_\epsilon) \leq J(u)$. Comme d'autre part

$$J_\epsilon(u_\epsilon) \geq J(u_\epsilon)$$

on a l'inégalité $J(u_\epsilon) \leq J(u)$. Comme J est $\alpha-$convexe, la suite u_ε est bornée. On peut extraire une sous-suite convergeant vers une limite $\tilde{u}$. De l'inégalité $J(u_\varepsilon) \leq J(u_\varepsilon) + \frac{G(u_\varepsilon)}{\varepsilon} \leq J(u)$, on déduit l'inégalité $G(u_\varepsilon) \leq \varepsilon(J(u) - J(u_\varepsilon))$, ce qui implique que $G(\tilde{u}) = 0$ (car G est continue donc $G(u_\varepsilon)$ tend vers $G(\tilde{u})$ pour la suite extraite et que $\varepsilon \to 0$). Cela exprime que $\tilde{u} \in K$. Ainsi comme $J(u_\varepsilon) \leq J(u)$, en considérant

toujours la même suite extraite et la continuité de J, on trouve $J(\tilde{u}) \leq J(u)$. On a démontré que $\tilde{u} = u$ et donc la suite u_ε admet une seule valeur d'adhérence.

Pour les multiplicateurs de Lagrange, on trouve, par définition de la dérivée en un point x de $(\max(x,0))^2$ qui vaut $2\max(x,0)$, l'égalité

$$J'(u_\varepsilon) + \frac{1}{\varepsilon}\sum_{j=1}^{j=M} 2\max(F_j(u_\varepsilon),0)F_j'(u_\varepsilon) = 0.$$

Comme J', F_j' sont continues, on trouve $J'(u_\varepsilon) \to J'(u)$ et $F_j'(u_\varepsilon) \to F_j'(u)$. On suppose que pour un élément j, on ait $F_j(u_\varepsilon) \to F_j'(u) < 0$. Alors il existe ε_0 tel que, pour $\varepsilon < \varepsilon_0$, $F_j(u_\varepsilon) < 0$ et donc on trouve $\max(F_j(u_\varepsilon, 0) = 0$. L'égalité devient, pour ε assez petit

$$J'(u_\varepsilon) + \frac{1}{\varepsilon}\sum_{j\in I(u)} 2\max(F_j(u_\varepsilon),0)F_j'(u_\varepsilon) = 0.$$

D'autre part, pour $j \in I(u)$, on vérifie qu'il existe une suite $\lambda_1, ..\lambda_M$, avec $\lambda_j = 0$ si $j \notin I(u)$, telle que $J'(u) + \sum \lambda_j F_j'(u) = 0$. Ainsi on trouve

$$J'(u_\varepsilon) - J'(u) + (\frac{1}{\varepsilon}\sum_{j\in I(u)} 2\max(F_j(u_\varepsilon),0) - \lambda_j)F_j'(u_\varepsilon) = 0.$$

La famille $(F_j'(u)$ est libre, donc, par continuité, pour ε assez petit, la famille $(F_j'(u_\varepsilon)$ est libre. De plus, en formant le produit scalaire avec tous les $F_j'(u_\varepsilon)$, le déterminant du système obtenu est, toujours pour ε petit, minoré par une constante. Ceci permet d'assurer le fait que $\frac{2}{\varepsilon}\max(F_j'(u_\varepsilon, 0)$ est borné et donc que

$$\frac{2}{\varepsilon}\max(F_j'(u_\varepsilon),0)(F_j'(u_\varepsilon) - F_j'(u))$$

tend vers 0 pour tout j. On en conclut sur la convergence, sur la base fixe des $F_j'(u)$, de $J'(u_\varepsilon) + \frac{2}{\varepsilon}\max(F_j'(u_\varepsilon),0)F_j'(u)$, d'où le résultat de convergence des coefficients.

6.8.3 Algorithme d'Uzawa

En fait, il s'agit d'une méthode de recherche de point selle.

On sait que, pour $\mathcal{L}(v,q) = J(v) + (q, F(v))$,

$$\forall q \geq 0, \mathcal{L}(u,q) \leq \mathcal{L}(u,p) \leq \mathcal{L}(v,p)$$

Ainsi

$$\forall q, q \geq 0, (p-q, F(u)) \geq 0.$$

Il vient, pour $\mu > 0$

$$(p-q, p-(p+\mu F(u))) \leq 0 \forall q \in (\mathbb{R}_+)^M.$$

Ceci indique que, pour tout $\mu > 0$, la projection de $p + \mu F(u)$ est p sur l'espace $(\mathbb{R}_+)^M$.

On définit alors, pour μ paramètre fixé, la suite (u^n, p^n) donnée par

$$\mathcal{L}(u^n, p^n) = \inf_{v \in V} \mathcal{L}(v, p^n)$$

et le multiplicateur p^{n+1} est la projection sur $(\mathbb{R}_+)^m$ de $p^n + \mu F(u^n)$.

Cette projection se fait très simplement: pour chaque coordonnée de $p^n + \mu F(u^n)$, si la coordonnée est positive ou nulle, on ne la change pas, mais si elle est strictement négative, on la met à 0. Cet algorithme converge: ce qui s'écrit dans le

Théorème 6.10 *On suppose J $\alpha-$convexe différentiable, Lipschitz de constante C et que le lagrangien $\mathcal{L}$ admet un point selle (u, p). Alors, pour $0 < \mu < \frac{2\alpha}{C^2}$, la suite u^n donnée par l'algorithme d'Uzawa converge vers u .*

On admettra la démonstration de ce théorème.

Chapter 7

Introduction aux méthodes de discrétisation des équations aux dérivées partielles

On souhaite étudier les équations aux dérivées partielles suivantes:
i) Equation de la chaleur $\partial_t u - \partial^2_{x^2} u = 0$
ii) Equation des ondes $\partial^2_{t^2} u - \partial^2_{x^2} u = 0$
iii) Equation de Laplace avec condition de Dirichlet

$$-\Delta u = f \text{ sur } \Omega, u|_{\partial\Omega} = 0.$$

7.1 Les différences finies

Pour les deux premières équations, on souhaite ramener ce problème continu à un problème discrétisé, c'est-à-dire faisant intervenir les valeurs de la solution u aux points $(j\Delta x, n\Delta t)$. Pour cela, il s'agit de calculer la dérivée première et la dérivée seconde en fonction des points voisins, sur le modèle de $\frac{u(x+h)-u(x)}{h} \simeq u'(x)$.

On écrit pour cela $u^j_n = u(j\Delta x, n\Delta t)$ pour u de classe C^4, sur laquelle on applique la formule de Taylor-Young.

$$\begin{aligned} u^{j+1}_n = \quad & u^j_n + \Delta x \partial_x u(j\Delta x, n\Delta t) + \tfrac{1}{2}(\Delta x)^2 \partial^2_{x^2} u(j\Delta x, n\Delta t) + \tfrac{1}{6}(\Delta x)^3 \partial^3_{x^3} u(j\Delta x, n\Delta t) \\ & + \tfrac{1}{24}(\Delta x)^4 \partial^4_{x^4} u((j+\theta)\Delta x, n\Delta t). \end{aligned}$$

Il ne suffit pas de u^{j+1}_n et de u^j_n pour connaitre la dérivée seconde; il faut un troisième point. On prend u^{j-1}_n, et on a

$$\begin{aligned} u^{j-1}_n = \quad & u^j_n - \Delta x \partial_x u(j\Delta x, n\Delta t) + \tfrac{1}{2}(\Delta x)^2 \partial^2_{x^2} u(j\Delta x, n\Delta t) - \tfrac{1}{6}(\Delta x)^3 \partial^3_{x^3} u(j\Delta x, n\Delta t) \\ & + \tfrac{1}{24}(\Delta x)^4 \partial^4_{x^4} u((j-\theta')\Delta x, n\Delta t). \end{aligned}$$

En additionnant les deux relations, on trouve ainsi

$$u^{j+1}_n + u^{j-1}_n - 2u^j_n = (\Delta x)^2 \partial^2_{x^2} u(j\Delta x, n\Delta t) + \frac{(\Delta x)^4}{24}[\partial^4_{x^4} u(j+\theta)\Delta x, n\Delta t) + \partial^4_{x^4} u(j-\theta')\Delta x, n\Delta t)],$$

ainsi

$$\partial^2_{x^2}u(j\Delta x,n\Delta t)=\frac{u_n^{j+1}+u_n^{j-1}-2u_n^j}{(\Delta x)^2}-\frac{(\Delta x)^2}{24}[\partial^4_{x^4}u(j+\theta)\Delta x,n\Delta t)+\partial^4_{x^4}u(j-\theta')\Delta x,n\Delta t)],$$

ce qui donne, sur un compact K:

$$|\partial^2_{x^2}u(j\Delta x,n\Delta t)-\frac{u_n^{j+1}+u_n^{j-1}-2u_n^j}{(\Delta x)^2}|\leq\frac{(\Delta x)^2}{12}||\partial^4_{x^4}u(j\Delta x,n\Delta t)||.$$

On utilise aussi la relation

$$u^j_{n+1}-u^j_n=\Delta t\partial_t u(j\Delta x,n\Delta t)+O((\Delta t)^2)$$

qui nous permet d'écrire des schémas pour l'équation des ondes et pour l'équation de la chaleur.

Pour l'équation des ondes, on écrit par exemple

$$\frac{u^j_{n+1}-2u^j_n+u^j_{n-1}}{(\Delta t)^2}-\frac{u_n^{j+1}-2u_n^j+u_n^{j-1}}{(\Delta x)^2}=0 \tag{7.1.1}$$

qui s'appelle un schéma **explicite** puisque u^j_{n+1} est connu explicitement en fonction des valeurs de u^l_k pour $k\leq n$, c'est-à-dire que l'on connait les valeurs aux points situés au temps $(n+1)\Delta t$ en fonction des temps précédents.

On écrit aussi

$$\frac{u^j_{n+1}-2u^j_n+u^j_{n-1}}{(\Delta t)^2}-\frac{u_{n+1}^{j+1}-2u_{n+1}^j+u_{n+1}^{j-1}}{(\Delta x)^2}=0 \tag{7.1.2}$$

qui s'appelle un schéma **implicite** car on ne peut pas déterminer les valeurs au temps $(n+1)\Delta t$ en fonction des valeurs aux temps précédents.

On suppose que l'on se place sur un compact, par exemple $x\in[0,1]$. On vérifie que la discrétisation correspond aux $\Delta x=\frac{1}{N}$ et $j\in[0,N]$. En ajoutant des conditions aux extrémités, on se ramène à un système de la forme

$$A\begin{pmatrix}u^1_{n+1}\\u^2_{n+1}\\.\\.u^N_{n+1}\end{pmatrix}=\begin{pmatrix}2u_{n-1}-u_{n-1}\end{pmatrix}.$$

C'est un système linéaire de la forme $Ax=b$ qui peut se résoudre par des méthodes d'approximation du cours d'optimisation, sur la fonctionnelle

$$J(x)=\frac{1}{2}(Ax,x)-(b,x).$$

Pour l'équation de la chaleur, on écrit les mêmes schémas:

$$\frac{u^j_{n+1}-u^j_n}{\Delta t}-\frac{u_n^{j+1}-2u_n^j+u_n^{j-1}}{(\Delta x)^2}=0 \tag{7.1.3}$$

qui est un schéma **explicite**, et

$$\frac{u^j_{n+1}-u^j_n}{\Delta t}-\frac{u^{j+1}_{n+1}-2u^j_{n+1}+u^{j-1}_{n+1}}{(\Delta x)^2}=0 \tag{7.1.4}$$

qui est un schéma implicite.

Pour affiner l'analyse, nous introduisons les fonctions, polynômiales de degré 3 au plus, qui soient de classe C^2 sur $[0,1]$ et qui coincident avec tous les u^j_n en tous les points $j\Delta x$ pour $\Delta x=\frac{1}{N}$. Pour ces fonctions là, on vérifie que la dérivée seconde sur tous les intervalles $[j\Delta x,(j+1)\Delta x]$ est exactement égale à $\frac{u^{j+1}_{n+1}-2u^j_{n+1}+u^{j-1}_{n+1}}{(\Delta x)^2}$, puisque la fonction est de dérivée quatrième nulle sur chaque intervalle. On peut donc déduire une formulation continue de cette formulation discrète, en remplacant le terme $\frac{u^{j+1}_{n+1}-2u^j_{n+1}+u^{j-1}_{n+1}}{(\Delta x)^2}$ par $\frac{u^{n+1}(x+\Delta x)+u^{n+1}(x-\Delta x)-2u^{n+1}(x)}{(\Delta x)^2}$. On emploiera en permanence cette notation désormais (utilisant l'indice pour la position en espace et l'exposant pour l'incrément en temps). On écrit les schémas sous la forme

$$\frac{u^{n+1}(x)-u^n(x)}{\Delta t}=\frac{u^{n+1}(x+\Delta x)+u^{n+1}(x-\Delta x)-2u^{n+1}(x)}{(\Delta x)^2}$$

$$\frac{u^{n+1}(x)-u^n(x)}{\Delta t}=\frac{u^{n}(x+\Delta x)+u^{n}(x-\Delta x)-2u^{n}(x)}{(\Delta x)^2}$$

Considérant la transformée de Fourier en x des deux égalités ci-dessus et utilisant la relation

$$\frac{e^{i\xi\Delta x}+e^{-i\xi\Delta x}-2}{(\Delta x)^2}=-4\frac{\sin^2\frac{\xi\Delta x}{2}}{(\Delta x)^2}$$

on trouve respectivement, en notant

$$v^n(\xi)=\int_{-\infty}^{+\infty}e^{-ix\xi}u^n(x)dx$$

la relation pour le schéma implicite pour l'équation de la chaleur

$$(1+4\sin^2\frac{\xi\Delta x}{2}\frac{\Delta t}{(\Delta x)^2})v^{n+1}(\xi)=v^n(\xi)$$

et la relation pour le schéma explicite pour l'équation de la chaleur

$$v^{n+1}(\xi)=(1-4\sin^2\frac{\xi\Delta x}{2}\frac{\Delta t}{(\Delta x)^2})v^n(\xi).$$

Le but est d'assurer la convergence de la suite pour tout n (c'est à dire lorsque le temps devient grand).

• Dans le cas du schéma explicite, il est nécessaire pour cela que le coefficient $(1-4\sin^2\frac{\xi\Delta x}{2}\frac{\Delta t}{(\Delta x)^2})$ soit de module plus petit que 1, soit l'inégalité

$$4\sin^2\frac{\xi\Delta x}{2}\frac{\Delta t}{(\Delta x)^2}>-2$$

ce qui est possible lorsque le coefficient $\frac{\Delta t}{(\Delta x)^2}$ est plus petit que $\frac{1}{2}$. Cette condition s'appelle une condition CFL et doit être vérifiée pour que la suite n'explose pas lorsque Δt tend vers 0 (ce qui est imposé par $[0,T]=\cup_{k\leq\frac{T}{\Delta t}}[k\Delta t,(k+1)\Delta t]$).

• Dans le cas du schéma implicite, le coefficient $(1+4\sin^2\frac{\xi\Delta x}{2}\frac{\Delta t}{(\Delta x)^2})^{-1}$ est toujours plus petit que 1 et le schéma implicite converge toujours.

Pour l'équation des ondes, la situation est similaire, sauf que la relation de récurrence pour la suite est une relation d'ordre 2, et on doit étudier les racines de la relation caractéristique. On trouve par exemple, pour le schéma explicite

$$v^{n+1}(\xi) - 2(1 - 2\sin^2\frac{\xi\Delta x}{2}(\frac{\Delta t}{\Delta x})^2)v^{n+1}(\xi) + v^n(\xi) = 0$$

et pour le schéma implicite

$$v^{n+1}(\xi)(1 + 4\sin^2\frac{\xi\Delta x}{2}(\frac{\Delta t}{\Delta x})^2) - 2v^{n+1}(\xi) + v^n(\xi) = 0.$$

On constate pour le premier schéma que le produit des racines de l'équation caractéristique est 1, donc le produit des modules est égal à 1. Si le discriminant est négatif, les deux racines sont complexes conjuguées de module 1, si le discriminant est positif, une des racines est de module supérieur à 1, donc il n'y a pas convergence.

Pour le deuxième schéma, le produit des racines est $\frac{1}{1+4\sin^2\frac{\xi\Delta x}{2}(\frac{\Delta t}{\Delta x})^2}$ et le discriminant est négatif, elles sont donc complexes conjuguées de module inférieur à 1 (égal à 1 lorsque $\xi\Delta x = 2\pi n$), donc ce schéma est convergent.

Ce schéma n'est pas employé en général; les numériciens préfèrent employer le schéma de Cranck-Nicholson qui se présente de la manière suivante.

On introduit l'opérateur A_h qui est l'opérateur employé dans les algorithmes précédents (le h correspond à Δx). Cet opérateur s'écrit

$$(A_h\phi)_j = -\frac{\phi_{j+1} - 2\phi_j + \phi_{j-1}}{(\Delta x)^2} \tag{7.1.5}$$

sur une suite ϕ_j.

Le schéma utilisé habituellement est alors

$$\frac{u_j^{n+1} + u_j^{n-1} - 2u_j^n}{(\Delta t)^2} + (A_h(\theta u^{n+1} + (1-2\theta)u^n + \theta u^{n+1}))_j = 0.$$

où $\theta \in [0, \frac{1}{2}]$. Le choix $\theta = 0$ correspond à un schéma explicite comme vu précédemment.

La transformée de Fourier appliquée à ce schéma comme cela a été fait précedemment conduit à la relation de récurrence

$$(1 + \alpha(\xi)\theta)v^{n+1}(\xi) - (2 - (1-2\theta)\alpha(\xi))v^n(\xi) + (1 + \alpha(\xi)\theta)v^{n-1}(\xi) = 0,$$

où

$$\alpha(\xi) = 4(\frac{\Delta t}{\Delta x})^2\sin^2\frac{\xi\Delta x}{2}$$

associée à l'équation caractéristique

$$(1 + \alpha(\xi)\theta)r^2 - (2 - (1-2\theta)\alpha(\xi))r + (1 + \alpha(\xi)\theta) = 0,$$

Comme précédemment, le produit des racines est 1, donc si les deux racines sont réelles et ne sont pas égales, le schéma est instable car une des racines est plus grande

que 1. Il vient alors qu'une condition nécessaire de stabilité est donnée par le fait que les deux racines sont complexes conjuguées, donc de module 1. Ceci s'écrit

$$(2(1+\theta\alpha(\xi))-\alpha(\xi))^2-4(1+\alpha(\xi)\theta)^2\leq 0$$

soit $-\alpha(\xi)(4(1+\alpha(\xi)\theta)-\alpha(\xi))\leq 0$ ou encore

$$(4\theta-1)\alpha+4\geq 0.$$

Lorsque $\theta\geq\frac{1}{4}$, cette inégalité est tout le temps vraie. Lorsque $\theta\in[0,\frac{1}{2}]$, on trouve que cette inégalité est vraie pour

$$(\frac{\Delta t}{\Delta x})^2\sin^2\frac{\xi\Delta x}{2}\leq\frac{1}{1-4\theta}$$

ce qui est vrai sous la condition

$$\frac{\Delta t}{\Delta x}\leq\frac{1}{\sqrt{1-4\theta}}.$$

On résume les résultats de cette section dans:

Théorème 7.1 *Soit A_h l'opérateur d'approximation donné par (7.1.5).*

1) Cet opérateur d'approximation vérifie l'inégalité, pour $\phi=(u(j\Delta x))_j$ et u de classe C^4 sur $[0,1]$ et $j\leq N$, $\Delta x=\frac{1}{N}$:

$$|(A_hu)_j+u"(j\Delta x)|\leq\frac{(\Delta x)^2}{12}||u^{(4)}||_{C^0([0,1])}.$$

2) Un schéma explicite pour l'équation de la chaleur s'écrit

$$\frac{u^{n+1}-u^n}{\Delta t}+A_hu^n=0.$$

Il est stable lorsque la condition suivante est satisfaite:

$$\frac{\Delta t}{(\Delta x)^2}\leq\frac{1}{2}.$$

3) Un schéma implicite pour l'équation de la chaleur s'écrit

$$\frac{u^{n+1}-u^n}{\Delta t}+A_hu^{n+1}=0.$$

Il est tout le temps stable.

4) Un schéma explicite pour l'équation des ondes s'écrit

$$\frac{u_j^{n+1}+u_j^{n-1}-2u_j^n}{(\Delta t)^2}+(A_hu^n)_j=0.$$

Il est tout le temps instable

5) Un schéma implicite pour l'équation des ondes s'écrit

$$\frac{u_j^{n+1}+u_j^{n-1}-2u_j^n}{(\Delta t)^2}+(A_hu^{n+1})_j=0.$$

Il est tout le temps stable.

6) Un schéma implicite pour l'équation des ondes **respectant l'invariance par renversement du temps** *est*

$$\frac{u_j^{n+1}+u_j^{n-1}-2u_j^n}{(\Delta t)^2}+(A_h(\theta u^{n+1}+(1-2\theta)u^n+\theta u^{n+1}))_j=0.$$

Il est tout le temps stable pour $\frac{1}{4}\leq\theta\leq\frac{1}{2}$.
Pour $0\leq\theta\leq\frac{1}{4}$, *il est stable sous la condition CFL*

$$\frac{\Delta t}{\Delta x}\leq\frac{1}{\sqrt{1-4\theta}}.$$

7.2 Les éléments finis

Nous terminons par une introduction à l'étude des éléments finis en utilisant l'équation $-\Delta u=f$ $u\in H^1(\Omega)$ avec condition au bord de Dirichlet sur un ouvert Ω borné.

On vérifie que, si cette équation est vraie au sens des distributions, alors on a

$$\forall\phi\in\in C^\infty(\Omega),<-\Delta u,\phi>=<f,\phi>.$$

On utilise la définition de la dérivée au sens des distributions pour obtenir

$$<\nabla u,\nabla\phi>=<f,\phi>.$$

Comme on suppose $u\in H_0^1(\Omega)$, la forme linéaire

$$\phi\rightarrow<\nabla u,\nabla\phi>$$

est continue sur $C_0^\infty(\Omega)$ pour la norme de $H_0^1(\Omega)$ donc peut se prolonger par densité. Si on suppose $f\in L^2(\Omega)$, le second membre a les mêmes propriétés, donc

$$<\nabla u,\nabla v>=\int f(x)v(x)dx$$

pour $v\in H_0^1(\Omega)$. Cette égalité s'écrit donc

$$\forall v\in H_0^1(\Omega),\int_\Omega u(x)v(x)dx=\int_\Omega f(x)v(x)dx. \tag{7.2.6}$$

On reconnait dans le membre de gauche la dérivée de Fréchet de la fonctionnelle $1-$convexe $\frac{1}{2}\int_\Omega(\nabla u)^2dx$, et l'égalité est l'écriture de la condition d'Euler pour la minimisation sur $H_0^1(\Omega)$ (dont l'espace des directions admissibles est lui-même) de

$$J(u)=\frac{1}{2}\int_\Omega(\nabla u)^2dx-\int_\Omega f(x)u(x)dx.$$

On utilise alors les théorèmes d'approximation, en supposant par exemple que $\Omega=[0,1]\times[0,1]$, pour lequel on construit des sous espaces adaptés de fonctions H_0^1, donnés par $(h=\frac{1}{n})$

$$P_h=\{u(x,y)\in H_0^1([0,1]\times[0,1]),\text{continues, polynômes de degré 1 sur}[ph,(p+1)h]\times[qh,(q+1)h]\}.$$

On détermine alors une base de P_h en définissant la valeur au bord et la valeur des dérivées $\partial_x u$ et $\partial_y u$ sur chacun des pavés du plan. On écrit alors un élément de P_h

sur une base, et on écrit la minimisation de J sur $P_h \subset H_0^1([0,1]\times[0,1])$. Alors on trouve, de l'égalité variationnelle (7.2.6) écrite pour $v_h \in P_h$ et $u_h \in P_h$, un système en dimension finie de la forme $A_h u_h = F_h$, que l'on résout par les méthodes usuelles du cours (en minimisant par exemple $\frac{1}{2}(A_h X, X) - (F_h, X)$), et on essaie d'avoir un résultat en faisant tendre h vers 0.

Par exemple, la base de polynômes sur chaque pavé est $(1, X, Y)$ donc tout polynôme de degré au plus 1 s'écrit

$$a_{p,q} + b_{p,q}(X - ph) + c_{p,q}(Y - qh)$$

Son gradient est approché par $(b_{p,q}, c_{p,q})$ et sa valeur sur $X = ph$ est donnée par $a_{p,q} + c_{p,q}(Y - qh)$, sur $X = (p+1)h$ est donnée par $a_{p,q} + h + c_{p,q}(Y - qh)$, sur $Y = qh$ est $a_{p,q} + b_{p,q}(X - ph)$ et sur $Y = (q+1)h$ par $a_{p,q} + h + b_{p,q}(X - ph)$. On peut alors calculer l'intégrale du produit d'éléments de la base:

$$\begin{array}{c}
\int_0^h \int_0^h 11 dxdy = h^2 \\
\int_0^h \int_0^h 1x dxdy = \frac{h^3}{2} \\
\int_0^h \int_0^h 1y dxdy = \frac{h^3}{2} \\
\int_0^h \int_0^h x^2 dxdy = \frac{h^4}{3} \\
\int_0^h \int_0^h xy dxdy = \frac{h^4}{4} \\
\int_0^h \int_0^h y^2 dxdy = \frac{h^4}{3}
\end{array}$$

ce qui fait que le produit de deux éléments $a + bx + cy$ et $a' + b'x + c'y$ donne

$$h^2[aa' + (ab' + a'b + ac' + a'c)\frac{h}{2} + (bc' + b'c)\frac{h}{3} + (bb' + cc')\frac{h^2}{4}]$$

ainsi la matrice de la forme quadratique associée (en divisant par h^2 pour plus de simplicité) est

$$\begin{pmatrix} 1 & \frac{h}{2} & \frac{h}{2} \\ \frac{h}{2} & \frac{h^2}{4} & \frac{h^2}{3} \\ \frac{h}{2} & \frac{h^2}{3} & \frac{h^2}{4} \end{pmatrix}.$$

Il est clair que c'est une forme quadratique définie positive puisque

$$\int_0^h \int_0^h (a + bx + cy)^2 dxdy = 0 \Rightarrow a = b = c = 0.$$

On utilise donc cette représentation des fonctions de H^1 par des des polynômes de degré 1.

La présentation ainsi faite n'est pas satisfaisante; en effet un carré ou un rectangle a quatre sommets, et un polynôme de degré 1 a trois coefficients. Ainsi on ne pourra pas construire une fonction générale prenant quatre valeurs données en tous les coins $ABCD$; il faut nécessairement que

$$u(A) + u(D) = u(B) + u(C)$$

Si on veut construire une famille qui conduise à toutes les valeurs possibles aux points du carré, il faut considérer les fonctions de la forme

$$u(x, y) = u(0, 0) + bx + cy + dxy$$

qui sont des polynômes de degré 1 dans chacune des variables x, y. Alors on aura

$$u(1,0) = u(0,0) + b, u(0,1) = u(0,0) + c, u(1,1) = u(0,0) + b + c + d$$

donc $b = u(1,0) - u(0,0), c = u(0,1) - u(0,0), d = u(1,1) + u(0,0) - u(0,1) - u(1,0)$, et cette famille permet de construire une solution dont les valeurs données sont les valeurs au coin.

Les valeurs aux sommets s'appellent les **degrés de liberté** d'une fonction de l'espace d'approximation. Dans le pavé $[0,1] \times [0,1]$, on construit les sommets de l'approximation $a_{ij} = (ih, jh)$ et la base de l'espace d'approximation V_h (ϕ_{ij}) des fonctions telles que

$$\phi_{ij}(a_{i'j'}) = \delta_{ii'}\delta_{jj'}$$

qui coincident avec les fonctions décrites ci-dessus sur tous les pavés élementaires de côté h. La fonction ϕ_{ij} est la fonction nulle sur tout pavé dont un coin n'est pas a_{ij} est est construite comme la fonction valant 1 au coin a_{ij} et 0 à tout autre coin pour un pavé ayant a_{ij} comme coin. Toute fonction de V_h s'écrit

$$u = \sum u(a_{ij})\phi_{ij}$$

et il suffit d'évaluer $\int \nabla u \nabla v dx = \sum a_{ij} b_{i'j'} \int \nabla \phi_{ij} \nabla \phi_{i'j'} dx$ pour obtenir la forme quadratique.

Cette présentation fait partie d'un cadre plus général d'approximation, dont on résume les résultats:

Proposition 7.1 *La formulation variationnelle d'un système d'équations aux dérivées partielles avec conditions aux limites prescrites est l'équation d'Euler associée à la minimisation sur un espace de Hilbert H de la fonctionnelle quadratique d'énergie associée au problème $\frac{1}{2}a(u,u) - L(u)$,.*

Elle s'écrit

$$\forall v \in H, a(u,v) = L(v).$$

Une méthode d'approximation s'obtient par le processus suivant: on définit une suite d'espaces vectoriels de dimension finie V_h, associée à un paramètre h tendant vers 0, dont on connait une base simple $\mathcal{B}_h$, ayant les propriétés suivantes

i) pour tout élément v de H on peut construire une suite $v_h \in V_h$ telle que

$$|v - v_h|_H \to 0 \text{ lorsque } h \to 0$$

ii) Le calcul de $a(\phi, \psi)$ pour ϕ et ψ dans $\mathcal{B}_h$ est simple.

Alors si u_h est le minimum de $\frac{1}{2}a(u,u) - L_h(u)$ sur V_h, dans certaines conditions $u_h \to u$.

Chapter 8

Problèmes d'examens

Dans cette partie, nous donnons les sujets d'examens posés les années précédentes. La solution sommaire est donnée en italique à la suite de chaque question.

8.1 Problème des splines: texte du problème de 1999

Dans ce long problème, on cherche à présenter une théorie d'optimisation pour construire les fonctions spline cubiques, qui sont, rappelons le, des polynômes de degré 3 qui se raccordent sur une subdivision. Dans un premier temps, on étudie des problèmes semblables au calcul des variations, en imposant les valeurs en $t = 0$ et en $t = 1$. Dans une deuxième partie, on étudiera une subdivision $t_0 = 0, t_1, ..t_N = 1$ de $[0, 1]$. Les questions marquées d'une * sont soit un peu plus difficiles soit présentent des calculs compliqués. Elles sont à considérer comme des questions facultatives, donnant un bonus lorsqu'elles sont résolues.

PARTIE I; Optimisation en deux points

On introduit $y(t) \in H^2(0,1)$, $v = (v_0, v_1) \in \mathbb{R}^2$. On définit

$$
\begin{aligned}
&J_0(y) = \tfrac{1}{2}\int_0^1 (\tfrac{d^2y}{dt^2})^2(t)dt \\
&J(y,v) = J_0(y) + \tfrac{1}{2}(y(1)-v_1)^2 + \tfrac{1}{2}(y(0)-v_0)^2 \\
&J_\varepsilon(y) = \tfrac{1}{2}\int_0^1 (\tfrac{d^2y}{dt^2})^2(t)dt + \tfrac{\varepsilon}{2}\int_0^1 (\tfrac{dy}{dt})^2(t)dt + \tfrac{\varepsilon}{2}\int_0^1 y^2(t)dt
\end{aligned}
$$

1. On veut résoudre

$$
(A) \left| \begin{array}{l} \inf J_0(y) \\ y(0) = v_0 \\ y(1) = v_1. \end{array} \right.
$$

On note $K = \{y \in H^2(0,1), y(0) = v_0, y(1) = v_1\}$. Montrer que K est fermé.

On peut par exemple utiliser $y(0) = y(\frac{1}{2}) - \int_0^1 y'(s)ds$. On se donne une suite y_n dans K qui converge vers y. Comme H^2 est complet, $y \in H^2$. Le point $\frac{1}{2}$ est intérieur donc comme la norme C^0 est majorée par la norme H^2 sur tout compact inclus dans $]0,1[$, , $y_n(\frac{1}{2})$ converge vers $y(\frac{1}{2})$. On en déduit que $y_n(0)$ tend vers $y(0)$ donc $y(0) = v_0$ et K est fermé. Deuxième solution élégante $y(x) - v_0 - (v_1 - v_0)x$ est dans H_0^2 qui est un espqce complet inclus dans C^1.

1.1. Calculer la dérivée de Gâteaux de J_0 en $y \in H^2(0,1)$ suivant la direction $w \in H^2(0,1)$.

On a la relation $J_0(y+\epsilon w) - J_0(y) = \frac{1}{2}\epsilon^2 J_0(w) + \epsilon \int_0^1 \frac{d^2y}{dt^2}\frac{d^2w}{dt^2}dt$. *Ainsi*

$$(J_0'(y), w) = \int_0^1 \frac{d^2y}{dt^2}\frac{d^2w}{dt^2}dt.$$

1.2. Pour $y \in K$ déterminer le cône des directions admissibles $K(y)$.

Le cône des directions admissibles est $K(y) = H_0^2([0,1])$.

1.3. Ecrire l'équation d'Euler et donner les conditions nécessaires sur y. Calculer la solution générale dans $H^4(0,1)$ de l'équation différentielle obtenue.

L'équation d'Euler est $\forall w \in H^2(0,1), \int_0^1 \frac{d^2y}{dt^2}\frac{d^2w}{dt^2}dt = 0$. *On prend* $w \in C_0^\infty(0,1)$, *ce qui implique que, au sens de* $\mathcal{D}'(0,1)$, $y^{(4)} = 0$. *On ne peut pas aller plus loin car on n'a aucune information sur la continuité de y" pour* $y \in H^2$, *donc on ne peut pas utiliser la formule d'intégration par parties.*

La solution générale de l'équation différentielle dans H^4 *est* $y = a_0 + a_1x + a_2x^2 + a_3x^3$.

1.4. Calculer la solution y_0 de (A) et donner $J_0(y_0)$.

Toute solution au sens des distributions de cette équation différentielle est alors un polynôme de degré 3. En effet, on montre que si z est une distribution de dérivée nulle et ψ *une fonction test, en utilisant une fonction test* θ *donnée d'intégrale égale à 1, la fonction* $\psi(x) - (\int \psi(x)dx)\theta(x)$ *est une fonction à support compact d'intégrale nulle, donc sa primitive* $\phi(x)$ *est une fonction à support compact. Ainsi* $< z, \psi >=< z, \psi - (\int \psi(x)dx)\theta) > + < z, \theta > \int \psi(x)dx =< z, \phi' > + < z, \theta > \int \psi(x)dx =< z, \theta > \int \psi(x)dx$. *On en déduit que z est constante.*

Maintenant, si y est de dérivée quatrième nulle, alors $y^{(3)} = 6a_3$, *donc* $(y - a_3x^3)^{(3)} = 0$. *On reprend le raisonnement de proche en proche pour aboutir à la conclusion. Maintenant, on peut appliquer, pour la solution de l'équation d'Euler, qui est (condition nécessaire) un polynôme de degré 3 donc est dans* H^4, *les formules d'intégration par parties. Alors, utilisant* $w(0) = w(1) = 0$, *on trouve, utilisant des fonctions test telles que* $w'(0) \neq 0$ *et* $w'(1) \neq 0$, *les relations* $y"(0) = y"(1) = 0$. *On trouve donc* $6a_3 + 2a_2 = 0$ *et* $a_2 = 0$, *donc la solution est* $y_0(x) = v_0 + v_1x$, *pour laquelle* $J_0(y_0) = 0$, *donc c'est bien un minimum et il est unique.*

2. On cherche à résoudre

$$(B) \left| \begin{array}{l} \inf J_\varepsilon(y) \\ y(0) = v_0 \\ y(1) = v_1. \end{array} \right.$$

2.1. Identifier α tel que J_ε est $\alpha-$convexe sur $H^2(0,1)$ muni de sa norme usuelle

$$||u|| = (\int_0^1 [(\frac{d^2u}{dt^2})^2 + (\frac{du}{dt})^2 + u^2]dt)^{\frac{1}{2}}.$$

Il suffit de prendre $\alpha = \min(\varepsilon, 1)$.

2.2. Justifier le fait que (B) admet une solution unique. Donner les conditions nécessaires sur la solution y_ε, supposée encore ici dans $H^4(0,1)$. *Montrer que cette solution peut se décomposer sur une base de fonctions de la forme $e^{\lambda t}$ et donner le système vérifié par les coefficients. **Ne Pas le résoudre**.

On applique le théorème 4.1. L'équation d'Euler s'écrit

$$\forall w \in H^2(0,1), \int_0^1 y"w" + \varepsilon(y'w' + yw) = 0.$$

L'équation différentielle ordinaire est alors

$$y^{(4)} - \varepsilon y" + \varepsilon y = 0.$$

Si la solution est dans H^4, par intégrations par parties, on trouve $y"(1) = y"(0) = 0$. On a donc l'équation différentielle ordinaire + quatre conditions aux limites $y(0) = v_0, y(1) = v_1, y"(0) = y"(1) = 0$.

D'autre part, il est facile de voir que l'équation différentielle ordinaire a, dans H^4, les solutions (pour $\epsilon < 4$)

$$a_+ e^{\lambda_1 x + i\lambda_2 x} + a_- e^{\lambda_1 x - i\lambda_2 x} + b_+ e^{-\lambda_1 x + i\lambda_2 x} + b a_- e^{-\lambda_1 x - i\lambda_2 x} = y$$

où $\lambda_1 = (\sqrt{\varepsilon} + \frac{\varepsilon}{2})^{\frac{1}{2}}$, $\lambda_2 = (\sqrt{\varepsilon} - \frac{\varepsilon}{2})^{\frac{1}{2}}$. Les quatre conditions aux limites conduisent à un système sur les coefficients.

2.3. * Montrer que, en utilisant y_0, on a l'inégalité $J_\varepsilon(y_\varepsilon) \leq C\varepsilon$ où C est une constante dépendant de v_0 et de v_1. Peut-on en déduire la limite, lorsque $\varepsilon \to 0$, de y_ε? On pourra utiliser la formule de Taylor avec reste intégral.

On a $J_\varepsilon(y_\varepsilon) \leq J_\varepsilon(y_0)$, ce qui implique $J_\varepsilon(y_\varepsilon) \leq \frac{\varepsilon}{2}[v_0^2 - 2v_0v_1 + v_1^2 + v_0^2 + v_0v_1 + v_1^2] = \varepsilon[v_0^2 - \frac{v_0v_1}{2}]$.

On en déduit $J_0(y_\varepsilon) \leq C\varepsilon$, ce qui démontre, puisque $y"_\varepsilon$ est une suite de L^2, que $y"_\varepsilon$ tend vers 0 dans L^2. On écrit alors

$$y_\varepsilon(x) = v_0 + y'_\varepsilon(0)x + x^2 \int_0^1 (1-t) y"_\varepsilon(tx) dt$$

égalité valable car y_ε est dans H^4, et, de plus, on a la relation

$$y'_\varepsilon(0) = v_1 - v_0 - \int_0^1 (1-t) y"_\varepsilon(t) dt$$

De ces deux égalités, on déduit que $y'_\varepsilon(0)$ converge vers $v_1 - v_0$, en utilisant l'inégalité de Cauchy-Schwartz sur l'intégrale, puis que $y_\varepsilon(x)$ converge vers $v_0 + (v_1 - v_0)x$ en tout point. On montre même, utilisant la formule de Taylor avec reste intégral sur y'_ε, que y_ε tend vers y_0 dans H^2.

3. On veut résoudre

$$(C) \left| \begin{array}{l} \inf J(y,v) \\ y \in H^2(0,1). \end{array} \right.$$

3.1. Montrer que, pour tout $v \in \mathbb{R}^2$, il existe $y(v)(t)$ telle que $y"(v)(t) = 0 \forall t$ et $J(y, v) = J(y - y(v), 0)$.

Comme y" est nulle, $y(v)(x) = a + bx$. Dire que l'égalité demandée est vraie se traduit en

$$J(y - y(v)) = J_0(y - y(v)) + \frac{1}{2}[(y(1) - a - b - v_1)^2 + (y(0) - a - v_0)^2]$$

donc $y(v)(x) = -v_0 - (v_1 - v_0)x$ et l'égalité est vérifiée.

3.2. Démontrer que, pour $(y, z) \in H^2(0, 1)$

$$(J'(y, 0) - J'(z, 0), y - z) = 2J(y - z, 0).$$

On admet que $z \to (J(z, 0))^{\frac{1}{2}}$ est une norme sur $H^2(0, 1)$, équivalente à $||z||$.

En déduire que $J(y, 0)$ est une fonctionnelle $\alpha-$convexe.

L'égalité vient de $(J'(y), w) = \int_0^1 y"w"dt + yw(1) + yw(0)$. Pour montrer l'inégalité de coercivité, on montre que $\int_0^1 y^2 dx$ et $\int_0^1 (y')^2 dx$ sont majorés par $C[(y(0))^2 + (y(1))^2 + \int_0^1 (y")^2 dx]$, ce qui implique que $||y||^2_{H^2} \leq (C + 1)J(y, 0)$.

On démontre par exemple que $y'(0) = y(1) - y(0) - \int_0^1 (1 - t)y"(t)dt$, donc

$$y(x) = y(0) + (y(1) - y(0))x + x^2 \int_0^1 (1 - t)y"(tx)dt - x \int_0^1 (1 - t)y"(t)dt$$

$$y'(x) = y(1) - y(0) + x \int_0^1 y"(xt)dt - \int_0^1 y"(t)dt$$

On en déduit ($(a + b)^2 \leq 2(a^2 + b^2)$)

$$\begin{aligned}(y(x))^2 \leq\ & 2[(y(0) + (y(1) - y(0))x)^2 + (x^2 \textstyle\int_0^1 (1 - t)y"(tx)dt - x \int_0^1 (1 - t)y"(t)dt)^2] \\ & \leq 2[(y(0) + (y(1) - y(0))x)^2 + 2(x^2 \textstyle\int_0^1 (1 - t)y"(tx)dt)^2 + 2x^2(\int_0^1 (1 - t)y"(t)dt)^2] \\ & \leq 2[(y(0) + (y(1) - y(0))x)^2 + 2(\tfrac{x^3}{3} + \tfrac{x^2}{3})||y"||^2_{L^2}]\end{aligned}$$

On en déduit

$$\int_0^1 (y(x))^2 dx \leq 2((y(0))^2 + y(0)y(1) + (y(1))^2) + \frac{7}{9}||y"||^2_{L^2} \leq 3((y(0))^2 + (y(1))^2) + \frac{7}{9}||y"||^2_{L^2}$$

On a un résultat identique pour l'intégrale de y', donc on a la coercivité de J par l'équivalence des normes. On applique alors la proposition 4.3.

3.3. Démontrer que le problème (C) admet une solution unique dans $H^2(0, 1)$. En écrivant la condition d'Euler, déterminer la solution de (C).

Comme il s'agit d'une fonctionnelle $\alpha-$convexe, on a l'existence et l'unicité du minimum. Les équations d'Euler sont

$$\forall w \in H^2, \int_0^1 y"w" + y(0)w(0) + y(1)w(1) = 0.$$

En prenant $w \in C_0^\infty$, on trouve que y est un polynôme. Alors la formule d'intégrations par parties est licite, et on trouve

$$\forall w \in H^2, y"(1)w'(1) - y"(0)w'(0) + (y(0) - y^{(3)}(0))w(0) + (y(1) - y^{(3)}(1))w(1) = 0$$

ce qui donne quatre relations sur les coefficients $6a_3 + 2a_2 = 0, a_2 = 0, a_0 - 6a_3 = 0, a_0+a_1+a_2+a_3-6a_3 = 0$, donc la solution est 0. On aurait pu le trouver directement en rappelant qu'il y a une solution unique, que la valeur de $J(y, 0)$ en $y = 0$ est le minimum, donc le minimum est 0.

4. Résultat général de calcul des variations:

Soit $L(t, u, \dot{u}, \ddot{u})$ une fonction de classe C^2 de toutes ses variables $t \in [0, 1], u \in \mathbb{R}, \dot{u} \in \mathbb{R}, \ddot{u} \in \mathbb{R}$.

On introduit, pour $y \in C^2([0, 1], \mathbb{R})$, $J(y) = \int_0^1 L(s, y(s), y'(s), y"(s))ds$. Déterminer l'équation d'Euler associée à la minimisation de $J(y)$ pour $y(0) = v_0$ et $y(1) = v_1$. Donner les conditions aux limites sur y_0, qui est le point où J est supposée être extremum.

En généralisant l'approche de l'équation d'Euler pour la mécanique, on écrit

$$\forall w \in C^\infty, \int_0^1 [\partial_y L(s, y, y', y")w + \partial_{y'} L(s, y, y', y")w' + \partial_{y"} L(s, y, y', y")w"]ds = 0.$$

Au sens des distributions, on trouve ainsi

$$\partial_y L(t, y_0(t), y_0'(t), y"_0(t)) - \frac{d}{dt}(\partial_{y'} L(t, y_0(t), y_0'(t), y"_0(t))) + \frac{d^2}{dt^2}(\partial_{y"} L(t, y_0(t), y_0'(t), y"_0(t))) = 0.$$

En supposant la solution de classe C^4 par exemple et en réalisant les intégrations par parties, on obtient les quatre relations

$$\partial_{y"} L(1, v_1, y_0'(1), y_0"(1)) = 0, \partial_{y"} L(0, v_0, y_0'(0), y"_0(0)) = 0, y_0(1) = v_1, y_0(0) = v_0.$$

PARTIE II; Optimisation en $N + 1$ points

On donne $(v_0, ..., v_N) \in \mathbb{R}^{N+1}$, et $0 = t_0 < t_1 < ... < t_{N-1} < t_N = 1$. On introduit

$$S(y, v) = \frac{1}{2}\int_0^1 (\frac{d^2y}{dt^2})^2 dt + \frac{1}{2}\sum_{j=0}^{j=N} (y(t_j) - v_j)^2.$$

On cherche les solutions de

$$(D) \left| \begin{array}{l} \inf S(y, v) \\ y \in H^2(0, 1) \end{array} \right. \quad (E) \left| \begin{array}{l} \inf J_0(y) \\ y \in H^2(0, 1), y(t_0) = v_0, ..., y(t_j) = v_j ... \end{array} \right.$$

5. Spline d'ajustement.

5.1. On suppose $N \geq 2$. Déterminer les relations sur $t_1, ..., t_{N_1}, v_1, ..., v_{N_1}$ en fonction de v_0 et de v_N de sorte que $S(y, v) = 0$ admette une solution y.

Si $S(y, v) = 0$, alors y est un polynôme de degré 1, entièrement déterminé par $y(t_0) = v_0$ et $y(t_N) = v_N$: $y(t) = v_0 + \frac{v_N - v_0}{t_N - t_0}(t - t_0)$. Alors les conditions de compatiblité sont

$$(v_j - v_0)(t_N - t_0) = (v_N - v_0)(t_j - t_0), \forall j.$$

5.2. Montrer que, pour $N \geq 1$, la fonctionnelle $y \to S(y, v)$ est une fonctionnelle $\alpha-$convexe sur $H^2(0, 1)$. On pourra remarquer que

$$S(y, v) = J(y, v_0, v_N) + \frac{1}{2} \sum_{i=1}^{i=N-1} (y(t_i) - v_i)^2$$

la somme étant vide si $N = 1$. On utilisera alors les questions **3.1.**, **3.2.**.

On sait alors que $J(y, v_0, v_N) = J(y - y(v_0, v_N), 0) \geq \alpha ||y - y(v_0, v_N)||^2_{H^2}$, ce qui implique la coercivité de S dans H^2. L'$\alpha-$convexité s'en déduit.

5.3. En déduire que (D) admet une solution unique $\tilde{y}$, pour laquelle on donnera les conditions nécessaires d'optimalité. On remarquera, pour obtenir ces équations, qu'il n'est pas licite de supposer $\tilde{y} \in H^4(0, 1)$, mais on démontrera en utilisant des fonctions test adéquates que l'on pourra prendre $\tilde{y} \in H^4(]t_i, t_{i+1}[)$ pour $i \leq N - 1$.

Le fait qu'il y a une solution unique provient de l'$\alpha-$convexité. La condition d'Euler s'écrit

$$\int_0^1 y" w" dt + \sum_j w(t_j)(y(t_j) - v_j) = 0 \forall w \in H^2.$$

On en déduit, prenant $w \in C_0^\infty(]t_i, t_{i+1}[)$, que $y^{(4)}$ est nulle dans $\mathcal{D}'(]t_i, t_{i+1}[)$, ainsi $y \in H^4(]t_j, t_{j+1}[)$.

5.4. Démontrer que $\tilde{y}$ est une fonction spline cubique de classe C^2 sur $[0, 1]$. On l'appelle spline d'ajustement.

Comme y est dans H^2, y est de classe C^1 sur $[0, 1]$ par inclusion d'espaces de Sobolev. Ceci se démontre car $y'(x) - y'(z) = \int_x^z y"(t)dt$ donc $|y'(x) - y'(z)|| \leq (|x - z|)^{\frac{1}{2}} ||y||_{H^2}$. Cette simple inégalité ne suffit pas. On montre d'abord que, pour f de classe C^2, on a l'inégalité $|f'(x) - f'(z)| \leq (|x - z|)^{\frac{1}{2}} ||f"||$, ainsi on en déduit $|f'(x)| \leq |f'(z)| + (|x - z|)^{\frac{1}{2}} ||f"||_2$, donc en intégrant en z sur $[0, 1]$ on trouve $|f'(x)| \leq ||f'||_2 + \frac{4}{3}||f"||_2$. On voit donc que si y_n est une suite de fonctions de classe C^2 convergeant vers y au sens H^2, alors $|y'_n(x) - y'_m(x)|$ vérifie le critère de Cauchy, donc la suite $y'_n(x)$ converge pour tout x, uniformément en x, vers une fonction continue notée $g(x)$. On montre ainsi que, de même, la suite $y_n(x)$ converge uniformément. Soit y la limite uniforme de y_n. Alors de l'égalité $y_n(x) - y_n(a) = \int_a^x y'_n(s)ds$ on déduit que $y(x) - y(a) = \int_a^x g(t)dt$, donc $y' = g$.

De plus, grâce à l'équation d'Euler, en effectuant l'intégration par parties sur $]t_i, t_{i+1}[$ et sur $]t_{i-1}, t_i[$, on trouve

$$\int_{t_{i-1}}^{t_{i+1}} y" w" dt = y"(t_{i+1} - 0)w'(t_{i+1}) + w'(t_i)(y"(t_i - 0) - y"(t_i + 0)) - w'(t_{i-1})y"(t_{i-1} - 0)$$

en ayant utilisé $w \in H^2$ donc w' continue, le -0 ou $+0$ étant une notation indiquant la limite de la dérivée seconde du polynôme de degré 3 représentant y dans chaque intervalle, pris dans l'intervalle considérée. Dire que l'équation d'Euler est vraie pour toute fonction w dans H^2 implique que $y"(t_i-0) = y"(t_i+0)$ pour tout i, $1 \leq i \leq N-1$ et $y"(0) = y"(1) = 0$. On en conclut que y" est affine par morceaux admettant la même limite à droite et à gauche en chaque point intérieur; elle est donc continue, donc y est de classe C^2.

Attention: sa valeur en un point t_j n'est **pas** *v_j. En effet, ce qui provient de l'équation d'Euler est la relation $y(t_i) = v_i + (y'''(t_i - 0) - y'''(t_i + 0))$.*

5.5. Que se passe-t-il si on étudie le problème

$$(D') \left| \begin{array}{l} \inf \frac{1}{2}\int_0^1 (\frac{d^2y}{dt^2})^2 dt + \sum_{j=0}^{j=N} (y(t_j) - v_j)^2 \\ y \in H^2(0,1) \end{array} \right.$$

Réponse: on change la spline d'ajustement car on change la relation en $y(t_i) = v_i + \frac{1}{2}(y'''(t_i - 0) - y'''(t_i + 0))$.

6. Spline d'interpolation.

6.1 Montrer que (E) admet une solution, lorsque $N \geq 1$. Donner les conditions d'optimalité. On note $\bar{y}$ une solution de l'équation d'Euler.

Attention: on ne peut pas dire que J_0 est infini à l'infini dans H^2 car toute fonction de la forme $y_{a,b}(x) = ax + b$ vérifie $J_0(y) = 0$ et pourtant $||y||^2_{H^2} = a^2 + a + 2b$, et il suffit de prendre $b = 0$ et a infini pour avoir y tend vers l'infini. On trouve aussi que pour tout y, $J_0(y + y_{a,b}) = J_0(y)$.

Lorsque $N \geq 1$, on considère $z(x) = y(x) - v_0 - (v_1 - v_0)x$. Lorsque y est dans l'espace des contraintes, cette fonction est dans H^2_0. Elle vérifie les contraintes $z(t_i) = v_i - v_0 - (v_1 - v_0)t_i$. On voit que

$$z(t) = \int_0^x (x-t)z"(t)dt - x\int_0^1 (1_t)z"(t)dt, z'(t) = \int_0^x tz"(t)dt - \int_x^1 (1-t)z"(t)dt$$

ce qui donne les majorations $|z(x)| \leq \frac{1}{\sqrt{3}}||z"||_{L^2} x(1-x)(\sqrt{x} + \sqrt{(1-x)})$ et $|z'(x)| \leq \frac{1}{\sqrt{3}}||z"||_{L^2}(x^{\frac{3}{2}} + (1-x)^{\frac{3}{2}})$. Ainsi, intégrant sur $(0,1)$ le carré de ces fonctions pour trouver la norme H^2, on trouve

$$||z||_{H^2} \leq (\frac{1}{45} + \frac{2}{3} + 1)^{\frac{1}{2}}||z"||_{L^2}.$$

6.2. En supposant $\bar{y} \in H^4(]t_i, t_{i+1}[)$, trouver les équations différentielles vérifiées par $\bar{y}$. Donner les conditions aux limites aux points t_i.

Ainsi, soit $K_0 = \{y, y(0) = v_0, y(1) = v_1\}$. On a l'inégalité, pour tout $y \in K_0$, $\frac{\sqrt{61}}{6\sqrt{5}}||y - y_0||^2_{H^2} \leq J_0(y)$, ce qui permet d'en déduire l'existence et l'unicité d'un minimum, puisque l'on a une fonctionnelle convexe sur un convexe. Ensuite, les équations sur $\bar{y}$ sont bien $\bar{y}^{(4)} = 0$ sur $]t_I, t_{i+1}[$. Comme l'équation d'Euler est $\int_0^1 y"w"dt = 0$ pour $w \in H^2$, $w(t_i) = 0 \forall i$, on trouve que $\bar{y}"(0) = 0, \bar{y}"(1) = 0$ et

$\bar{y}"(t_i+0) - \bar{y}"(t_i-0) = 0$ *puisque l'on peut prendre une fonction* w *quelconque telle que* $w(t_{i_0}) = 0$, $w'(t_{i_0}) = 1$, *et* w *à support compact dans* $]t_{i_0-1}, t_{i_0+1}[$ *pour* $i_0 \neq 0, N$. *Ainsi les conditions aux limites sont* $\bar{y}(t_i) = v_i$, $\bar{y}"$ *continue. On a répondu à la question suivante.*

6.3. Démontrer que la solution est unique* et que c'est une spline cubique de classe C^2.

6.4. Ecrire les conditions d'optimalité avec multiplicateurs de Lagrange, et retrouver les résultats précédents.

On trouve que

$$\begin{aligned} J_0'(\bar{y}) = \quad & \bar{y}^{(4)} - y"(1)\delta_1' + y"(0)\delta_0' + \textstyle\sum_{i=1}^{N-1}(y"(t_i+0) - y"(t_i-0))\delta_{t_i}' \\ & + \textstyle\sum_{i=1}^{N-1}(y'''(t_i+0) - y'''(t_i-0))\delta_{t_i} - y'''(1)\delta_1 + y'''(0)\delta_0 \end{aligned}$$

Il existe donc $N+1$ *valeurs* λ_i *telles que*

$$\begin{aligned} & \bar{y}^{(4)} - y"(1)\delta_1' + y"(0)\delta_0' + \textstyle\sum_{i=1}^{N-1}(y"(t_i+0) - y"(t_i-0))\delta_{t_i}' \\ & + \textstyle\sum_{i=1}^{N-1}(y'''(t_i+0) - y'''(t_i-0))\delta_{t_i} - y'''(1)\delta_1 + y'''(0)\delta_0 + \textstyle\sum_i \lambda_i \delta_{t_i} = 0 \end{aligned}$$

ce qui redonne les conditions d'optimalité.

6.5. Comparer $S(\tilde{y}, v)$ et $J_0(\bar{y})$. En déduire une comparaison des deux types d'approximation.

On voit que $S(\bar{y}, v) = J_0(\bar{y})$, *donc, comme le minimum de* S *est atteint en* $y = \tilde{y}$, *on a* $S(\tilde{y}, v) \leq J_0(\bar{y})$. *On se place dans le cas* $N \geq 1$. *Alors, si* $S(\tilde{y}, v) = J_0(\tilde{y})$, *on en déduit,* $\forall y, S(y, v) \geq J_0(\tilde{y})$ *et donc* $\tilde{y} = \bar{y}$. *Donc si* $\tilde{y} \neq \bar{y}$, *alors* $S(\tilde{y}, v) < J_0(\bar{y})$.

6.6. *Dans le cas $N = 2$, $t_1 = \frac{1}{2}$, vérifier que, pour $t \leq \frac{1}{2}$

$$\tilde{y}(t) = v_0 - \frac{1}{6+\frac{1}{24}}(v_0+v_2-2v_1) + t[v_1 - v_0 - \frac{1}{8}\frac{1}{6+\frac{1}{24}}(v_0+v_2-2v_1)] + \frac{t^3}{3}\frac{1}{6+\frac{1}{24}}(v_0+v_2-2v_1)$$

et pour $t \geq \frac{1}{2}$

$$\begin{aligned} \tilde{y}(t) \quad & = v_0 - (2+\tfrac{1}{2})\tfrac{1}{6+\frac{1}{24}}(v_0+v_2-2v_1) \\ & + t[v_1 - v_0 - \tfrac{1}{8}\tfrac{1}{6+\frac{1}{24}}(v_0+v_2-2v_1)] + \tfrac{1}{6+\frac{1}{24}}(v_0+v_2-2v_1)\tfrac{t^2}{2} \\ & - \tfrac{t^3}{3}\tfrac{1}{6+\frac{1}{24}}(v_0+v_2-2v_1). \end{aligned}$$

De même, vérifier que, pour $t \leq \frac{1}{2}$

$$\bar{y}(t) = v_0 + t[v_2 - v_0 - \frac{3}{2}(v_2+v_0-2v_1)] + 2t^3(v_0+v_2-2v_1)$$

et pour $t \geq \frac{1}{2}$ que

$$\begin{aligned} \bar{y}(t) \quad & = v_0 + \tfrac{1}{2}(v_0+v_2-2v_1) + (v_1 - v_0 - (4+\tfrac{1}{2})(v_0+v_2-2v_1))t \\ & + 6t^2(v_0+v_2-2v_1) - 2t^3(v_0+v_2-2v_1). \end{aligned}$$

8.2 Texte du problème 2000

Dans ce sujet, on considère le système suivant d'équations aux dérivées partielles

$$\begin{cases} -\Delta y + y^3 = u \text{ dans } \Omega \\ y = 0 \text{ sur } \partial\Omega \end{cases} \tag{8.2.1}$$

où Ω est un ouvert borné régulier de $\mathbb{R}^3$.

On note $||y||_{H_0^1(\Omega)} = (\int_\Omega |\nabla y(x)|^2 dx)^{\frac{1}{2}}$ et $||y||_{H^1(\Omega)} = (\int_\Omega |\nabla y(x)|^2 dx + \int_\Omega |y(x)|^2 dx)^{\frac{1}{2}}$.

On suppose que $u \in L^2(\Omega)$.

On rappelle que, pour tout p entier inférieur à 6, il existe une constante c_p telle que

$$||y||_{L^p(\Omega)} \leq c_p ||y||_{H^1(\Omega)}$$

et que on a l'inégalité de Poincaré pour $y \in H_0^1(\Omega)$:

$$||y||_{H^1(\Omega)} \leq C ||y||_{H_0^1(\Omega)}.$$

Les questions marquées d'une * sont facultatives car plus difficiles, elles donnent droit à un bonus.

0) **Généralités et fonctions homogènes**

On suppose que $J(y)$ est une application d'un espace de Hilbert V dans $\mathbb{R}$, telle que

$$J(y) = J_2(y) + J_1(y) + J_\lambda(y)$$

où λ est un réel positif et où on a, pour tout $p \in 1, 2, \lambda$, l'égalité d'homogénéité:

$$J_p(ky) = k^p J(y).$$

On suppose que J est de classe C^2 et on considère sa dérivée J' et sa dérivée seconde J". **Montrer** les égalités:

$$\forall y \in V, (J'_p(y), y) = pJ_p(y), (J"_p(y), y, y) = p(p-1)J_p(y).$$

On constate que $J_p((k+\epsilon)y) = J_p(ky + \epsilon y) = J_p(ky) + \epsilon(J'_p(ky), y) + o(\epsilon)$. D'autre part, $J_p((k+\epsilon)y) = (k+\epsilon)^p J_p(y) = k^p J_p(y) + pk^{p-1}\epsilon J_p(y) + o(\epsilon)$, donc finalement $(J'_p(ky), y) = pk^{p-1}J_p(y)$. Il suffit de prendre $k = 1$ pour obtenir la première égalité.

De plus, $J_p(k(y+w)) = J_p(ky + kw) = J_p(ky) + k(J'_p(ky), w) + o(w)$, donc $(J'_p(ky), w) = k^{p-1}(J'_p(y), w)$. De cette dernière égalité, on déduit que J'_p est homogène de degré $p-1$ donc $(J"_p(y)y, w) = (p-1)(J'_p(y), w)$. Il suffit de prendre $w = y$ et d'appliquer le résultat précédent.

1) a) **Montrer** que, si $y \in H_0^1(\Omega)$ est solution de (8.2.1) au sens des distributions, alors on a

$$\forall \phi \in C_0^\infty(\Omega), L(y, \phi) = \int_\Omega \nabla y(x) \nabla \phi(x) dx + \int_\Omega y^3 \phi(x) dx = \int_\Omega u(x)\phi(x) dx. \tag{8.2.2}$$

Ceci provient du calcul de la formulation variationnelle associée à l'équation. Dans tous les cas, on multiplie par une fonction ϕ et on utilise la formule d'intégration par parties $\int_\Omega (-\Delta y \phi) dx = \int_\Omega \nabla y \nabla \phi - \int_{\partial\Omega} \partial_n y \phi d\sigma$. Lorsque $\phi \in C_0^\infty(\Omega)$, le terme de bord vaut 0, et on obtient l'égalité ci-dessus.

b) **Démontrer** que cette égalité est vraie pour $\phi \in C^\infty(\mathbb{R}^3)$, ainsi que pour $\phi \in H_0^1(\Omega)$.

Lorsque ϕ est dans $H_0^1(\Omega)$, c'est la limite d'une suite de fonctions de $C_0^\infty(\Omega)$, notée ϕ_n et on a $L(y,\phi_n) = \int_\Omega u\phi_n dx$. La limite lorsque ϕ_n tend vers ϕ dans $H_0^1(\Omega)$ de $\int_\Omega u\phi_n$ est $\int_\Omega u\phi dx$ car c'est une limite dans L^2, et de même dans $H^1(\Omega)$. Un détail cependant: comme $y \in H_0^1(\Omega)$, on a l'inégalité

$$|\int_\Omega y^3(\phi_n - \phi_m)dx| \leq (\int_\Omega y^6(x)dx)^{\frac{1}{2}}||\phi_n - \phi_m||_{L^2}.$$

Cette inégalité assure la convergence de ce terme car y est dans L^6.

Pour ϕ dans $C^\infty(\mathbb{R}^3)$, l'égalité est fausse (contrairement à l'énoncé) car $\partial_n y$ n'est pas nul.

c) **Montrer** que, si $y \in H_0^1(\Omega)$ est solution de (8.2.2) pour tout $\phi \in H_0^1(\Omega)$, alors y est solution de (8.2.1).

On a, au sens des distributions, $\int_\Omega \nabla y \nabla\phi =< \Delta y, \phi >$. Pour le démontrer, on peut par exemple prendre une suite de fonctions y_n de $C_0^\infty(\Omega)$ qui converge vers y. Alors, comme $\phi|_{\partial\Omega} = 0$, on a $\int_\Omega \nabla y_n \nabla\phi$ tend vers $\int_\Omega \nabla y \nabla\phi$, et donc l'égalité est vraie. Ainsi on trouve

$$< -\Delta y + y^3, \phi >= \int_\Omega u\phi dx, \forall\phi \in H_0^1(\Omega).$$

On en déduit $-\Delta y + y^3 = u$. Comme $y \in H_0^1(\Omega)$, $y = 0$ sur le bord.

2) En utilisant la question 0), **trouver*** p et $J_p(y)$ fonction de classe C^2 sur $H_0^1(\Omega)$ de sorte que $(J_p'(y), z) = \int_\Omega (y(x))^3 z(x)dx$. *On vérifie que $(J_p'(y), y) = pJ_p(y)$, ce qui nous donnerait $pJ_p(y) = \int_\Omega (y(x))^4 dx$. On en déduit $p = 4$ car $\int_\Omega (ky(x))^4 dx = k^4 \int_\Omega (y(x))^4 dx$, donc $J_4(y) = \frac{1}{4}\int_\Omega (y(x))^4 dx$.*

3) On introduit la fonctionnelle

$$\Phi(y) = \frac{1}{2}\int_\Omega |\nabla y(x)|^2 dx - \int_\Omega y(x)u(x)dx + \frac{1}{4}\int_\Omega (y(x))^4 dx.$$

a) **Montrer** que Φ est une application $\alpha-$convexe continue de $H_0^1(\Omega)$ dans $\mathbb{R}$, et qu'elle possède un minimum unique, noté $y(u)$.

On calcule $(\Phi'(y), v) = \int_\Omega [\nabla y \nabla v + y^3 v]dx$. On trouve alors $(\Phi'(y) - \Phi'(z), y - z) = \int_\Omega [(\nabla y - \nabla z).(\nabla y - \nabla z) + (y^3 - z^3)(y - z)]dx = \int_\Omega [|\nabla(y - z)|^2 + (y - z)^2(y^2 + yz + z^2)]dx$. On trouve alors, sachant que la norme sur H_0^1 est $\int(\nabla\phi)^2$, la relation $(\Phi'(y) - \Phi'(z), y - z) \geq \int_\Omega (\nabla y - \nabla z)^2 dx = ||y - z||^2_{H_0^1}$, donc l'application est $\alpha-$convexe continue de $H_0^1(\Omega)$ dans $\mathbb{R}$ (la continuité est une conséquence de l'inégalité $\int y^4 \leq (\int y^6)^{\frac{1}{2}}(\int y^2)^{\frac{1}{2}} \leq (c_6)^3||y||^4_{H^1}$). On utilise l'inégalité de Poincaré, d'où la continuité du terme $\int uydx$. L'existence du minimum et l'unicité est alors une conséquence d'un théorème du cours.

b) **Donner l'équation d'Euler** associée à $y(u)$. En effectuant un choix adéquat de ϕ dans l'égalité $L(y(u), \phi) = 0$, démontrer qu'il existe une constante c_1, telle que

$$||y(u)||_{H_0^1(\Omega)} \leq c_1||u||_{L^2(\Omega)}.$$

L'équation d'Euler est alors $\forall w, \int_\Omega (\nabla y(u)\nabla w + (y(u))^3 w - uw)dx = 0$. On prend $w = y(u)$ donc $\int_\Omega (\nabla y(u))^2 + \int (y(u))^4 = \int uy(u)dx$. On en déduit, utilisant l'inégalité de Cauchy-Schwartz, et $\int (y(u))^4 dx \geq 0$:

$$||y(u)||^2_{H^1_0(\Omega)} \leq (\int_\Omega u^2 dx)^{\frac{1}{2}} (\int_\Omega (y(u))^2 dx)^{\frac{1}{2}} \leq (\int_\Omega u^2 dx)^{\frac{1}{2}} \sqrt{C} ||y(u)||_{H^1_0(\Omega)},$$

d'où on déduit l'inégalité

$$||y(u)||_{H^1_0(\Omega)} \leq \sqrt{C} ||u||_{L^2(\Omega)}.$$

c) **Calculer**, pour tout y les expressions

$$(\Phi'(y), y), (\Phi"(y), y, y).$$

On applique le résultat de la question 0). Alors $(\Phi'(y), y) = \int_\Omega ((\nabla y)^2 + y^4) dx$, $(\Phi"(y)y, y) = \int_\Omega ((\nabla y)^2 + 3y^4) dx$.

4) **Montrer*** que la solution unique de

$$\mathrm{Inf}_{y,w} (\frac{1}{2} \int_\Omega (w(x) + (y(x))^3)^2 dx)$$

sous la contrainte $-\Delta y = u + w, y \in H^1_0(\Omega), w \in L^2(\Omega)$ est le couple $(y(u), -(y(u))^3)$. *On remarque que ce couple vérifie* $\frac{1}{2} \int_\Omega (w + y^3)^2 dx = 0$. *On a donc l'existence d'un minimum. D'autre part, si on a un autre point de minimum, alors* $w + y^3$, *qui est dans* L^2, *est nul donc* $w = -y^3$ *et la contrainte s'écrit* $-\Delta y + y^3 = u$, *dont la solution unique est* $y(u)$.

On note que l'on s'est donc ramené à la résolution d'un laplacien et ensuite d'une minimisation sur w.

5) On considère u et v dans $L^2(\Omega)$. On désigne par $y(u)$ et $y(v)$ les deux solutions précédentes associées. On note

$$m(x) = (y(u)(x))^2 + y(u)(x) y(v)(x) + (y(v)(x))^2$$

et $z(x) = y(u)(x) - y(v)(x)$. **Montrer** que $m(x) \geq 0$.

Montrer que z est solution H^1_0 de l'équation

$$-\Delta z(x) + m(x) z(x) = u(x) - v(x).$$

En multipliant cette équation par $z_+(x) = \max(0, z(x))$ et en intégrant sur Ω, (on admettra l'égalité $\int_\Omega \nabla z(x) \nabla z_+(x) dx = \int_\Omega |\nabla z_+|^2 dx$), **montrer*** que si $v - u \leq 0$ sur Ω, alors $z(x) \leq 0$.

On intègre l'égalité $(-\Delta z(x) + m(x) z(x)) z_+(x) = (u(x) - v(x)) z_+(x)$. *On vérifie que* $\int \nabla z_+|^2 + \int m(x) z(x) z_+(x) dx = \int_\Omega (u - v) z_+ dx$. *D'autre part,* $\int m(x) z_+ z dx = \int m(x) (z_+)^2 dx$ *et* $m \geq 0$ *donc nécessairement de* $\int (u - v) z_+ dx \leq 0$ *on déduit* $\int m z_+^2 = 0$ *et* $\int (\nabla z_+)^2 dx = 0$ *donc* $z_+ = 0$. *On en déduit que* $\max(z, 0) = 0$ *donc* $z \leq 0$.

8.3 Texte du problème 2000-2001

Avertissement

Cet examen se compose de deux parties totalement indépendantes, et n'est pas fait pour être fini. Une première partie concerne les conditions aux limites et une formulation lagrangienne de l'équation des ondes pour des cordes vibrantes. Une

deuxième partie étudie un système électrique et introduit des contraintes de type isopérimétrique.

Toute égalité énoncée dans le texte peut être utilisée même si elle n'a pas été établie.

8.4 Partie I

1) Résultat général

On considère une fonction de $C^2(\mathbb{R}^4)$ dans $\mathbb{R}$, notée $L(p_1, p_2, q_1, q_2)$. On notera parfois p ou $\vec{p}$ le vecteur de composantes (p_1, p_2) (de même pour q).

On introduit une fonction $\vec{u}(x,t) = (u_1(x,t), u_2(x,t))$ une fonction de classe $C^2(\mathbb{R}^2)$ dans $\mathbb{R}^2$. On la notera aussi u (omettant le vecteur). On veut minimiser

$$I(u) = \int_0^T \int_0^a L(\partial_t \vec{u}, \partial_x \vec{u}) dx dt$$

On note que $p_1 = \partial_t u_1, p_2 = \partial_t u_2$...

a) Etablir les équations d'Euler en tout point $(x,t) \in]0,a[\times]0,T[$ pour une solution u_0 de

$$\inf I(u)$$

(on ne cherche pas à préciser les conditions aux limites sur le bord du rectangle Ω dans $\mathbb{R}^2$).

On considère $w \in C_0^\infty([0,a] \times [0,T])$. Alors on trouve

$$I(\vec{u} + \epsilon \vec{w}) - I(\vec{u}) = \int_0^T \int_0^a (L(\partial_t \vec{u} + \epsilon \partial_t \vec{w}, \partial_x \vec{u} + \epsilon \vec{w}) - L(\partial_t \vec{u}, \partial_x \vec{u})) dx dt$$

En effectuant un développement limité en $\epsilon \to 0$, on trouve que la limite du taux d'accroissement est

$$\int_0^T \int_0^a [\partial_p L(\partial_t \vec{u}, \partial_x \vec{u}).\partial_t \vec{w} + \partial_q L(\partial_t \vec{u}, \partial_x \vec{u}).\partial_x \vec{w}] dt dx.$$

En effectuant une intégration par parties en t pour le premier terme, et une intégration par parties en x pour le deuxième terme, on trouve

$$(I'(u), w) = -\int_0^T \int_0^a [w_1 [\frac{d}{dt}(\partial_{p_1} L) + \frac{d}{dx}(\partial_{q_1} L)] + w_2 [\frac{d}{dt}(\partial_{p_2} L) + \frac{d}{dx}(\partial_{q_2} L)]] dt dx$$

et la condition d'Euler conduit aux deux équations

$$\begin{cases} \frac{d}{dt}(\partial_{p_1} L) + \frac{d}{dx}(\partial_{q_1} L) = 0 \\ \frac{d}{dt}(\partial_{p_2} L) + \frac{d}{dx}(\partial_{q_2} L) = 0. \end{cases}$$

b) Soit u_0 une solution des équations d'Euler précédentes. Montrer que

$$\frac{d}{dt}(\int_0^a [L(\partial_t u_0, \partial_x u_0) - \partial_t u_0 \partial_p L(\partial_t u_0, \partial_x u_0)](y,t) dy)$$
$$=$$
$$\partial_t u_0 \partial_q L(\partial_t u_0, \partial_x u_0)(a,t) - \partial_t u_0 \partial_q L(\partial_t u_0, \partial_x u_0)(0,t).$$

(on pourra pour cela dériver la fonction composée $\partial_t(L(\partial_t u_0, \partial_x u_0))$ et une autre expression)

On dérive la fonction composée. On trouve $\partial_t(L(\partial_t \vec{u}_0, \partial_x \vec{u}_0) = \partial^2_{t^2}\vec{u}_0 \cdot \partial_p L + \partial^2_{tx}\vec{u}_0 \partial_q L$.

En utilisant l'équation d'Euler, on trouve

$$\begin{gathered}\frac{d}{dt}(\int_0^a [L(\partial_t \vec{u}_0 \partial_x \vec{u}_0) - \partial_t \vec{u}_0 \cdot \partial_p L(\partial_t \vec{u}_0, \partial_x \vec{u}_0)](y,t)dy) \\ = \\ \int_0^a [\partial^2_{t^2}\vec{u}_0 \cdot \partial_p L + \partial^2_{tx}\vec{u}_0 \partial_q L - \partial^2_{t^2}\vec{u}_0 \partial_p L - \partial_t \vec{u}_0 \frac{d}{dt}(\partial_p L(\partial_t \vec{u}_0, \partial_x \vec{u}_0))]dy \\ = \\ \int_0^a [\partial^2_{tx}\vec{u}_0 \partial_q L + \partial_t \vec{u}_0 \frac{d}{dx}(\partial_q L(\partial_t \vec{u}_0, \partial_x \vec{u}_0))]dy\end{gathered}$$

On reconnait dans le crochet la dérivée par rapport à x de la fonction $\partial_t \vec{u}_0 \partial_q L$, ce qui donne le résultat demandé en intégrant en y.

c) On considère les trois problèmes

$$(P_1)\ \begin{array}{l}\inf I(u) \\ u(x,0) = u^0(x) \\ u(x,T) = u^f(x)\end{array} \qquad (P_2)\ \begin{array}{l}\inf I(u) \\ u(x,0) = u^0(x) \\ u(x,T) = u^f(x) \\ u(0,t) = 0\end{array} \qquad (P_3)\ \begin{array}{l}\inf I(u) \\ u(x,0) = u^0(x) \\ u(x,T) = u^f(x) \\ u(0,t) = 0 \\ u(a,t) = 0\end{array} \quad .$$

Ecrire les équations d'Euler et les conditions aux limites en $x = 0$ et $x = a$ pour chacun de ces problèmes.

Pour cela, l'équation d'Euler est celle obtenue ci-dessus et on ne se préoccupera que des conditions aux limites. Pour le problème (P_1)*, on trouve* $w(x,0) = w(x,T) = 0$*, ainsi quand on reprend l'égalité ci-dessus ayant abouti à* $(I'(u), w)$*, on trouve*

$$(I'(u), w) = \int_0^T \partial_q L \cdot \vec{w}(a,t)dt - \int_0^T \partial_q L \cdot \vec{w}(0,t)dt.$$

Comme cette quantité doit être nulle pour tout $\vec{w}$*, on en déduit* $\partial_q L(\partial_t \vec{u}_0(a,t), \partial_x \vec{u}_0(a,t)) = 0$ *et* $\partial_q L(\partial_t \vec{u}_0(0,t), \partial_x \vec{u}_0(0,t)) = 0$*. Ce sont les deux conditions aux limites que l'on doit ajouter à* $\vec{u}_0(x,0) = \vec{u}^0(x)$ *et* $\vec{u}_0(x,T) = \vec{u}^f(x)$*.*

Pour le problème (P_2) *on a la condition aux limites supplémentaire* $\partial_q L(\partial_t \vec{u}_0(a,t), \partial_x \vec{u}_0(a,t)) = 0$ *par l'équation d'Euler.*

Pour le problème (P_3)*, il n'y a aucune condition supplémentaire.*

Montrer, pour la solution u_0^j de P_j, pour tout j, la relation

$$\int_0^a [L(\partial_t u_0^j, \partial_x u_0^j) - \partial_t u_0^j \partial_p L(\partial_t u_0^j, \partial_x u_0^j)](y,t)dy = C_j$$

où C_j est une constante indépendante du temps.

On remplace les relations supplémentaires obtenues dans le second membre du b). Alors on trouve, pour le problème (P_1)*, que ce second membre est nul car les deux termes* $\partial_q L$ *sont nuls en* $x = 0$ *et* $x = a$*. Pour le problème* (P_2)*, on sait que le terme* $\partial_q L$ *est nul en* a *et comme* $\vec{u}(0,t) = 0$ *on trouve que* $\partial_t \vec{u}_0(0,t) = 0$*. Enfin, pour le problème* (P_3)*, il vient, d'après* $\vec{u}(0,t) = \vec{u}(a,t) = 0$ *que le terme* $\partial_t \vec{u}_0(0,t)$ *et le terme* $\partial_t \vec{u}_0(a,t)$ *sont nuls, d'où le résultat.*

2) Application à l'équation des ondes dans les cordes vibrantes

a) Etablissement de l'équation

On étudie les **petits** déplacements d'une corde autour de sa position d'équilibre (OA), O(0,0,0), A(a, 0,0).

La position d'un point de la courbe est $(x, u_1(x,t), u_2(x,t)) = (x, u(x,t))$.

La densité de la corde est ρ_0, et cette corde est soumise à la tension $\vec{T_0}$, de module constant T_0, dirigée suivant le vecteur tangent unitaire τ.

Ecrire le bilan des forces et la relation fondamentale de la dynamique pour un segment $[x, x+\Delta x]$ en négligeant tous les termes d'ordre au moins 2 en u. En faisant tendre Δx vers 0, en déduire l'équation

$$\rho_0 \frac{\partial^2 \vec{u}}{\partial t^2} = T_0 \frac{\partial^2 \vec{u}}{\partial x^2}.$$

laissé en exercice (voir méthodes mathématiques pour la physique, de L. Schwartz)

b) Etablir la relation, pour $\vec{u}_0$ solution de l'équation précédente

$$\frac{dE}{dt} = \frac{d}{dt}\int_0^a \frac{1}{2}(\rho_0(\frac{\partial \vec{u}}{\partial t})^2 + T_0(\frac{\partial \vec{u}}{\partial x})^2)(y,t)dy = \partial_t \vec{u} \partial_x \vec{u}(a,t) - \partial_t \vec{u} \partial_x \vec{u}(0,t).$$

il suffit de multiplier par $\partial_t \vec{u}$ et de remarquer que l'on a

$$\partial_t(\frac{1}{2}(\rho_0(\frac{\partial \vec{u}}{\partial t})^2) = T_0 \partial_t \vec{u}_0 \partial^2_{x^2} \vec{u}_0 = T_0 \partial_x(\partial_t \vec{u}_0 \partial_x \vec{u}_0) - T_0 \partial^2_{tx}(\vec{u}_0)\partial_x \vec{u}_0$$

et on intègre sur $[0,a]$, remarquant que le dernier terme est la dérivée par rapport à t de $\frac{1}{2}T_0(\vec{u}_0)^2$.

Donner les solutions $L(p,q)$ de l'égalité

$$\frac{1}{2}(\rho_0 p^2 + T_0 q^2) = L(p,q) - p\frac{\partial L}{\partial p}(p,q).$$

(on dérivera cette égalité par rapport à p_1 et p_2).

En dérivant par rapport à p, on trouve $\rho_0 p = -p\partial^2_{p^2} L$, ce qui donne $\rho_0 = -\partial^2_{p^2} L$. Ainsi $L = -\frac{1}{2}\rho_0 p^2 + C(q)p + D(q)$. On remplace dans l'équation et on trouve $-\frac{1}{2}\rho_0 p^2 + C(q)p + D(q) + \rho_0 p^2 - pC(q) = \frac{1}{2}(\rho_0 p^2 + T_0 q^2)$, donc $C(q)$ est indeterminé et $D(q) = \frac{1}{2}T_0 q^2$.

c) Montrer que l'équation des cordes vibrantes est le système des équations d'Euler pour le Lagrangien $L(p,q) = \frac{1}{2}T_0 q^2 - \frac{1}{2}\rho_0 p^2$. Peut-on appliquer la théorie classique de minimisation?

On applique le résultat du 1, a), car $\partial_p L = -\rho_0 p$, $\partial_q L = T_0 q$.

Déduire de 1) que

• lorsque les deux extrémités de la corde sont fixées, les conditions en 0 et a sont les conditions de Dirichlet homogènes $u = 0$

• lorsqu'une extrémité de la corde est libre, la condition à cette extrémité s'écrit $\frac{\partial \vec{u}}{\partial x} = 0$, qui est la condition de Neumann. En déduire que l'énergie E est conservée.

C'est la traduction des résultats de 1).

8.5 Partie II

On cherche à minimiser la valeur moyenne de la tension J:

$$J(v_0) = \frac{1}{T}\int_0^T v_0(t)dt$$

sous les conditions $v_0(0) = 0$, $v_0(T) = V$ (c'est à dire un système dans lequel on établit une tension V en un temps T)
et sous la contrainte d'énergie dissipée par effet Joule constante:

$$K = \int_0^T Ri^2(t)dt$$

où le courant électrique est produit par la mise sous tension $v_0(t)$ d'un condensateur C et d'une résistance R disposés en parallèle (même tension).

a) Peut-on résoudre ce problème en considérant une perturbation $\varepsilon w(t)$ de la tension $v_0(t)$? Justifier.

b) On se donne ε_1 et ε_2, et on perturbe la solution cherchée par $\varepsilon_1 w_1(t) + \varepsilon_2 w_2(t)$.

Ecrire les conditions d'optimalité.

Montrer qu'il existe un réel λ tel que ces conditions d'optimalité correspondent aux conditions d'optimalité du lagrangien **augmenté** $J + \lambda K$, K étant considéré comme une fonction de $v(t)$. On pourra supposer à cet effet w_2 fixé. On admettra pour la suite ce résultat si il n'a pas été démontré.

c) On considère $\lambda \in \mathbb{R}$. Déterminer v_0 qui réalise le minimum de $J(v)+\lambda K(v), v(0) = 0, v(T) = 0$.

d) Déterminer λ de sorte que le v_0 trouvé au c) conduise à $i_0(t)$ tel que $\int_0^T R(i_0(t))^2 = K$. Calculer la solution $v_0(t)$ et interpréter. En particulier, pour K, V et R, C donnés, identifier les temps T pour lesquels on peut trouver $v_0(t)$.

Calculer la valeur maximum de J en fonction de K, V, R, C.

Bibliography

[1] J.C. Culioli: Optimisation: Cours à l'Ecole des Mines publié aux éditions Ellipses (1994)

[2] P. Faurre: Optimisation Cours à l'X

[3] B. Larrouturou et P.L. Lions: Cours d'optimisation et d'Analyse Numérique.

[4] J. Cea: Lectures on optimization-theory and algorithms: Tata institute of fundamental research, Bombay, 1978.

[5] H. Sagan: Boundary and Eigenvalue Problems in Mathematical Physics John Wiley and Sons, 1961.

[6] V. M. Tichomirov: Fundamental Principles of the Theory of Extremal Problems: John Wiley and Sons, 1982, 1986.

[7] P. G. Ciarlet: Introduction à l'analyse numérique matricielle et à l'optimisation Mathématiques Appliquées pour la maitrise, Masson, 1982.