Optimisation et programmation dynamique

Master mention Mathématiques et Applications, 1ère année
Université Paris Dauphine

Pierre Cardaliaguet

(Version du 18 décembre 2020)

Introduction

Ces notes sont un support pour le cours

Optimisation et programmation dynamique

du Master 1 de mathématiques appliquées de l'Université Paris Dauphine. L'objectif est de présenter quelques notions sur deux types de problèmes d'optimisation : l'optimisation dans $\mathbb{R}^n$ sous contraintes, d'une part, et le contrôle optimal, d'autre part. Ces deux problématiques se rencontrent fréquemment dans toutes les questions liées à la décision. Si les méthodes de résolution diffèrent sensiblement, les deux domaines ont recours à des concepts similaires (conditions d'optimalité, fonction valeur, etc...).

Dans la pratique, les problèmes rencontrés ont souvent une structure spéciale qui sortira du cadre général abstrait du cours. Il faudra alors adapter les techniques développées au problème étudié. Aussi, la solution explicite est en général inaccessible et il faut recourir à des méthodes numériques, celles-ci s'appuyant fortement sur l'analyse mathématique du problème. Nous donnerons un aperçu de quelques unes de ces méthodes.

Pré-requis : Le cours utilise souvent sans rappel des notions de calcul différentiel et d'analyse convexe de L2, de topologie et d'équations différentielles de L3, et d'analyse fonctionnelle de L3 et de M1. Quelques rappels d'analyse convexe sont néanmoins donnés au début de ces notes.

Références bibliographiques : Quelques références sont données à la fin du polycopié. Pour la partie sur l'optimisation sous contraintes, deux bonnes références en langue française sont les livres respectifs de P. G. Ciarlet et G. Allaire. Pour la partie sur le contrôle optimal, le poly du cours à l'ENSAE de G. Carlier constitue une bonne introduction. Nous citons aussi le livre de Lucas et Stokey qui est une excellent référence pour les problèmes en temps discret et contient de nombreux exemples économiques. Pour la partie en temps continu, une référence possible est le livre de Fleming et Rishel.

Remarque : Ces notes sont entièrement reprises du polycopié écrit par Olga Mula (lui-même exhaustivement relu par Yannick Viossat) pour ce cours en 2018-2019. Elles sont basées sur plusieurs documents existants. Le chapitre sur l'optimisation sous contraintes est un résumé de certains chapitres du livre de G. Allaire.

Table des matières

Partie I

Rappels

A toutes fins utiles, nous rappelons quelques éléments de calcul différentiel, analyse convexe et extrema.

1 Convexité

1.1 Fonctions convexes, strictement convexes, fortement convexes

Définition 1.1. Un ensemble $K \subset \mathbb{R}^n$ est dit convexe si $\forall x,\, y \in K$ on a $tx + (1-t)y \in K$ pour tout $t \in [0,1]$ (quels que soient deux points dans K, tout le segment qui les unit est dans K).

Définition 1.2. Soit $K \subset \mathbb{R}^n$ un ensemble convexe et $f : K \to \mathbb{R}$ une fonction.

1. On dit que f est **convexe** sur K si

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y), \quad \forall x, y \in K,\ t \in [0,1].$$

2. On dit que f est **strictement convexe** sur K si

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y), \quad \forall x \neq y \in K,\ t \in]0,1[.$$

3. On dit que f est **fortement convexe** sur K s'il existe $\alpha > 0$ tel que

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{\alpha}{2}t(1-t)\|x-y\|^2, \quad \forall x, y \in K,\ t \in [0,1].$$

4. On dit que f est **concave** si $-f$ est convexe (idem pour strictement/fortement concave).

Il est facile de voir que

$$\text{fortement convexe} \Rightarrow \text{strictement convexe} \Rightarrow \text{convexe}$$

mais les réciproques ne sont pas vraies en général.

Voici deux critères classiques caractérisant la convexité pour une fonction de classe $\mathcal{C}^1$:

Proposition 1.3 (Caractérisation de la convexité)**.** *Soit $K \subset \mathbb{R}^n$ un ensemble convexe et $f : K \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$ dans un voisinage de K. Les trois assertions suivantes sont équivalentes :*

(i) f est convexe sur K

(ii) $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$, $\forall x, y \in K$

(iii) $\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0$, $\forall x, y \in K$.

La stricte convexité a la même caractérisation en remplaçant les inégalités par des inégalités strictes. Finalement, la forte convexité se caractérise de façon similaire.

Proposition 1.4 (Caractérisation de la forte convexité)**.** *Soit $K \subset \mathbb{R}^n$ un ensemble convexe et $f : K \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$ dans un voisinage de K. Les trois assertions suivantes sont équivalentes :*

(i) f est α-fortement convexe sur K (avec $\alpha > 0$)

(ii) $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|^2, \quad \forall x, y \in K$

(iii) $\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \alpha \|y - x\|^2, \quad \forall x, y \in K.$

Définition 1.5. On appelle **fonction elliptique** une fonction $f : K \to \mathbb{R}$ de classe $\mathcal{C}^1$ et fortement convexe. Ces fonctions se caractérisent par la proposition 1.4.

Proposition 1.6 (Quelques propriétés)**.**

1. *Toute combinaison linéaire à coefficients positifs d'une famille de fonctions convexes est convexe.*
2. *Toute composition d'une fonction $f : \mathbb{R}^n \to \mathbb{R}$ et d'une fonction convexe croissante $g : \mathbb{R} \to \mathbb{R}$ est convexe.*

   Preuve de (2). Comme $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$ pour tout x, $y \in \mathbb{R}$, alors, comme g est croissante,

   $$g \circ f(tx + (1-t)y) \leq g(tf(x) + (1-t)f(y)) \leq tg \circ f(x) + (1-t)g \circ f(y)$$

   où nous avons utilisé la convexité de g dans la dernière inégalité. □

1.2 Exemples de fonctions convexes, strictement convexes, fortement convexes

1. La fonction $f : \mathbb{R} \to \mathbb{R}$ donnée par $f(x) = x^2$ est fortement convexe.

   Preuve. Pour tout x, $y \in \mathbb{R}$,

   $$\langle \nabla f(y) - \nabla f(x), y - x \rangle = (f'(y) - f'(x))(y - x) = 2|y - x|^2$$

   donc la fonction est fortement convexe avec $\alpha = 2$. □

2. De façon similaire, la norme euclidienne au carré $f : \mathbb{R}^n \to \mathbb{R}$ avec $f(x) = \|x\|^2$ est aussi fortement convexe avec $\alpha = 2$.

   Preuve. Comme $\nabla f(x) = 2x$ pour tout $x \in \mathbb{R}^n$, on a $\langle \nabla f(y) - \nabla f(x), y - x \rangle = 2\|y - x\|^2$. □

3. Plus généralement, soit $f : \mathbb{R}^n \to \mathbb{R}$ la fonction donnée par

   $$f(x) = \frac{1}{2}\langle Ax, x \rangle + \langle b, x \rangle + c \tag{1.1}$$

   avec $\mathcal{M}_n(\mathbb{R})$ une matrice carrée réelle de taille n et $b \in \mathbb{R}^n$ et $c \in \mathbb{R}$. On a

   $$\langle \nabla f(y) - \nabla f(x), y - x \rangle = \langle A(y - x), (y - x) \rangle$$

   Par conséquent,

(a) A semi-définie positive $\Leftrightarrow$ f convexe.

(b) A définie positive de plus petite valeur propre $\lambda_{\min} > 0 \Leftrightarrow$ f fortement convexe avec $\alpha = \lambda_{\min}$.

1.3 Fonctions coercives

Définition 1.7. Soit $K \subset \mathbb{R}^n$ un ensemble non borné (par exemple, $\mathbb{R}^n$). Une fonction $f : K \to \mathbb{R}$ est coercive sur K si

$$\lim_{x \in K,\, \|x\| \to +\infty} f(x) = +\infty.$$

Ceci peut s'écrire de façon équivalente :

Pour toute suite $(x_k)_{k \in \mathbb{N}}$ d'éléments de K telle que $\|x_k\| \to \infty$, alors $f(x_k) \to \infty$.

Remarque : Comme toutes les normes sont équivalentes sur $\mathbb{R}^n$, en pratique, on choisit la norme la plus adaptée à la fonction f étudiée.

Proposition 1.8. *Soit $K \subset \mathbb{R}^n$ un ensemble convexe non borné. Si $f : K \to \mathbb{R}$ est de classe $\mathcal{C}^1$ sur un voisinage de K et fortement convexe sur K, alors f est coercive sur K.*

Preuve. La formule (ii) de la Proposition 1.4 pour les fonctions fortement convexes nous permet d'écrire pour tout $x, y \in K$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \|y - x\|^2. \tag{1.2}$$

De plus, par l'inégalité de Cauchy-Schwarz,

$$\langle \nabla f(x), x - y \rangle \leq \|\nabla f(x)\| \|x - y\|,$$

donc,

$$\langle \nabla f(x), y - x \rangle \geq -\|\nabla f(x)\| \|x - y\|.$$

En injectant la dernière inégalité dans (1.2), il vient

$$f(y) \geq f(x) - \|\nabla f(x)\| \|x - y\| + \frac{\alpha}{2} \|y - x\|^2.$$

En fixant $x \in K$ et en faisant $\|y\| \to \infty$ (ce qui implique $\|y - x\| \to \infty$), on obtient le résultat. □

Exemples de fonctions coercives :

1. Par la proposition 1.8, les fonctions fortement convexes de classe $\mathcal{C}^1$ sont coercives. En particulier, la fonction
$$f(x) = \frac{1}{2} \langle Ax, x \rangle + \langle b, x \rangle + c \tag{1.3}$$
est fortement convexe, donc coercive, si A est définie positive.
2. Toute fonction minorée par une fonction coercive est coercive.

3. Soit la fonction $f : \mathbb{R}^n \to \mathbb{R}$ de la forme

$$\forall x = (x_1, \ldots, x_n) \in \mathbb{R}^n,\ f(x) = \sum_{i=1}^{n} f_i(x_i)$$

où les fonctions $f_i : \mathbb{R} \to \mathbb{R}$ sont minorées et coercives. Alors f est coercive.

Preuve. Pour tout $i \in \{1, \ldots, n\}$, f_i est minoré par une constante m_i. Posons $m = \max_{1 \leq i \leq n} |m_i|$ et soit $M > 0$ fixé. Comme chaque f_i est coercif, il existe une constante $R_i > 0$ telle que

$$\forall x_i \in \mathbb{R}, \quad |x_i| \geq R_i \Rightarrow\ f_i(x_i) \geq M + nm\ .$$

Soit $R = \max_{1 \leq i \leq n} R_i$. Alors pour tout $x \in \mathbb{R}^n$ avec $\|x\|_\infty \geq R$, il existe un indice $i \in \{1, \ldots, n\}$ tel que $|x_i| \geq R \geq R_i$ et donc $f_i(x_i) \geq M + nm$. Pour $j \neq i$,

$$f_j(x_j) \geq m_j \geq -m.$$

Ces minorations permettent de conclure que

$$f(x) \geq M.$$

On a donc montré que

$$\forall M \geq 0,\ \exists R > 0 \text{ tel que } \forall x \in \mathbb{R}^n,\ \|x\|_\infty \geq R\ \Rightarrow\ f(x) \geq M\ .$$

Par conséquent,

$$\lim_{\|x\|_\infty \to \infty} f(x) = +\infty$$

ce qui prouve la coercivité de f. □

Partie II

Optimisation sous contraintes en dimension finie

Dans toute la suite, nous considérons $f : \mathbb{R}^n \to \mathbb{R}$ une fonction continue et K un sous-ensemble non vide de $\mathbb{R}^n$. Nous allons étudier le problème d'optimisation

$$\min_{x \in K} f(x). \tag{$\mathcal{P}$}$$

Nous verrons tout d'abord des conditions suffisantes garantissant que le minimum existe bien. L'expression de ces conditions varie en fonction de la structure de l'ensemble K et la régularité de la fonction f. On distinguera si K est ouvert, fermé, borné ou convexe et si f est seulement continu ou bien de classe $\mathcal{C}^1$.

Nous aborderons ensuite la question réciproque, c'est à dire les conditions nécessaires que satisfont les points de minimum. Cela amène naturellement la question de savoir si ces conditions nécessaires sont suffisantes. En utilisant des éléments de la théorie de la dualité, nous verrons que les conditions nécessaires sont suffisantes dans certains cas. La théorie de la dualité nous permettra aussi de construire des méthodes numériques efficaces que nous étudierons.

1 Terminologie

Infimum et minimum

Définition 1.1. On appelle **infimum** de f sur K la valeur $l \in [-\infty, +\infty[$ telle que

1. $\forall x \in K,\ f(x) \geq l$,
2. il existe une suite (x_n) d'éléments de $\mathbb{R}^n$ telle que

$$\forall n \geq 0,\ x_n \in K \text{ et } \lim_n f(x_n) = l$$

Cette valeur est notée $\inf_{x \in K} f(x)$:

$$l = \inf_{x \in K} f(x)\ .$$

Remarques :

1. L'infimum existe toujours. Il est fini (c'est-à-dire que $l \neq -\infty$) si et seulement si la fonction f est **minorée** sur K, c'est-à-dire s'il existe une constante $M \in \mathbb{R}$ telle que

$$\forall x \in K,\ f(x) \geq M\ .$$

2. Si f n'est pas minorée, alors l'infimum de f est $-\infty$.

3. Une suite (x_n) telle que $x_n \in K$ pour tout $n \in N$ et

$$\lim_n f(x_n) = \inf_{x \in K} f(x)$$

est appelée une **suite minimisante** du problème de minimisation.

Définition 1.2. On appelle **minimum** de f sur K la valeur $l \in]-\infty, +\infty[$ - si elle existe - pour laquelle il existe un élément $\bar{x} \in K$ tel que

1. $\forall x \in K,\ f(x) \geq l$.
2. $f(\bar{x}) = l$.

Cette valeur est notée $\min_{x \in K} f(x)$.

On dit alors que f atteint son minimum sur K en $\bar{x}$, ou que le problème $\min_{x \in K} f(x)$ admet $\bar{x}$ comme solution.

Remarques :

1. Par abus de language, on appelle aussi minimum (on dit aussi "minimum global") un élément $\bar{x} \in K$ satisfaisant les propriétés ci-dessus (en toute rigueur, $\bar{x}$ devrait s'appeler "argument du minimum".)
2. Contrairement à l'infimum, le minimum n'existe pas toujours. Des conditions suffisantes d'existence sont données ci-dessous.

Une autre notion très utile en optimisation est celle de minimum local :

Définition 1.3. On appelle **point de minimum local** de f sur K un point $\bar{x}$ de K pour lequel il existe un voisinage V de $\bar{x}$ tel que

$$\forall x \in V \cap K,\ f(x) \geq f(\bar{x})\ .$$

Bien sûr, un point de minimum est un minimum local, mais l'assertion inverse est fausse en général. Une fonction admet le plus souvent beaucoup plus de minima locaux que de minima globaux.

2 Conditions générales d'existence d'un minimum

2.1 Conditions suffisantes lorsque K est ouvert ou fermé

Il est possible d'énoncer des conditions suffisantes d'existence de minima pour le problème $(\mathcal{P})$ dans le cas très général où K est fermé ou bien ouvert.

Théorème 2.1 (Condition suffisante, K fermé)**.**
Soit $K \subset \mathbb{R}^n$ un ensemble non-vide et fermé et $f : K \to \mathbb{R}$ une fonction continue. Le problème

$$(\mathcal{P}) \qquad \min_{x \in K} f(x)$$

admet au moins une solution si l'une des deux conditions suivantes est satisfaite :

1. *la contrainte K est bornée (théorème de Weierstrass),*
2. *f est coercive sur K.*

Rappelons que dans le premier cas, f a aussi un maximum sur K.

Preuve. Dans le premier cas, K est fermé et borné, donc il est compact. Comme f est continue, le théorème de Weierstrass assure que f est bornée sur K et elle atteint ses bornes. Donc il existe au moins un point de minimum.

Prouvons le second point. Soit $x_0 \in K$ fixé. La coercivité de f sur K entraîne qu'il existe $r > 0$ tel que

$$\forall x \in K, \quad \|x\| \geq r \quad \Rightarrow f(x) > f(x_0) \tag{2.1}$$

Donc, en notant $\overline{B(0,r)}$ la boule fermée de $\mathbb{R}^n$ de centre 0 et de rayon r, nous avons

$$\inf_{x \in K} f(x) = \inf_{x \in K \cap \overline{B(0,r)}} f(x).$$

Comme l'ensemble $K \cap \overline{B(0,r)}$ est fermé et borné et que f est continue, le théorème de Weierstrass assure que f atteint ses bornes dans $K \cap \overline{B(0,r)}$. Ceci assure l'existence d'un minimum x^* dans $K \cap \overline{B(0,r)}$. Ce minimum est aussi le minimum sur K. En effet, pour tout $x \in K$:

1. soit $x \in K \cap \overline{B(0,r)}$, et alors $f(x) \geq f(x^*)$ car f atteint son minimum sur $K \cap \overline{B(0,r)}$ en x^*,
2. soit $x \notin K \cap \overline{B(0,r)}$, auquel cas on a $f(x) > f(x_0) \geq f(x^*)$ où la seconde inégalité vient du fait qu'on a forcément $\|x_0\| < r$ (en effet, si $\|x_0\| \geq r$, l'inégalité (2.1) donnerait $f(x_0) > f(x_0)$, ce qui est absurde).

Ceci montre donc que x^* est un minimum de f sur K. □

Si K est un ensemble ouvert, le problème est plus compliqué. Signalons la condition suffisante suivante :

Proposition 2.2 (Condition suffisante, K ouvert)**.**
On suppose que K est un ouvert borné, que f est continue sur $\overline{K}$, et qu'il existe un point x_0 de K tel que

$$\forall x \in \partial K, \ f(x) > f(x_0) \ .$$

où ∂K est la frontière de K. Alors le problème $(\mathcal{P})$ admet une solution.

Preuve. Comme l'ensemble $\overline{K}$ est compact, la fonction continue f admet un minimum x^* sur $\overline{K}$ par le thèorème 2.1. Cet élément est donc tel que

$$\forall x \in \overline{K}, \ f(x) \geq f(x^*) \ .$$

Montrons par l'absurde que x^* est bien dans l'ouvert K et ne se trouve pas au bord. Supposons que $x^* \in \partial K$. Alors $f(x^*) \leq f(x)$ pour tout $x \in K$. Or, par hypothèse, il existe $x_0 \in K$ tel que $f(x) > f(x_0)$ pour tout $x \in \partial K$, donc pour $x = x^*$, $f(x^*) > f(x_0)$. Cette contradiction nous permet de conclure que x^* est bien dans l'ouvert K. □

2.2 Conditions nécessaires et suffisantes lorsque K est convexe

à présent, rajoutons un peu plus de structure sur l'ensemble K et supposons qu'il est convexe. Ceci permet d'énoncer les conditions nécessaires d'optimalité suivantes, c'est-à-dire des conditions, portant sur la dérivée de f.

Théorème 2.3 (Condition nécessaire, K convexe : **Inéquation d'Euler**)**.**
Soit $\Omega \subset \mathbb{R}^n$ un ouvert, $K \subset \Omega$ un ensemble convexe et $f : \Omega \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$. Si x^* est un minimum local de f sur K, alors

$$\langle \nabla f(x^*), y - x^* \rangle \geq 0, \quad \forall y \in K. \tag{2.2}$$

Preuve. Soit $y \in K$ et $h \in]0, 1]$. Alors, $x^* + h(y - x^*) \in K$ car K est convexe. De plus, comme x^* est un minimum local de K,

$$\frac{f(x^* + h(y - x^*)) - f(x^*)}{h} \geq 0.$$

On trouve (2.2) en faisant un développement de Taylor de $f(x^*+h(y-x^*))$ autour de x^* au premier ordre, puis en faisant tendre h vers 0^+. □

Nous verrons dans le théorème suivant 2.5 que la condition nécessaire (2.2) devient suffisante lorsque f est convexe. Avant de présenter ce résultat, il est important de voir que la condition (2.2) se réduit à **l'équation d'Euler**

$$\nabla f(x^*) = 0.$$

lorsque K est un ouvert (en particulier, lorsque $K = \mathbb{R}^n$).

Théorème 2.4 (Condition nécessaire, K ouvert : équation d'Euler)**.**
Soit $K \subseteq \mathbb{R}^n$ un ouvert et $f : K \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$. Si x^* est un minimum local de f sur K, alors

$$\nabla f(x^*) = 0. \tag{2.3}$$

Preuve. Nous allons montrer que si x^* est un minimum local dans l'ouvert K, alors

$$\langle \nabla f(x^*), y - x^* \rangle \geq 0, \quad \forall y \in \mathbb{R}^n \quad \Longleftrightarrow \quad \nabla f(x^*) = 0.$$

L'implication inverse étant évidente, regardons l'implication directe. Soit y un vecteur quelconque de $\mathbb{R}^n$. Comme x^* appartient à K, qui est ouvert, il existe $h_0 > 0$ tel que, pour tout $h \in [0, h_0]$, le point $x^* + hy$ appartient à K. Or x^* étant un minimum local du problème, on a

$$f(x^* + hy) - f(x^*) \geq 0$$

Comme

$$f(x^* + hy) - f(x^*) = h\langle \nabla f(x^*), y \rangle + h\epsilon(hy) \ ,$$

en divisant l'inegalité ci-dessus par $h > 0$, et en faisant tendre h vers 0, on obtient

$$\langle \nabla f(x^*), y \rangle \ \geq \ 0$$

Comme cette inegalité est vraie pour tout $y \in \mathbb{R}^n$, elle est également vraie pour $-y$. Donc $\langle \nabla f(x^*), -y \rangle \ \geq \ 0$. D'où $\langle \nabla f(x^*), y \rangle \ = \ 0$. Donc finalement

$$\langle \nabla f(x^*), y \rangle \ = \ 0, \quad \forall y \in \mathbb{R}^n,$$

c'est-à-dire que $\nabla f(x^*) = 0$.

Remarquons qu'il existe une preuve beaucoup plus courte dans le cas $K = \mathbb{R}^n$. Comme $\mathbb{R}^n$ est convexe, l'inégalité d'Euler (2.2) donne $\langle \nabla f(x^*), y - x^* \rangle \geq 0$, $\forall y \in \mathbb{R}^n$. Mais comme dans ce cas $y - x^*$ engendre tout l'espace $\mathbb{R}^n$ car $y \in \mathbb{R}^n$, l'inégalité d'Euler (2.2) implique $\nabla f(x^*) = 0$. □

Prouvons maintenant que l'inéquation d'Euler est une conditions nécessaire et suffisante lorsque K et f sont tous les deux convexes.

Théorème 2.5 (Condition nécessaire et suffisante, K et f convexes : Inéquation d'Euler)**.** *Si, en plus des hypothèses du théorème 2.3, f est convexe, alors,*

$$\langle \nabla f(x^*), y - x^* \rangle \geq 0, \quad \forall y \in K \quad \Leftrightarrow \quad x^* \textit{ est un minimum global de } f \textit{ sur } K \tag{2.4}$$

Preuve. L'implication réciproque étant assurée par le théorème 2.3, nous n'avons à prouver que l'implication directe. Comme f est convexe sur K, pour tout $y \in K$,

$$f(y) \geq f(x^*) + \langle \nabla f(x^*), y - x^* \rangle$$

grâce à la formule (ii) de la Proposition 1.3. Comme $\langle \nabla f(x^*), y - x^* \rangle \geq 0$, on a donc $f(y) \geq f(x^*)$ pour tout $y \in K$ ce qui prouve que x^* est un minimum global de f sur K. □

Dans le cas où K est convexe et f est non seulement convexe, mais strictement ou fortement convexe, nous avons les deux résultats suivants qui permettront de garantir existence, unicité et caractérisation du point de minimum.

Théorème 2.6 (Unicité)**.** *Soit $K \subset \mathbb{R}^n$ un ensemble convexe et $f : K \to \mathbb{R}^n$ une fonction strictement convexe. Alors il existe au plus un point de minimum sur K.*

Preuve. Nous allons raisonner par l'absurde. Soient x_1 et $x_2 \in K$ avec $x_1 \neq x_2$ deux points de minimum de f sur K. Nous avons donc $f(x_1) = f(x_2) \leq f(x)$ pour tout $x \in K$. Comme f est strictement convexe et que $(x_1 + x_2)/2 \in K$ car K est convexe, nous avons

$$f\left(\frac{x_1 + x_2}{2}\right) < \frac{1}{2} f(x_1) + \frac{1}{2} f(x_2) = f(x_1)$$

Ceci contredit le fait que x_1 soit un minimum. □

Théorème 2.7 (Existence et unicité)**.** *Soit $K \subset \mathbb{R}^n$ un ensemble convexe fermé et $f : \mathbb{R}^n \to \mathbb{R}$ une fonctions fortement convexe de classe $\mathcal{C}^1$. Alors il existe un unique point x^* de minimum de f sur K et ce point vérifie l'inéquation d'Euler*

$$\langle \nabla f(x^*), y - x \rangle \geq 0, \quad \forall y \in K.$$

Preuve. Comme f est fortement convexe et de classe $\mathcal{C}^1$, elle est continue et coercive. L'existence du minimum est garantie par le théorème 2.1. Le théorème 2.6 permet d'assurer l'unicité et l'inéquation d'Euler est une conséquence directe du théorème 2.5. □

3 Le théorème de Kuhn et Tucker

Lorsque K n'est pas convexe, les conditions nécessaires sont plus difficiles à énoncer. Le théorème de Kuhn & Tucker (ou Karush, Kuhn et Tucker, KKT) donne des conditions nécessaires lorsque la contrainte K est de la forme

$$K = \{x \in \mathbb{R}^n,\ G(x) \leq 0,\ H(x) = 0\},$$

où

$$\begin{aligned} G : \mathbb{R}^n &\to \mathbb{R}^m \\ x &\to G(x) = (g_1(x), \ldots, g_m(x))^T \end{aligned}$$

avec les $g_i : \mathbb{R}^n \to \mathbb{R}$ et

$$\begin{aligned} H : \mathbb{R}^n &\to \mathbb{R}^p \\ x &\to H(x) = (h_1(x), \ldots, h_p(x))^T \end{aligned}$$

avec les $h_i : \mathbb{R}^n \to \mathbb{R}$. L'inégalité $G(x) \leq 0$ est à comprendre composante par composante et les fonctions g_i et h_i sont toutes supposées de classe $\mathcal{C}^1$ de $\mathbb{R}^n$ dans $\mathbb{R}$. Notons que K est fermé mais n'est pas forcément borné ni convexe. Si f était coercive, le théorème 2.1 assurerait l'existence d'au moins un minimiseur du problème.

Dans le but de bien comprendre le sens des hypothèses du théorème de Kuhn et Tucker, nous procédons progressivement en étudiant des contraintes K de difficulté croissante. Cela permettra aussi de comprendre de façon plus naturelle la notion essentielle de multiplicateurs de Lagrange qui intervient dans l'énoncé.

3.1 Contraintes d'égalité

K espace affine : Supposons que K est un espace affine de $\mathbb{R}^n$ (qui est un ensemble convexe fermé non borné dans $\mathbb{R}^n$). On peut donc exprimer $K = x_0 + V$ avec V un sous-espace de $\mathbb{R}^n$. Si $f : K \to \mathbb{R}^n$ est de classe $\mathcal{C}^1$, alors l'inéquation d'Euler (2.2) garantit que si x^* est un minimum du problème $(\mathcal{P})$, alors $\langle \nabla f(x^*), y - x^* \rangle \geq 0$ pour tout $y \in K$. Donc, quand y varie dans K, $y - x^*$ varie dans V. La condition d'Euler donne alors $\langle \nabla f(x^*), v \rangle \geq 0$ pour tout $v \in V$. D'où

$$\langle \nabla f(x^*), v \rangle = 0, \quad \forall v \in V.$$

Par conséquent $\nabla f(x^*) \in V^\perp$. En particulier, si V est une intersection de $p \leq n$ hyperplans de la forme

$$V = \{x \in \mathbb{R}^n \ :\ \langle v_i, x \rangle = 0,\ 1 \leq i \leq p\}$$

où les $v_i \in \mathbb{R}^n$, alors

$$V^\perp = \text{span}\{v_i\}_{i=1}^p.$$

Par conséquent, vu que $\nabla f(x^*) \in V^\perp$, les conditions nécessaires d'optimalité peuvent donc s'écrire sous la forme

$$x^* \in K \quad \text{et} \quad \exists\, \mu_1, \ldots, \mu_p \in \mathbb{R} \text{ tels que } \nabla f(x^*) + \sum_{i=1}^{p} \mu_i v_i = 0.$$

Les coefficients μ_i sont appelés multiplicateurs de Lagrange. Ils joueront un rôle important dans le cas plus général du théorème KKT.

K **avec contraintes d'égalité :** Considérons maintenant le cas où

$$K = \{v \in \mathbb{R}^n \,:\, H(v) = 0, \}. \tag{3.1}$$

L'ensemble n'étant pas convexe en général, il n'est pas possible d'utiliser l'inéquation d'Euler (2.2). Cependant, il est possible de généraliser ce résultat en cherchant quelles sont les directions admissibles pour lesquelles l'inéquation reste vraie. Dans le cas présent, nous pourrons utiliser le résultat suivant que l'on admettra.

Proposition 3.1. *Soit x^* un minimum local de f sur l'ensemble K défini en* (3.1)*. Si la famille de vecteurs $(\nabla h_i(x^*))_{i=1}^p$ est libre, alors*

$$\langle \nabla f(x^*), y\rangle = 0, \quad \forall y \in K(x^*)$$

avec

$$K(x^*) = \{v \in \mathbb{R}^n \,:\, \langle \nabla h_i(x^*), v\rangle = 0,\, 1 \le i \le p\} \tag{3.2}$$

L'ensemble $K(x^*)$ constitue ce que l'on appelle le **cone des directions admissibles** au point x^*. Dans le cas présent, $K(x^*)$ est plus qu'un cone, il s'agit du plan tangent à la variété K au point x^*. La proposition précédente nous permet d'énoncer le résultat suivant.

Théorème 3.2. *Soit x^* un point de l'ensemble K défini en* (3.1)*. Supposons que les fonctions h_i soient de classe $\mathcal{C}^1$ dans un voisinage de x^* pour tout $i = 1, \dots, p$. De plus, supposons que les vecteurs $(\nabla h_i(x^*))_{i=1}^p$ soient linéairement indépendants. Alors, si x^* est un minimum de f sur K, il existe $\mu_1, \dots \mu_p \in \mathbb{R}$ appelés multiplicateurs de Lagrange, tels que*

$$\nabla f(x^*) + \sum_{i=1}^{p} \mu_i \nabla h_i(x^*) = 0. \tag{3.3}$$

Preuve. Les hypothèses de la proposition 3.1 étant satisfaites, nous avons

$$\langle \nabla f(x^*), y\rangle = 0, \quad \forall y \in K(x^*)$$

où $K(x^*)$ est le plan tangent à la variété K au point x^* défini en (3.2). Donc $\nabla f(x^*)$ est engendré par les $(\nabla h_i(x^*))_{i=1}^m$, d'où la formule (3.3). □

3.2 Contraintes d'inégalité

K **cone :** Supposons que K est un cone convexe fermé, c'est à dire que K est un convexe fermé tel que pour tout $y \in K$, alors $\lambda y \in K$ pour tout $\lambda \ge 0$. En appliquant l'inéquation d'Euler (2.2) à $y = 0$ et $y = 2x^*$, il vient

$$\langle \nabla f(x^*), x^*\rangle = 0. \tag{3.4}$$

Par conséquent, la condition d'Euler (2.2) devient

$$\langle \nabla f(x^*), y\rangle = 0, \quad \forall y \in K.$$

Si

$$K = \{x \in \mathbb{R}^n \,:\, \langle v_i, x\rangle \le 0,\, 1 \le i \le m\}$$

où les $v_i \in \mathbb{R}^n$, il est possible de prouver par le lemme de Farkas (que l'on ne présentera pas par manque de temps) que les conditions nécessaires d'optimalité sont

$$x^* \in K \quad \text{et} \quad \exists \lambda_1, \ldots, \lambda_m \geq 0 \text{ tels que } \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i v_i = 0. \tag{3.5}$$

Compte tenu de l'égalité (3.4) et de la relation (3.5), on voit que si $\langle v_i, x^* \rangle < 0$, alors $\lambda_i = 0$. Les $\lambda_i \geq 0$ sont de nouveau appelés multiplicateurs de Lagrange.

K avec contraintes d'inégalité : Prenons maintenant le cas où

$$K = \{v \in \mathbb{R}^n \, : \, G(v) \leq 0\}. \tag{3.6}$$

De façon similaire au cas avec contraintes d'égalité, la démarche consiste à trouver pour chaque point x le cone de directions admissibles $K(x)$ où il est possible de formuler une équation ou inéquation d'Euler. Dans les cas précédents, les gradients des contraintes ont suffi pour identifier $K(x)$. Afin que cela reste vrai dans le cas présent, il est nécessaire d'ajouter des conditions supplémentaires sur les contraintes, appelées **conditions de qualification**. Ceci provient du fait que dans le cas actuel, toutes les contraintes ne vont pas jouer le même rôle en chaque point x et il est nécessaire d'ajouter des conditions supplémentaires dans le but de garantir que les gradients des contraintes déterminent entièrement les directions dans lesquelles il est possible de faire des variations autour d'un point. Dit d'une façon plus abstraite, les conditions de qualification garantissent que le cone des directions admissible peut s'obtenir en linéarisant les contraintes.

Définition 3.3. Soit $x \in K$. L'ensemble $I(x) = \{i \in \{1, \ldots m\} \, : \, g_i(x) = 0\}$ est appelé l'ensemble de contraintes actives au point x.

Définition 3.4. Les contraintes (3.6) sont qualifiées au point $x \in K$ si $I(x) = \emptyset$ ou bien s'il existe une direction $w \in \mathbb{R}^n$ telle que

$$\begin{aligned} &\text{soit } \langle \nabla g_i(x), w \rangle < 0, \quad \forall i \in I(x). \\ &\text{soit } \langle \nabla g_i(x), w \rangle = 0 \text{ et } g_i \text{ est affine.} \end{aligned} \tag{3.7}$$

Remarque 3.5. Les contraintes sont automatiquement qualifiées dans deux cas importants, ce qui facilite grandement l'analyse de nombreux cas pratiques :

- Si toutes les fonctions g_i sont affines, nous pouvons prendre $w = 0$, ce qui satisfait automatiquement la condition de qualification.
- Plus généralement, si toutes le contraintes g_i sont convexes et $\mathcal{C}^1$, la contrainte K est automatiquement qualifiée si l'intérieur de K est non vide.

  Preuve. Fixons un point x_0 à l'intérieur de K. Soit x un point du bord de K et posons $w = x_0 - x$. Le vecteur v est non nul car x_0 appartient à l'intérieur de K. Comme la contrainte g_i est convexe, on a pour tout $i \in I(x)$,

  $$\langle \nabla g_i(x), w \rangle \leq g_i(x_0) - g_i(x) = g_i(x) < 0$$

  car $g_i(x) = 0$ et x_0 est dans l'intérieur de K. Donc la contrainte est qualifiée d'après la définition (3.4). □

Théorème 3.6. *Soit x^* un point de l'ensemble K donné par (3.6) où les fonctions $g_1, \dots, g_m$ sont de classe $\mathcal{C}^1$. Supposons que les contraintes sont qualifiées en x^*. Alors, si x^* est un minimum local de f sur K, il existe $\lambda_1, \dots, \lambda_m \geq 0$ appelés multiplicateurs de Lagrange tels que*

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*) = 0, \quad \lambda_i \geq 0, \quad \lambda_i = 0 \text{ si } g_i(x^*) < 0, \, \forall i \in \{1, \dots, m\}. \tag{3.8}$$

Remarque 3.7. La condition (3.8) peut s'écrire de façon équivalente comme

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*) = 0, \quad \lambda_i \geq 0, \quad \lambda_i g_i(x^*) = 0, \, \forall i \in \{1, \dots, m\}.$$

Cette forme est particulièrement utile pour la résolution pratique de problèmes. Notons aussi la notation condensée

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*) = 0, \quad \lambda \geq 0, \quad \lambda \cdot G(x^*) = 0, \, \forall i \in \{1, \dots, m\},$$

où $\lambda \geq 0$ veut dire que toutes le composantes du vecteur $\lambda = (\lambda_1, \dots \lambda_m)^T$ sont positives ou nulles. La condition $\lambda_i g_i(x^*) = 0$, $\forall i = 1, \dots, m$, ou encore $\lambda \cdot G(x^*) = 0$, est appelée **condition d'exclusion**.

3.3 Contraintes d'égalité et d'inégalité

K avec contraintes d'égalité et inégalité : Dans le cas général combinant des contraintes d'égalité et d'inégalité, K est de la forme

$$K = \{x \in \mathbb{R}^n, \; G(x) \leq 0, \; H(x) = 0\}. \tag{3.9}$$

Il faut définir la qualification des contraintes dans ce contexte. Comme précédemment, nous notons

$$I(x) = \{i \in \{1, \dots, m\} \, : \, g_i(x) = 0\}$$

l'ensemble de contraintes d'inégalité actives au point x.

Définition 3.8. Les contraintes (3.9) sont qualifiées au point $x \in K$ si toutes les contraintes sont affines ou bien si les deux conditions suivantes sont satisfaites :

1. les vecteurs $(\nabla h_j(x))_{j=1}^p$ sont linéairement indépendants
2. il existe une direction $\omega \in \mathbb{R}^n$ telle que

$$\begin{aligned} &\langle \nabla h_j(x), \omega \rangle = 0, \quad \forall j \in \{1, \dots, p\} \\ &\langle \nabla g_i(x), \omega \rangle < 0, \quad \forall i \in I(x). \end{aligned}$$

Nous sommes en mesure d'énoncer les conditions nécessaires d'optimalité sur l'ensemble (3.9).

Théorème 3.9 (Théorème de Kuhn et Tucker)**.** *Soit x^* un point de l'ensemble K défini dans* (3.9)*. On suppose que f, G et H sont de classe $\mathcal{C}^1$ et que les contraintes sont qualifiées en x^* au sens de la définition 3.8. Alors, si x^* est un minimum local sur K, il existe des multiplicateurs de Lagrange $\mu_1, \dots, \mu_p \in \mathbb{R}$ et $\lambda_1, \dots, \lambda_m \geq 0$ tels que*

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*) + \sum_{j=1}^{p} \mu_j \nabla h_j(x^*) = 0, \quad \lambda_i \geq 0, \quad \lambda_i g_i(x^*) = 0, \, \forall i = 1, \dots, m. \tag{3.10}$$

En pratique, il est souvent commode de retrouver le système d'optimalité (3.10) en introduisant le lagrangien du problème qui est la fonction

$$\begin{aligned} \mathcal{L} : \mathbb{R}^n \times (\mathbb{R}_+)^m \times \mathbb{R}^p &\to \mathbb{R} \\ (x, \lambda) &\mapsto \mathcal{L}(x, \lambda, \mu) \coloneqq f(x^*) + \sum_{i=1}^{m} \lambda_i g_i(x^*) + \sum_{j=1}^{p} \mu_j h_j(x^*). \end{aligned}$$

Le nom de multiplicateurs de Lagrange vient du fait que les λ_i et μ_i multiplient les contraintes dans le lagrangien. Les conditions nécessaires d'optimalité d'écrivent alors de façon condensée

$$\nabla_x \mathcal{L}(x^*, \lambda, \mu) = 0, \quad \lambda \geq 0, \quad \lambda \cdot G(x^*) = 0$$

Cette fonction joue donc un rôle très important dans l'expression des conditions nécessaires. Dans les sections suivantes, nous allons voir qu'il sera aussi essentiel pour développer des méthodes numériques efficaces de résolution.

3.4 Le cas convexe

Dans cette section, nous nous concentrons sur le cas convexe, entendu au sens de la définition suivante.

Définition 3.10 (Problème convexe)**.** On dit que le problème de minimisation $\min_{x \in K} f(x)$ est convexe si f est convexe et de classe $\mathcal{C}^1$, si les contraintes d'inégalité sont aussi convexes et de classe $\mathcal{C}^1$ et si les les contraintes d'égalité sont affines.

Nous allons prouver que dans ce cas, les conditions KKT données en (3.10) sont non seulement nécessaires, mais aussi suffisantes, et elles sont suffisantes sans aucune condition de qualification. La qualification reste malgré tout nécessaire pour énoncer la condition nécessaire, mais nous allons voir que dans ce cas elle s'exprime de façon très simple. En résumé, dans le cas convexe, nous avons :

Théorème 3.11 (CNS Cas Convexe)**.** *Dans le cas convexe,*

$$(KKT) \begin{cases} \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*) + \sum_{j=1}^{p} \mu_j \nabla h_j(x^*) = 0, \\ \lambda_i \geq 0, \quad \lambda_i g_i(x^*) = 0, \, \forall i = 1, \dots, m. \end{cases} \quad \begin{array}{c} \xrightarrow{\text{sans qualif.}\Rightarrow} \\ \xleftarrow[\text{qualif.}]{} \end{array} \quad x^* \text{ solution}$$

Preuve. L'implication réciproque étant garantie par le théorème 3.9 de Kuhn et Tucker, il suffit de montrer l'implication directe. Pour cela, soit $x^* \in K$ un point satisfaisant la relation KKT. On a donc

$$\nabla f(x^*) = -\sum_{i=1}^{m} \lambda_i \nabla g_i(x^*) - \sum_{j=1}^{p} \mu_j \nabla h_j(x^*). \tag{3.11}$$

La fonction f étant convexe sur K,

$$f(x) - f(x^*) \geq \langle \nabla f(x^*), x - x^* \rangle, \quad \forall x \in K$$

et en remplaçant $\nabla f(x^*)$ par la formule (3.11), on a

$$f(x) - f(x^*) \geq \sum_{i=1}^{m} \lambda_i \langle \nabla g_i(x^*), x^* - x \rangle + \sum_{i=1}^{p} \mu_i \langle \nabla h_i(x^*), x^* - x \rangle. \tag{3.12}$$

Les fonctions h_i étant affines, elles peuvent s'exprimer comme

$$h_i(x) = h_i(x^*) + \langle \nabla h_i(x^*), x - x^* \rangle.$$

Or, comme x et x^* sont dans K, on a $h_i(x) = h_i(x^*) = 0$ car ce sont des contraintes d'égalité. Donc

$$\langle \nabla h_i(x^*), x - x^* \rangle = 0$$

et l'inégalité (3.12) devient

$$\begin{aligned} f(x) - f(x^*) &\geq \sum_{i=1}^{m} \lambda_i \langle \nabla g_i(x^*), x^* - x \rangle \\ &\geq \sum_{i=1}^{m} \lambda_i \left(g_i(x^*) - g_i(x) \right), \end{aligned}$$

où nous avons utilisé le fait que les contraintes g_i sont convexes. Par ailleurs, comme chaque $\mu_i \geq 0$ et $g_i(x) \leq 0$ pour tout $x \in K$, on a donc $\sum_{i=1}^{m} \lambda_i g_i(x) \leq 0$ et donc

$$f(x) - f(x^*) \geq \sum_{i=1}^{m} \lambda_i g_i(x^*).$$

Finalement, par la condition d'exclusion, si $g_i(x^*) < 0$ alors $\mu_i = 0$ et si $g_i(x^*) = 0$ alors $\mu_i \geq 0$, d'où la conclusion

$$f(x) \geq f(x^*).$$

□

Les conditions de qualification dans le cas convexe sont en général très simples à examiner comme le montre la proposition suivante.

Proposition 3.12. *Si l'intérieur de K est non vide, alors les contraintes sont qualifiées au point $x \in K$ si les vecteurs $(\nabla h_j(x))_{j=1}^{p}$ sont linéairement indépendants.*

Preuve. Soit $x \in K$ fixé et soit x_0 un point de l'intérieur de K. En vue de la définition 3.8 sur la qualification, il suffit de vérifier qu'il existe une direction $\omega \in \mathbb{R}^n$ telle que

$$\begin{aligned} &\langle \nabla h_j(x), \omega \rangle = 0, \quad \forall j \in \{1, \dots, p\} \\ &\langle \nabla g_i(x), \omega \rangle < 0, \quad \forall i \in I(x). \end{aligned}$$

Prenons $\omega = x_0 - x$. La remarque 3.5 permet d'affirmer que la condition sur les g_i est vérifiée. De plus, suivant un raisonnement similaire à celui donné dans la preuve précédente, nous avons $\langle \nabla h_j(x), \omega \rangle = h(x_0) - h(x) = 0$ car les h_i sont affines et que $h(x) = h(x_0) = 0$ vu que x, $x_0 \in K$. □

3.5 Résumé : Démarche pour résoudre un problème de minimisation

Au vue des sections précédentes, la démarche pour résoudre un problème de minimisation sous contraintes se décompose en quatre étapes :

1. On montre *a priori* que le problème admet une solution. Pour cela, on pourra utiliser les conditions suffisantes d'existence de la section 2. Il sera parfois nécessaire d'utiliser des raisonnements spécifiques au problème donné.
2. On cherche les points où la contrainte n'est pas qualifiée. On appelle cet ensemble E_1. En pratique, on détermine l'ensemble des points qui ne vérifient pas la définition 3.8.
3. On cherche ensuite les points satisfaisant les conditions nécessaires de Kuhn et Tucker. On note cet ensemble E_2.
4. Si le problème a une solution, le minimum appartient à $E_1 \cup E_2$. Un minimum est donc un point de critère minimal dans $E_1 \cup E_2$.

Remarque : Dans le cas convexe, il suffit de trouver des points vérifiant les conditions nécessaires d'optimalité pour pouvoir conclure à l'existence d'un minimum (voir théorème 3.11).

3.6 Quelques exemples

1. **Une contrainte d'égalité :** On considère le problème suivant dans $\mathbb{R}^2$:

$$\min_{x^2+y^2=1} 2x + y$$

Avec les notations précédentes, nous avons $f(x, y) = 2x + y$ et $K = \{(x, y) : h(x, y) = 0\}$, avec $h(x, y) = x^2 + y^2 - 1$. Comme K est un ensemble compact et non vide (car $(1, 0) \in K$) et que f est continue, alors le théorème de Weierstrass assure que le problème admet au moins une solution $(x^*, y^*) \in K$.

La contrainte est qualifiée en tout point car $\nabla h(x, y) = (2x, 2y)^T$ est non nul pour tout $(x, y) \in K$. Le théorème 3.2 permet d'affirmer que si (x^*, y^*) est un minimum de f sur K, alors il existe $\mu \in \mathbb{R}$ tel que

$$\nabla f(x^*, y^*) + \mu \nabla h(x^*, y^*) = 0 \quad \Leftrightarrow \quad \begin{cases} 2 + 2\mu x^* = 0 \\ 1 + 2\mu y^* = 0 \\ (x^*)^2 + (y^*)^2 = 1 \end{cases}$$

On déduit des deux premières égalités que μ est non nul, et que $x^* = -1/\mu$, $y^* = -1/(2\mu)$. En reportant dans la dernière égalité, on a

$$(1/\mu)^2 + (1/(2\mu))^2 = 1$$

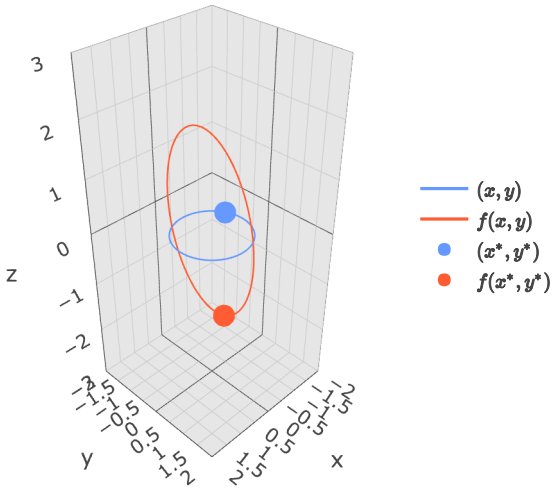

FIGURE II.1 – Illustration du problème $\min_{x^2+y^2=1} 2x + y$.

c'est-à-dire, $\mu = \sqrt{5}/2$ ou $\mu = -\sqrt{5}/2$. Dans le premier cas, $(x^*, y^*) = (-2\sqrt{5}/5, -\sqrt{5}/5)$, et dans le second $(x^*, y^*) = (2\sqrt{5}/5, \sqrt{5}/5)$. L'ensemble des points vérifiant les conditions nécessaires d'optimalité est donc $\{(2\sqrt{5}/5, \sqrt{5}/5), (-2\sqrt{5}/5, -\sqrt{5}/5)\}$. On sait (c'est le théorème) que le ou les minima du problème se situent parmi ces points. On calcule la valeur de f pour déterminer le minimum :

$$f(2\sqrt{5}/5, \sqrt{5}/5) = \sqrt{5} \text{ et } f(-2\sqrt{5}/5, -\sqrt{5}/5) = -\sqrt{5}$$

Il n'y a donc qu'un seul minimum : c'est le point $(-2\sqrt{5}/5, -\sqrt{5}/5)$. Une illustration de cet exemple est donnée en figure 1.

2. **Une contrainte d'inégalité :** On considère maintenant le problème suivant dans $\mathbb{R}^2$:

$$\min_{x^2+y^2 \leq 1} 2 + xy$$

Ici, $f(x, y) = 2 + xy$ et $K = \{(x, y) : g(x, y) \leq 0\}$, avec $g(x, y) = x^2 + y^2 - 1$. Comme K est compact et non vide (car $(0, 0) \in K$), le problème admet bien un minimum par le théorème de Weierstrass.

Pour tout $(x, y) \in K$, la contrainte est qualifiée. En effet, les points (x, y) dans l'intérieur du disque unité sont tels que $g(x, y) < 0$ strictement donc ils sont automatiquement qualifiés. Les points (x, y) sur le cercle unité sont tels que $g(x, y) = 0$ et il faut donc vérifier s'il existe une direction $\omega \in \mathbb{R}^2$ telle que $\langle \nabla g(x, y), \omega \rangle < 0$. On vérifie facilement que le choix $\omega = -\nabla g(x, y) = -(2x, 2y)$ donne bien l'inégalité de qualification. Donc tous les points de K sont qualifiés.

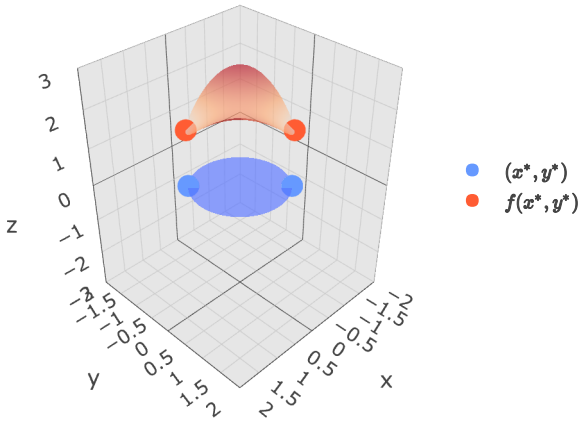

FIGURE II.2 – Illustration du problème $\min_{x^2+y^2\leq 1} 2 + xy$.

Le théorème 3.6 permet alors d'affirmer que si (x^*, y^*) est un minimum de f sur K, alors il existe $\lambda \geq 0$ tel que

$$\nabla f(x^*, y^*) + \lambda \nabla g(x^*, y^*) = 0 \quad \text{et} \quad \lambda g(x^*, y^*) = 0$$

c'est-à-dire que

$$\begin{cases} y^* + 2\lambda x^* = 0 \\ x^* + 2\lambda y^* = 0 \\ \lambda\left((x^*)^2 + (y^*)^2 - 1\right) = 0 \end{cases}$$

Ce système a cinq solutions possibles : $(\lambda = 0,\ x^* = y^* = 0)$, $(\lambda = 1/2,\ x^* = -y^* = \sqrt{2}/2)$, $(\lambda = 1/2,\ x^* = -y^* = -\sqrt{2}/2)$, $(\lambda = -1/2,\ x^* = y^* = \sqrt{2}/2)$, $(\lambda = -1/2,\ x^* = y^* = -\sqrt{2}/2)$. Les deux dernières solutions ne sont pas admissibles car le multiplicateur associé λ est strictement négatif. Reste les trois premières. On calcule alors le critère en ces points :

$$f(\sqrt{2}/2, -\sqrt{2}/2) = f(-\sqrt{2}/2, \sqrt{2}/2) = 3/2 \text{ et } f(0,0) = 2.$$

Donc il y a donc deux solutions : $(\sqrt{2}/2, -\sqrt{2}/2)$ et $(-\sqrt{2}/2, \sqrt{2}/2)$. Une illustration de cet exemple est donnée en figure 2.

3. **Plusieurs contraintes d'égalité :** On considère le problème suivant dans $\mathbb{R}^3$:

$$\begin{array}{ll} \min & x + z \\ x^2 + y^2 = 1 & \\ y^2 + z^2 = 4 & \end{array}$$

Ici, $f(x, y, z) = x+z$ et $K = \{(x, y, z) \ : \ H(x, y, z) = 0\}$, avec $H(x, y, z) = (h_1(x, y, z), h_2(x, y, z))$ et $h_1(x, y, z) = x^2 + y^2 - 1$, $h_2(x, y, z) = y^2 + z^2 - 4$.

Le théorème de Weierstrass permet à nouveau d'assurer l'existence d'un minimum car K est compact (fermé et contenu dans la boule $B(0,2)$ de $\mathbb{R}^3$), non vide (car $(1,0,2) \in K$) et la fonction f est continue.

Les points $(x,y,z) \in K$ où la contrainte est qualifiée sont ceux pour lesquels la famille $\{\nabla h_1(x,y,z), \nabla h_2(x,y,z)\}$ est libre. Or

$$\nabla h_1(x,y,z) = \begin{pmatrix} 2x \\ 2y \\ 0 \end{pmatrix} \ , \ \nabla h_2(x,y,z) = \begin{pmatrix} 0 \\ 2y \\ 2z \end{pmatrix}$$

Ces deux vecteurs sont liés, si et seulement si au moins deux coordonnées x, y, z sont nulles. Prenons $x = z = 0$ et $y \neq 0$, qui correspond à l'ensemble de points de la forme $(0,y,0)$. Mais ces points n'appartiennent pas à K puisque sinon on aurait $y^2 = 1$ et $y^2 = 4$. Donc le système de vecteurs est toujours libre pour tout $(x,y,z) \in K$ et la contrainte est qualifiée. Le théorème 3.2 permet d'affirmer que si (x^*, y^*) est un minimum de f sur K, alors il existe μ_1, $\mu_2 \in \mathbb{R}$ tels que

$$\nabla f(x^*,y^*,z^*) + \mu_1 \nabla h_1(x^*,y^*,z^*) + \mu_2 \nabla h_2(x^*,y^*,z^*) = 0 \quad \Leftrightarrow \quad \begin{cases} 1 + 2\mu_1 x^* = 0 \\ 2\mu_1 y^* + 2\mu_2 y^* = 0 \\ 1 + 2\mu_2 z^* = 0 \\ (x^*)^2 + (y^*)^2 = 1 \\ (y^*)^2 + (z^*)^2 = 4 \end{cases}$$

Noter que μ_1 et μ_2 sont non nuls et que

$$x^* = -1/(2\mu_1) \text{ et } z^* = -1/(2\mu_2)$$

D'autre part, soit $\mu_1 + \mu_2 = 0$, soit $y^* = 0$. Le premier cas est impossible, car alors $x^* = -z^*$ et

$$(y^*)^2 = 1 - (x^*)^2 = 1 - (z^*)^2 = 4 - (z^*)^2,$$

ce qui conduit à $1 = 4$ et on aboutit à une contradiction. Donc $y^* = 0$, ce qui impose $x^* = \pm 1$ et $z^* = \pm 2$. On voit alors facilement que le minimum du problème est $(-1,0,-2)$.

4. **Plusieurs contraintes d'inégalité :** On considère le problème suivant dans $\mathbb{R}^3$:

$$\begin{array}{ll} \min & x+y+z \\ x^2+y^2+z^2 \leq 1 & \\ x \geq 0 & \end{array}$$

Ici, $f(x,y,z) = x+y+z$ et $K = \{(x,y,z) \ : \ G(x,y,z) \leq 0\}$, avec $G(x,y,z) = (g_1(x,y,z), g_2(x,y,z))$ et $g_1(x,y,z) = x^2+y^2+z^2-1$, $g_2(x,y,z) = -x$.

La contrainte K étant compacte et non vide (car $(0,0,0) \in K$) et f étant continue, le problème admet une solution (x^*, y^*, z^*).

Montrons que la contrainte est qualifiée en tout point. Soit (x,y,z) appartenant au bord de la contrainte. Si $I(x,y,z) = \{1\}$, alors $\nabla g_1(x,y,z)$ est non nul car $x^2+y^2+z^2-1 = 0$ et $\nabla g_1(x,y,z)$ ne s'annule que pour $x = y = z = 0$. Si $I(x,y,z) = \{2\}$, alors $\nabla g_2(x,y,z) \neq 0$. Supposons maintenant que $I(x,y,z) = \{1,2\}$. Choisissons $v = (1,-y,-z)$. Alors

$$\langle \nabla g_1(x,y,z), v \rangle = -2y^2 - 2z^2 = -2 < 0 \text{ et } \langle \nabla g_2(x,y,z), v \rangle = -1 < 0 \ .$$

Donc la contrainte est qualifiée en tout point. Au point (x^*, y^*, z^*), la condition nécessaire suivante est satisfaite : il existe $\lambda_1 \geq 0$ et $\lambda_2 \geq 0$ tels que

$$\begin{cases} 1 + 2\lambda_1 x^* - \lambda_2 = 0 \\ 1 + 2\lambda_1 y^* = 0 \\ 1 + 2\lambda_1 z^* = 0 \end{cases}$$

avec la condition d'exclusion

$$\begin{cases} \lambda_1 \left((x^*)^2 + (y^*)^2 + (z^*)^2 - 1 \right) & = 0 \\ \lambda_2 (-x^*) & = 0. \end{cases}$$

D'après les dernières équations, on a $\lambda_1 > 0$. Donc $(x^*)^2 + (y^*)^2 + (z^*)^2 - 1 = 0$, et $y^* = z^* = -1/(2\lambda_1)$. Supposons $x^* > 0$. Alors $\lambda_2 = 0$ et $x^* = -1/(2\lambda_1)$, ce qui est impossible car $x^* > 0$ par hypothèse. Donc $x^* = 0$, ce qui impose

$$0 + (-1/(2\lambda_1))^2 + (-1/(2\lambda_1))^2 = 1$$

c'est-à-dire $\lambda_1 = \sqrt{2}/2$ car $\lambda_1 > 0$. D'où $x^* = 0$ et $y^* = z^* = -\sqrt{2}/2$. Noter enfin que $\lambda_2 = 1 > 0$.

5. **Un peu de tout :** On mélange maintenant les difficultés. Soit le problème dans $\mathbb{R}^3$:

$$\begin{array}{ll} \min & x + 2y + 3z \\ x^2 + y^2 + z^2 = 1 & \\ x + y + z \leq 0 & \end{array}$$

La contrainte est compacte, non vide (car $(-1, 0, 0) \in K$) et le critère continu, donc le problème admet au moins une solution. Le critère est $f(x, y, z) = x + 2y + 3z$, il y a une contrainte d'égalité $h(x, y, z) = x^2 + y^2 + z^2 - 1 = 0$ et une contrainte d'inégalité $g(x, y, z) = x + y + z \leq 0$. Vérifions que la contrainte est qualifiée. Le premier point de la définition 3.8 est satisfaite car $\nabla h(x, y, z) = 0$ si et seulement si $x = y = z = 0$, ce qui est impossible car $h(x, y, z) = 0$. D'autre part, pour un point $(x, y, z) \in K$ tel que $g(x, y, z) = 0$, si on prend $v = (-1, -1, -1)$, on a

$$\langle \nabla h(x, y, z), v \rangle = -2x - 2y - 2z = -2(x + y + z) = 0 \text{ et } \langle \nabla g(x, y, z), v \rangle = -3 < 0.$$

Donc la contrainte K est qualifiée en tout point. Si (x^*, y^*, z^*) est un minimum du problème, les conditions de Kuhn & Tucker s'écrivent

$$\begin{cases} 1 + \lambda + 2\mu x^* = 0 \\ 2 + \lambda + 2\mu y^* = 0 \\ 3 + \lambda + 2\mu z^* = 0 \\ (x^*)^2 + (y^*)^2 + (z^*)^2 = 1 \\ \lambda(x^* + y^* + z^*) = 0 \end{cases}$$

où $\lambda \geq 0$ et $\mu \in \mathbb{R}$. On déduit des trois premières équations que μ est non nul et que

$$x^* = -\frac{1 + \lambda}{2\mu}, \quad y^* = -\frac{2 + \lambda}{2\mu}, \quad z^* = -\frac{3 + \lambda}{2\mu},$$

de sorte que, d'après la condition d'exclusion, on a

$$\lambda\left(\frac{1+\lambda}{2\mu}+\frac{2+\lambda}{2\mu}+\frac{3+\lambda}{2\mu}\right)=0$$

D'où $\lambda(6+3\lambda)=0$. Comme $\lambda \geq 0$, cela impose $\lambda = 0$. On reporte alors les expressions de x^*, y^* et z^* en fonction de μ dans la contrainte d'égalité pour trouver $\mu = \sqrt{14}/2$ ou $\mu = -\sqrt{14}/2$. Il y a donc deux points vérifiant les conditions de Kuhn & Tucker. On voit facilement que la solution est $\frac{-1}{\sqrt{14}}(1,2,3)$.

4 Dualité

4.1 Introduction

Dans toute cette partie, nous nous concentrons sur le cas où K ne contient que des contraintes d'inégalité. Dans ce cas, le lagrangien est défini comme la fonction

$$\begin{aligned}\mathcal{L} : \mathbb{R}^n \times (\mathbb{R}_+)^m &\to \mathbb{R} \\ (x,\lambda) &\mapsto \mathcal{L}(x,\lambda) \coloneqq \mathcal{L}(x,\lambda) \coloneqq f(x^*) + \sum_{i=1}^{m} \lambda_i g_i(x^*)\end{aligned}$$

Si x^* est minimum local sur K et la contrainte est qualifiée, alors le théorème de Kuhn et Tucker 3.6 assure l'existence de multiplicateurs de Lagrange $\lambda_1, \ldots, \lambda_m \geq 0$ tels que

$$\nabla_x \mathcal{L}(x^*, \lambda) = 0$$

De plus,

$$\nabla_\lambda \mathcal{L}(x^*, \lambda) = G(x^*) \leq 0.$$

Le couple (x^*, λ) est presque un point stationnaire du lagrangien. Il ne l'est pas entièrement car $\nabla_\lambda \mathcal{L}(x^*, \lambda)$ n'est pas nécessairement nul. Dans cette section, nous allons voir que (x^*, λ) est en réalité un point selle du lagrangien. Cette vision du minimum en tant que point selle de $\mathcal{L}$ permettra essentiellement deux choses :

1. Trouver une preuve alternative au fait que, dans le cas convexe, les conditions KKT sont aussi suffisantes (preuve alternative au théorème 3.11 de la section 3.4).
2. Reformuler le problème de minimisation sous une approche dite duale qui est en général bien adaptée pour la résolution numérique des problèmes de minimisation.

4.2 Théorie générale du point selle et de la dualité

Dans cette section, nous nous plaçons dans un cadre abstrait pour introduire la théorie générale du point selle. La section suivante montrera comment l'appliquer pour résoudre le problème de minimisation dans le cas convexe sans contraintes d'égalité.

Soient $U \subset \mathbb{R}^n$ et $P \subset (\mathbb{R}_+)^m$ deux ensembles et soit $L : U \times P \to \mathbb{R}$ une application.

Définition 4.1. On dit que le couple $(u^*, p^*) \in U \times P$ est un point selle de L sur $U \times P$ si et seulement si

$$L(u^*, p) \leq L(u^*, p^*) \leq L(u, p^*), \quad \forall (u,p) \in U \times P. \tag{4.1}$$

Autrement dit, u^* est un minimum de la fonction $u \in U \to L(u, p^*)$ et p^* est un maximum de la fonction $p \in P \to L(u^*, p)$.

Nous introduisons maintenant la notion de problème primal et dual. Pour cela, pour $u \in U$ et $p \in P$, soient les fonctions

$$\mathcal{J}(u) \coloneqq \sup_{p \in P} L(u, p), \qquad \mathcal{G}(p) \coloneqq \inf_{u \in U} L(u, p). \tag{4.2}$$

Définition 4.2. On appelle problème primal le problème de minimisation

$$\inf_{u \in U} \mathcal{J}(u) = \inf_{u \in U} \sup_{p \in P} L(u, p)$$

et problème dual le problème de maximation

$$\sup_{p \in P} \mathcal{G}(p) = \sup_{p \in P} \inf_{u \in U} L(u, p).$$

Le résultat suivant montre que les problèmes primal et dual sont intimement liés au point selle (u^*, p^*).

Théorème 4.3 (Théorème de dualité)**.** *Le couple (u^*, p^*) est un point selle de L sur $U \times P$ si et seulement si*

$$L(u^*, p^*) = \mathcal{J}(u^*) = \min_{u \in U} \mathcal{J}(u) = \max_{p \in P} \mathcal{G}(p) = \mathcal{G}(p^*).$$

Preuve. Prouvons d'abord l'implication directe. Soit (u^*, p^*) un point selle de L sur $U \times P$ et soit

$$L^* = L(u^*, p^*).$$

Nous allons montrer que

$$\min_{u \in U} \mathcal{J}(u) = \mathcal{J}(u^*) = L^* \tag{4.3}$$

et que

$$\max_{p \in P} \mathcal{G}(p) = \mathcal{G}(p^*) = L^*. \tag{4.4}$$

Pour tout $u \in U$, par la définition (4.2) de $\mathcal{J}(u)$, nous avons $\mathcal{J}(u) \geq L(u, p)$ pour tout $p \in P$, donc en particulier pour $p = p^*$. Alors $\mathcal{J}(u) \geq L(u, p^*)$. Or $L(u, p^*) \geq \inf_{u \in U} L(u, p^*) = L(u^*, p^*) = L^*$ par la définition (4.1) du point selle. De plus, $L(u^*, p^*) = \sup_{p \in P} L(u^*, p) = \mathcal{J}(u^*)$ par la définition (4.2) de $\mathcal{J}(u^*)$. De cette chaîne de minorations, il vient que $\mathcal{J}(u) \geq \mathcal{J}(u^*)$, ce qui prouve (4.3). Par une démarche similaire, on prouve (4.4).

Pour prouver l'implication réciproque, en utilisant (4.2), on a pour tout $(u, p) \in U \times P$, $L^* = L(u^*, p^*) = \mathcal{J}(u^*) \geq L(u^*, p)$ et $L^* = \mathcal{G}(p^*) \leq L(u, p^*)$. Par conséquent $L(u^*, p) \leq L(u^*, p^*) \leq L(u, p^*)$ pour tout $(v, p) \in U \times P$, ce qui veut dire que (u^*, p^*) est un point selle de L sur $U \times P$. □

4.3 Application de la théorie du point selle à l'optimisation

Nous allons montrer le lien étroit entre la notion de point selle et les problèmes de minimisation avec contraintes d'inégalité

$$\min_{x\in K} f(x), \quad K = \{x \in \mathbb{R}^n \,:\, G(x) \leq 0\}. \tag{4.5}$$

Les domaines abstraits U et P de la section précédente vont être particulariés au cas où $U = \Omega$ un ouvert de $\mathbb{R}^n$ contenant K et $P = (\mathbb{R}_+)^m$ est le domaine de définition des multiplicateurs de Lagrange. La fonction L considérée par la suite est le lagrangien associé au problème (4.5), c'est à dire,

$$\mathcal{L}(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x), \quad \forall (x,\lambda) \in \Omega \times (\mathbb{R}_+)^m.$$

Le résultat suivant montre que s'il existe un point selle (x^*, λ^*) du lagrangien, alors x^* est un **minimum global** au problème (4.5).

Proposition 4.4. *Soit Ω un ouvert contenant K et soit le lagrangien*

$$\mathcal{L}(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x), \quad \forall (x,\lambda) \in \Omega \times (\mathbb{R}_+)^m.$$

Si (x^*, λ^*) est un point selle de $\mathcal{L}$ sur $\Omega \times P$, alors $x^* \in K$ et x^* est un minimum global de f sur K. De plus, si f et les g_i sont $\mathcal{C}^1$, alors

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) = 0.$$

Preuve. La condition de point selle de traduit par

$$f(x^*) + \lambda \cdot G(x^*) \leq f(x^*) + \lambda^* \cdot G(x^*) \leq f(x) + \lambda^* \cdot G(x), \quad \forall (x,\lambda) \in \Omega \times P.$$

La première inégalité entraine que $G(x^*) \cdot (\lambda - \lambda^*) \leq 0$ pour tout $\lambda \in P$. En prenant $\lambda = \lambda^* + e_i$ avec e_i le i-ème vecteur canonique de $\mathbb{R}^m$, on a $G(x^*) \cdot (\lambda - \lambda^*) = g_i(x^*) \leq 0$ donc $x^* \in K$.

De plus, en prenant $\lambda = 0$ et puis $\lambda = 2\lambda^*$, on en déduit que $G(x^*) \cdot \lambda^* = 0$. Par conséquent, la deuxième inégalité devient $f(x^*) \leq f(x) + \lambda^* \cdot G(x)$ pour tout $x \in \Omega$. Mais comme λ^* a toutes ses composantes positives vu que c'est un élément de P et que $G(x) \leq 0$, alors $\lambda^* \cdot G(x) \leq 0$. Par conséquent $f(x^*) \leq f(x)$ pour tout $x \in \Omega$ ce qui veut dire que x^* est un minimum global de f sur Ω (donc sur K aussi).

Finalement, si f et les g_i sont $\mathcal{C}^1$, la deuxième inégalité du point selle montre que x^* est un minimum de la fonction $x \to f(x) + \lambda^* \cdot G(x)$ définie sur l'ouvert Ω. Par conséquent, l'équation d'Euler (théorème 2.4) assure que le gradient de cette fonction au point x^* s'annule, c'est à dire, $\nabla f(x^*) + \sum_{i=1}^m \lambda_i^* g_i(x^*) = 0$. □

Nous utilisons maintenant la théorie du point selle pour montrer que les conditions KKT sont nécessaires et suffisantes dans la cas convexe (sous condition de qualification pour la condition nécessaire). Il s'agit du même résultat déjà donné dans le théorème 3.11 de la section 3.4 mais prouvé par des arguments de dualité.

Théorème 4.5 (CNS cas convexe)**.** *On suppose que f et les contraintes g_i sont convexes et de classe $\mathcal{C}^1$ sur l'ensemble convexe $K = \{x \in \mathbb{R}^n : g_i(x) \leq 0,\ i = 1, \ldots, m\}$. Supposons aussi que les contraintes soient qualifiées au point $x^* \in K$ au sens de la définition 3.4. Alors,*

$$x^* \text{ est un minimum global de } f \text{ sur } K$$
$$\Longleftrightarrow$$
$$\text{Il existe } \lambda^* \in (\mathbb{R}_+)^m \text{ tel que } (x^*, \lambda^*) \text{ est un point selle du lagrangien } \mathcal{L} \text{ sur } \mathbb{R}^n \times (\mathbb{R}_+)^m$$
$$\Longleftrightarrow$$

$$\exists \lambda^* \in (\mathbb{R}_+)^m, \quad \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) = 0, \quad G(x^*) \leq 0, \quad \lambda^* \geq 0, \quad \lambda^* \cdot G(x^*) = 0. \tag{4.6}$$

Preuve. La démarche est la suivante :

$$\begin{aligned} &x^* \text{ est un minimum de } f \text{ sur } K \\ &\xRightarrow[\textit{Th. } 3.6]{} \text{Conditions KKT (4.6)} \\ &\xRightarrow[\text{Convexité}]{} \exists \lambda^* \in (\mathbb{R}_+)^m \text{ tel que } (x^*, \lambda^*) \text{ point selle de } \mathcal{L} \text{ sur } \mathbb{R}^n \times (\mathbb{R}_+)^m \\ &\xRightarrow[\textit{Prop. } 4.4]{} x^* \text{ est un minimum de } f \text{ sur } K \end{aligned}$$

Toutes les étapes sont justifiées par des résultats précédents, sauf le passage des conditions KKT à l'existence du point selle que nous détaillons maintenant. Soit le lagrangien

$$\mathcal{L}(x, \lambda) = f(x) + \lambda \cdot G(x), \quad \forall (x, \lambda) \in \mathbb{R}^n \times (\mathbb{R}_+)^m.$$

Pour $\lambda \in (\mathbb{R}_+)^m$ fixé, la fonction $x \mapsto \mathcal{L}(x, \lambda)$ est convexe sur $\mathbb{R}^n$ car f et les g_i sont convexes. Donc, comme $\nabla_x \mathcal{L}(x^*, \lambda^*) = 0$ par la condition KKT, le théorème 2.5 garantit que x^* est minimum global de $x \mapsto \mathcal{L}(x, \lambda^*)$ sur $\mathbb{R}^n$ et donc $\mathcal{L}(x^*, \lambda^*) \leq \mathcal{L}(x, \lambda^*)$. De plus $\mathcal{L}(x^*, \lambda^*) \geq \mathcal{L}(x^*, \lambda)$ pour tout $\lambda \in (\mathbb{R}_+)^m$ car $\lambda^* \cdot G(x^*) = 0$ par KKT et $\lambda \cdot G(x^*) \leq 0$ pour tout $\lambda \in (\mathbb{R}_+)^m$. Par conséquent, pour tout $(x, \lambda) \in \mathbb{R}^n \times (\mathbb{R}_+)^m$, $\mathcal{L}(x^*, \lambda) \leq \mathcal{L}(x^*, \lambda^*) \leq \mathcal{L}(x, \lambda^*)$ et (x^*, λ^*) est bien un point selle de $\mathcal{L}$. □

Nous allons à présent utiliser le théorème de dualité pour construire des méthodes pratiques de résolution du problème

$$\min_{x \in \mathbb{R}^n,\ G(x) \leq 0} f(x)$$

avec f et g_i convexes. On considère à nouveau le lagrangien

$$\mathcal{L}(x, \lambda) \coloneqq f(x) + \lambda \cdot G(x), \quad \forall (x, \lambda) \in \mathbb{R}^n \times (\mathbb{R}_+)^m.$$

Nous avons

$$\mathcal{J}(x) = \sup_{\lambda \in (\mathbb{R}_+)^m} \mathcal{L}(x, \lambda) = \begin{cases} f(x) \text{ si } G(x) \leq 0 \\ +\infty \text{ sinon.} \end{cases} \tag{4.7}$$

Donc le problème primal primal associé à $\mathcal{L}$ est exactement le problème d'origine,

$$\inf_{x \in \mathbb{R}^n} \mathcal{J}(x) = \min_{x \in \mathbb{R}^n,\ G(x) \leq 0} f(x).$$

Pour le problème dual, nous avons

$$\mathcal{G}(\lambda) = \inf_{x \in \mathbb{R}^n} f(x) + \lambda \cdot G(x). \tag{4.8}$$

On peut montrer que la fonction $\mathcal{G}$ est concave donc le problème dual peut s'exprimer comme un problème de minimisation convexe en remarquant que

$$\sup_{\lambda \in (\mathbb{R}_+)^m} \mathcal{G}(\lambda) = - \inf_{\lambda \in (\mathbb{R}_+)^m} -\mathcal{G}(\lambda).$$

Les contraintes du problème dual étant linéaires, elles sont en général beaucoup plus simples que celles du problème primal. Cela est utilisé de façon cruciale dans certains algorithmes que nous allons détailler dans la section suivante.

D'un point de vue pratique, on utilise la dualité de la manière suivante :

Corollaire 4.6 (Méthode de dualité)**.** *On suppose que f et les contraintes g_i sont convexes et de classe $\mathcal{C}^1$. Supposons que les contraintes soient qualifiées en tout point et que le problème primal admet au moins une solution. Alors le problème dual admet une solution λ^* et la fonction $x \to \mathcal{L}(x, \lambda^*)$ admet pour minimum sur $\mathbb{R}^n$ n'importe quel minimum de f sur K (et λ^* est un multiplicateur associé).*

Pour trouver un minimum de f sur K **par dualité**, il suffit donc :

1. De calculer, pour tout $\lambda \in (\mathbb{R}_+)^m$, $\mathcal{G}(\lambda) := \min_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda)$; soit $x(\lambda)$ un point de minimum ;
2. De résoudre le problème dual : $\max_{\lambda \in (\mathbb{R}_+)^m} \mathcal{G}(\lambda)$;
3. Alors tout minimum de f sur K se trouve parmi le ou les minima de $\mathcal{L}(\cdot, \lambda^*)$ sur $\mathbb{R}^n$. *En particulier, si f est strictement convexe, alors $\mathcal{L}(\cdot, \lambda^*)$ a un unique minimum sur $\mathbb{R}^n$ et ce minimum est le minimum de f sur K.*

Preuve. On sait que, comme le problème primal admet une solution x^*, alors il existe $\lambda^* \in (\mathbb{R}_+)^I$ tel que (x^*, λ^*) est un point selle de $\mathcal{L}$ sur $\mathbb{R}^d \times (\mathbb{R}_+)^I$. Donc le problème dual admet comme solution λ^*.

Supposons maintenant que λ^* est un maximum du problème dual. On va montrer que toute solution du primal est un minimum de $\mathcal{L}(\cdot, \lambda^*)$. Soit x^* une solution du problème primal et $\lambda \in (\mathbb{R}_+)^I$ un multiplicateur associé. Alors, (x^*, λ) est un point selle de $\mathcal{L}$, donc x^* est un minimum de $\mathcal{L}(\cdot, \lambda)$ sur $\mathbb{R}^d$. Alors, pour tout $x \in \mathbb{R}^d$, on a :

$$\begin{aligned}
\mathcal{L}(x, \lambda^*) &\geq \mathcal{G}(\lambda^*) && \text{(par déf de } \mathcal{G}) \\
&\geq \mathcal{G}(\lambda) && \text{(par optimalité de } \lambda^*) \\
&= \mathcal{L}(x^*, \lambda) = f(x^*) && \text{(par complémentarité)} \\
&\geq f(x^*) + \sum_i \lambda_i^* g_i(x^*) = \mathcal{L}(x^*, \lambda^*).
\end{aligned}$$

Donc x^* est un point de minimum de $\mathcal{L}(\cdot, \lambda^*)$. □

Exemple : Pour illustrer l'intérêt du corollaire 4.6 et de l'intérêt de l'approche par dualité, considérons le problème de minimisation quadratique

$$\min_{x\in\mathbb{R}^n,\, G(x)=Bx-c\leq 0} \left\{ f(x) = \frac{1}{2}Ax\cdot x - b\cdot x \right\}, \tag{4.9}$$

où $A \in \mathcal{M}_n(\mathbb{R})$ est une matrice carrée symétrique définie positive, $b \in \mathbb{R}^n$, $B \in \mathcal{M}_{m,n}(\mathbb{R})$ et $c \in \mathbb{R}^m$. Le lagrangien est

$$\mathcal{L}(x,\lambda) = \frac{1}{2}Ax\cdot x - b\cdot x + \lambda\cdot(Bx-c), \quad \forall (x,\lambda)\in\mathbb{R}^n\times(\mathbb{R}_+)^m.$$

Nous avons déjà observé que le problème primal associé à $\mathcal{L}$ est le problème d'origine donc examinons le problème dual. Pour $\lambda \in (\mathbb{R}_+)^m$ fixé, le problème $\min_{x\in\mathbb{R}^n}\mathcal{L}(x,\lambda)$ admet une unique solution car $x \to \mathcal{L}(x,\lambda)$ est fortement convexe (cf. discussion autour de la formule (1.1)). Cette solution satisfait

$$\nabla_x\mathcal{L}(x,\lambda) = Ax - b + B^T\lambda = 0 \quad \Leftrightarrow \quad x = A^{-1}(b - B^T\lambda)$$

On obtient donc

$$\mathcal{G}(\lambda) = \mathcal{L}\left(A^{-1}(b-B^T\lambda),\lambda\right)$$

et le problème dual s'écrit

$$\sup_{\lambda\geq 0}\left(-\frac{1}{2}\lambda^T BA^{-1}B^T\lambda + (BA^{-1}b - c)^T\lambda - \frac{1}{2}b^TA^{-1}b\right).$$

La fonctionnelle à maximiser n'a pas une forme particulièrement facile mais l'essentiel est qu'il s'agit d'une forme quadratique et les contraintes sont linéaires. Le premier point du corollaire 4.6 assure que le problème admet une solution mais elle n'est pas forcément unique à moins que B soit de rang m. Dans ce cas, la matrice $BA^{-1}B^T$ est définie positive et donc on a unicité. L'avantage du problème dual provient du fait que les contraintes sont exprimées de façon plus simple que dans le problème primal. Dans la section suivante, nous allons utiliser ce point de façon cruciale pour construire des algorithmes de résolution.

5 Méthodes numériques

Dans cette section, nous étudions quelques algorithmes permettant de calculer une solution approchée au problème d'optimisation

$$\min_{x\in K} f(x), \tag{5.1}$$

où K est convexe et f est fortement convexe et $\mathcal{C}^1$ sur K. Avec ces hypothèses, le théorème 2.7 assure l'existence et l'unicité d'un minimum $x^* \in K$ au problème et nous avons la caractérisation

$$\langle \nabla f(x^*), v - x^*\rangle \geq 0, \quad \forall v \in K. \tag{5.2}$$

Les algorithmes présentés sont itératifs, c'est à dire que partant d'un point initial x_0, nous construisons une suite $(x_n)_{n\in\mathbb{N}}$ qui converge vers x^*. Nous donnerons aussi quelques résultats quant au taux de convergence. Comme les algorithmes reposent sur la projection de points de $\mathbb{R}^n$ sur l'ensemble K, nous commençons la section en rappelant cette notion.

L'hypothèse de forte convexité pour la fonction f est forte, mais nous verrons qu'elle joue un rôle crucial dans les preuves de convergence. La minimisation de fonctions qui ne sont ni strictement convexes, ni différentiables est possible. Il s'agit d'un sujet de recherche actuel très actif dont la présentation dépasse malheureusement les contraintes en temps de ce cours. La minimisation de fonctions non convexes est aussi possible sous certaines conditions mais elle constitue aussi un sujet de recherche actuel. Le principal défi est de trouver dans ce cas des algorithmes qui convergent vers le minimum global et ne stagnent pas dans des minima locaux.

5.1 Projection sur un ensemble convexe fermé

Théorème 5.1. *Soit K un ensemble convexe fermé de $\mathbb{R}^n$. Pour tout $x \in \mathbb{R}^n$, il existe un unique $x_K \in K$ tel que*

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

De façon équivalente, x_K se caractérise par

$$x_K \in K, \quad \langle x_K - x, x_K - y \rangle \leq 0, \quad \forall y \in K. \tag{5.3}$$

Le vecteur x_K est appelé la projection orthogonale de x sur K.

Preuve. Comme la fonction $v \to \|x - y\|^2$ est strictement convexe, le problème $\min_{y \in K} \|x - y\|^2$ admet une unique solution en vertu du théorème 2.7. Montrons que x_K vérifie la relation énoncée. En effet, pour tout $y \in K$, le point $ty + (1-t)x_K$ appartient à K si $t \in [0, 1]$. Donc

$$\|x - x_K\|^2 \leq \|x - (ty + (1-t)x_K)\|^2 = \|x - x_K\|^2 - 2t\langle x - x_K, y - x_K \rangle + t^2 \|y - x_K\|^2$$

En retranchant $\|x - x_K\|^2$ des deux côtés, puis en divisant par $t > 0$ et faisant tendre t vers 0^+, nous obtenons l'inégalité désirée,

$$\langle x - x_K, y - x_K \rangle \leq 0.$$

Réciproquement, montrons que si un point $u \in K$ vérifie l'inégalité

$$\forall y \in K, \ \langle x - u, y - u \rangle \leq 0$$

alors $u = x_K$. Comme

$$\forall y \in K, \ \|x - y\|^2 - \|x - u\|^2 = \|x - u + u - y\|^2 - \|x - u\|^2 = 2\langle x - u, u - y \rangle + \|u - y\|^2 \geq 0,$$

alors $u \in K$ est l'unique minimum de la fonction $x \to \|x - y\|^2$. Il s'agit donc bien du projeté orthogonal de x sur K. □

Remarque 5.2. La caractérisation (5.3) du projeté se traduit visuellement par le fait que les directions $x_K - x$ et $x_K - y$ forment un angle obtus pour tous les éléments $y \in K$.

Le théorème 5.1 permet de définir l'opérateur de projection $P_K : \mathbb{R}^n \to K$ défini pour tout $x \in \mathbb{R}^n$ par $P_K(x) = x_K$. Cet opérateur est faiblement contractant au sens où

$$\|P_K(x) - P_K(y)\| \leq \|x - y\|, \quad \forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n.$$

Preuve. On a

$$\langle x - P_K(x), P_K(y) - P_K(x)\rangle \leq 0$$

et

$$\langle P_K(y) - y, P_K(y) - P_K(x)\rangle \leq 0$$

On additionne pour obtenir

$$\langle x - y - (P_K(x) - P_K(y)), P_K(y) - P_K(x)\rangle \leq 0$$

c'est-à-dire

$$\|P_K(y) - P_K(x)\|^2 \ \leq \ \langle x - y; P_K(x) - P_K(y)\rangle \ \leq \ \|x - y\|\|P_K(x) - P_K(y)\|.$$

En divisant par $\|P_K(y) - P_K(x)\|$ on obtient l'inégalité désirée (le cas $\|P_K(y) - P_K(x)\| = 0$ est trivialement vérifié). □

Remarque 5.3. Dans le cas particulier où K est un sous-espace vectoriel de $\mathbb{R}^n$, la caractérisation (5.3) devient

$$x_K \in K, \quad \langle x_K - x, y\rangle = 0, \quad \forall y \in K.$$

Cette relation découle de (5.3) en prenant $z = x_K \pm y$ avec $z \in K$.

5.2 Algorithme du gradient projeté à pas fixe

L'algorithme de base pour résoudre numériquement (5.1) part de l'observation suivante. Pour tout $\mu > 0$, il découle de (5.2) que

$$\langle x^* - (x^* - \mu\nabla f(x^*)), v - x^*\rangle \geq 0, \quad \forall v \in K.$$

Donc, par la caractérisation (5.3) de la projection orthogonale, nous avons

$$x^* = P_K(x^* - \mu\nabla f(x^*)).$$

Donc x^* est un point fixe de la fonction $v \to P_K(v - \mu\nabla f(v))$. L'algorithme du gradient projeté à pas fixe est donc défini comme les itérations de point fixe suivantes.

Gradient projeté à pas fixe avec $\mu > 0$ fixé :

$$\begin{cases} x_{n+1} = P_K(x_n - \mu\nabla f(x_n)) \\ x_0 \text{ donné} \end{cases} \tag{5.4}$$

Le résultat suivant prouve la convergence de cet algorithme vers x^*.

Théorème 5.4. *Si f est α-fortement convexe et si ∇f est Lipschitzienne de constante C, alors, si $0 < \mu < 2\alpha/C^2$, l'algorithme de gradient à pas fixe converge : pour tout $x_0 \in K$, la suite (x_n) converge vers le minimum x^*.*

Preuve. Pour tout $n \geq 0$,

$$\begin{aligned}
\|x^* - x_{n+1}\|^2 &= \|P_K(x^* - \mu\nabla f(x^*)) - P_K(x_n - \mu\nabla f(x_n))\|^2 \\
&\leq \|x^* - x_n - \mu\left(\nabla f(x^*) - \nabla f(x_n)\right)\|^2 \qquad (\text{car } P_K \text{ faiblement contractant}) \\
&= \|x^* - x_n\|^2 + \mu^2\|\nabla f(x^*) - \nabla f(x_n)\|^2 - 2\mu\left\langle x^* - x_n, \nabla f(x^*) - \nabla f(x_n)\right\rangle \\
&\leq (1 + \mu^2 C^2 - 2\mu\alpha)\|x^* - x_n\|^2 \qquad (\text{car } f \ \alpha\text{-conv et } \nabla f \text{ Lip.})
\end{aligned}$$

Comme $\mu < 2\alpha/C^2$, alors $(1 + \mu^2C^2 - 2\mu\alpha) < 1$, et par conséquent $x_n \to_{n\to\infty} x^*$. □

5.3 Algorithme d'Uzawa

L'algorithme précédent n'est pas facilement implémentable en pratique car en général l'opérateur de projection P_K n'est pas connu explicitement : la projection d'un élément $x \in \mathbb{R}^n$ dans un convexe fermé K quelconque peut être très difficile à déterminer. Il existe néanmoins une exception importante qui est le cas des ensembles de la forme

$$K = \prod_{i=1}^{n}[a_i, b_i], \tag{5.5}$$

avec potentiellement des $a_i = -\infty$ et des $b_i = +\infty$ pour certains indices i. Dans ce cas, on vérifie aisément que si $x = (x_1, \ldots, x_n)^T$, alors $y = P_K(x)$ a pour composantes

$$y_i = \min(\max(a_i, x_i), b_i) = \begin{cases} a_i \text{ si } x_i \leq a_i \\ x_i \text{ si } a_i \leq x_i \leq b_i \\ b_i \text{ si } x_i \geq b_i \end{cases}, \quad 1 \leq i \leq n.$$

En d'autres mots, dans ce cas il suffit de "tronquer" les composantes de x pour obtenir $P_K(x)$. Cette simple observation peut se combiner avec les résultats de dualité pour construire un nouvel algorithme beaucoup plus facile à implémenter en pratique : l'algorithme d'Uzawa. En effet, lorsque la projection P_K n'est pas explicite, le problème dual est souvent posé sur un ensemble de la forme (5.5), typiquement sur $(\mathbb{R}_+)^m$. Dans ce cas, le problème dual peut se calculer avec l'algorithme du gradient projeté à pas fixe et le problème primal peut se trouver par résolution d'un problème de minimisation sans contraintes.

Sous les hyothèses du corollaire 4.6, la solution du problème (5.1) revient à trouver le point selle (x^*, λ^*) du lagrangien

$$\mathcal{L}(x, \lambda) = f(x) + \langle\lambda, G(x)\rangle,$$

sur $\mathbb{R}^n \times (\mathbb{R}_+)^m$. Nous avions montré dans la preuve de la proposition 4.4 que le point selle vérifie l'inégalité

$$\langle\lambda^* - \lambda, G(x^*)\rangle \geq 0, \quad \forall\lambda \in (\mathbb{R}_+)^m.$$

Donc pour tout $\mu > 0$,

$$\langle\lambda^* - \lambda, \lambda^* - (\lambda^* + \mu G(x^*))\rangle \leq 0, \quad \forall\lambda \in (\mathbb{R}_+)^m.$$

Par la caractérisation (5.3) de la projection, l'inégalité précédente montre que

$$\lambda^* = P_{(\mathbb{R}_+)^m}\left(\lambda^* + \mu G(x^*)\right). \tag{5.6}$$

Donc λ^* est un point fixe de l'application $\lambda \to P_{(\mathbb{R}_+)^m}(\lambda + \mu G(x^*))$. En vue de cette propriété, l'algorithme d'Uzawa se construit de la façon suivante. On part d'un point $\lambda_0 \in (\mathbb{R}_+)^m$ et on construit les suites (x_n) et (λ_n) définies par récurrence pour tout $n \geq 0$ comme suit.

Algorithme d'Uzawa :

$$x_n \in \operatorname{argmin}_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda_n) \tag{5.7}$$
$$\lambda_{n+1} = P_{(\mathbb{R}_+)^m}(\lambda_n + \mu G(x_n)). \tag{5.8}$$

Nous avons le résultat de convergence suivant.

Théorème 5.5. *Supposons que f est α-fortement convexe, que G est convexe et Lipschitzienne de constante C et qu' il existe un point selle (x^*, λ^*) du lagrangien $\mathcal{L}$ sur $\mathbb{R}^n \times (\mathbb{R}_+)^m$. Alors, si $0 < \mu < 2\alpha/C^2$, la suite (x_n) de l'algorithme d'Uzawa converge vers la solution x^* du problème* (5.1).

Preuve. Soit $n \geq 0$. En soustrayant (5.6) et (5.8), on a

$$\begin{aligned}
\|\lambda^* - \lambda_{n+1}\|^2 &= \|P_{(\mathbb{R}_+)^m}(\lambda^* + \mu G(x^*)) - P_{(\mathbb{R}_+)^m}(\lambda_n + \mu G(x_n))\|^2 \\
&\leq \|\lambda^* - \lambda_n + \mu(G(x^*) - G(x_n))\|^2 \\
&\leq \|\lambda^* - \lambda_n\|^2 + \mu^2 C^2 \|x^* - x_n\|^2 + 2\mu \langle \lambda^* - \lambda_n, G(x^*) - G(x_n) \rangle .
\end{aligned} \tag{5.9}$$

où nous avons utilisé que $P_{(\mathbb{R}_+)^m}$ est faiblement contractant et que G est Lipschitz pour passer de la première à la deuxième ligne.

Trouvons une borne pour le produit scalaire $\langle \lambda^* - \lambda_n, G(x^*) - G(x_n) \rangle$. Pour ce faire, nous partons des conditions nécessaires d'optimalité pour x^* et x_n qui s'écrivent respectivement

$$\nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(x^*) = 0$$

et

$$\nabla f(x_n) + \sum_{i=1}^m \lambda_i^{(n)} \nabla g_i(x_n) = 0.$$

Leur soustraction donne

$$\nabla f(x^*) - \nabla f(x_n) + \sum_{i=1}^m \lambda_i^* \nabla g_i(x^*) - \sum_{i=1}^m \lambda_i^{(n)} \nabla g_i(x_n) = 0 \tag{5.10}$$

Comme f est α-fortement convexe, alors $\langle \nabla f(x^*) - \nabla f(x_n), x^* - x_n \rangle \geq \alpha \|x^* - x_n\|^2$. Donc en prenant le produit scalaire de (5.10) par $x^* - x_n$, on déduit l'inégalité

$$\sum_{i=1}^m \lambda_i^* \langle \nabla g_i(x^*), x_n - x^* \rangle + \sum_{i=1}^m \lambda_i^{(n)} \langle \nabla g_i(x_n), x^* - x_n \rangle \geq \alpha \|x^* - x_n\|^2$$

Mais comme les g_i sont convexes, $\langle \nabla g_i(x^*), x_n - x^* \rangle \leq g_i(x_n) - g_i(x^*)$ et $\langle \nabla g_i(x_n), x^* - x_n \rangle \leq g_i(x^*) - g_i(x_n)$, d'où

$$\sum_{i=1}^{m} (\lambda_i^* - \lambda_i^{(n)}) \left(g_i(x_n) - g_i(x^*)\right) = \langle \lambda^* - \lambda_n, G(x_n) - G(x^*) \rangle \geq \alpha \|x^* - x_n\|^2$$

En multipliant cette inégalité par -1 et en l'injectant dans (5.9), il vient

$$\|\lambda^* - \lambda_{n+1}\|^2 \leq \|\lambda^* - \lambda_n\|^2 + (\mu^2 C^2 - 2\mu\alpha)\|x^* - x_n\|^2.$$

Comme $\mu < 2\alpha/C^2$, alors il existe $\beta > 0$ tel que $\mu^2 C^2 - 2\mu\alpha < -\beta$. Donc

$$\beta \|x^* - x_n\|^2 \leq \|\lambda^* - \lambda_n\|^2 - \|\lambda^* - \lambda_{n+1}\|^2.$$

Ceci implique d'une part que la suite des erreurs $\|\lambda^* - \lambda_n\|$ est décroissante et minorée par zéro, donc elle converge. D'autre part, la différence $\|\lambda^* - \lambda_{n+1}\|^2 - \|\lambda^* - \lambda_n\|^2$ tend vers 0 lorsque $n \to \infty$ donc $x_n \to x^*$. □

Remarque 5.6. Le théorème précédent n'assure pas la convergence de λ_n vers λ^* mais seulement la convergence de la suite vers une certaine valeur finie. De plus, l'unicité de λ^* n'est pas assurée sous les hypothèses du théorème.

Exemple : On considère une collection de $N+2$ corps ponctuels ayant tous la même masse m. On note

$$(x_0, y_0),\ (x_1, y_1),\ \ldots,\ (x_{N+1}, y_{N+1})$$

les positions (voir figure II.3). Chaque corps i est relié au corps $i+1$ par un ressort de raideur k et de longueur au repos nulle, de telle sorte que l'énergie élastique associée est

$$\frac{k}{2}\left((x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2\right).$$

L'énergie potentielle de chaque corps est mgy_i où g est la constante de pesanteur.

On suppose que les masses extrêmes sont attachées en $(0,1)$ et $(1,1)$ respectivement, de telle sorte que le vecteur de degrés de libertés peut s'écrire

$$z = (x_1, \ldots, x_N, y_1, \ldots, y_N) \in \mathbb{R}^{2N}.$$

L'énergie totale du système $E(z)$ du système des $N+2$ ressorts étant la somme de l'énergie élastique et de pesanteur, elle est donc égale à

$$E(z) = \frac{k}{2} \sum_{i=0}^{N} \left((x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2\right) + mg \sum_{i=0}^{N+1} y_i \tag{5.11}$$

$$= \frac{k}{2} Az \cdot z - b \cdot z + c \tag{5.12}$$

où

$$A = \begin{pmatrix} \tilde{A}_{N,N} & 0_{N,N} \\ 0_{N,N} & \tilde{A}_{N,N} \end{pmatrix} \in \mathbb{R}^{2N \times 2N}, \quad \text{avec } \tilde{A}_{N,N} = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ 0 & 0 & \cdots & -1 & 2 \end{pmatrix} \in \mathbb{R}^{N \times N} \tag{5.13}$$

et $b = (b_1, \dots, b_{2N})^T \in \mathbb{R}^{2N}$ avec

$$b_i = \begin{cases} 0 & \text{si } i = 1, \dots, N-1 \\ k & \text{si } i = N \\ k - mg & \text{si } i = N+1, 2N \\ -mg & \text{si } i = N+2, \dots, 2N-1 \end{cases}$$

La quantité $c \in \mathbb{R}$ est une constante qui est non nulle avec les calculs effectués jusqu'à présent mais, comme l'énergie est une quantité définie à une constante près, nous fixerons $c = 0$ dans la suite pour simplifier légèrement les calculs.

La position d'équilibre z^* de ce système est celle qui minimise l'énergie totale donc il faut résoudre

$$\min_{z \in \mathbb{R}^{2N}} E(z) = \min_{z \in \mathbb{R}^{2N}} \frac{k}{2} Az \cdot z - b \cdot z$$

où nous cherchons tout d'abord z^* dans $\mathbb{R}^{2N}$ tout entier. Il s'agit d'un problème de minimisation sans contraintes. Comme A est symétrique définie positive, $z \mapsto E(z)$ est α-fortement convexe avec $\alpha = \lambda_{\min}(A)$. Donc, par le théorème 2.7, il existe un unique point z^* d'énergie minimale et ce point vérifie l'équation d'Euler

$$\nabla E(z^*) = 0 \quad \Leftrightarrow \quad z^* = \frac{1}{k} A^{-1} b.$$

Supposons maintenant la présence d'un "plancher" au niveau de l'axe des x, de telle sorte qu'il faut maintenant minimiser l'énergie totale sous les contraintes

$$y_i \geq 0, \quad i = 1, \dots, N.$$

Le problème peut donc s'écrire comme

$$\min_{z \in \mathbb{R}^{2N},\ G(z) = Bz \leq 0} \frac{k}{2} Az \cdot z - b \cdot z$$

avec

$$B = \begin{pmatrix} 0_{N,N} & -\mathrm{Id}_{N,N} \end{pmatrix} \in \mathbb{R}^{N \times 2N}. \tag{5.14}$$

Ce problème a la même forme que celui analysé en (4.9) et nous pouvons le résoudre avec l'algorithme d'Uzawa pour obtenir la position d'équilibre. Nous encourageons le lecteur à tenter de l'implémenter. Une illustration est donnée en figure II.3.

5.4 Programmation linéaire et algorithme du simplexe

Cette courte partie (très largement empruntée à l'ouvrage de Ciarlet) est une introduction aux problèmes de programmation linéaire, c'est-à-dire des problèmes d'optimisation sous contraintes avec critère et contraintes affines.

Structure de l'ensemble des solutions

Soit C une matrice de format $m \times n$, $a \in \mathbb{R}^n$ et $d \in \mathbb{R}^m$. On considère le problème

$$(P) \qquad \min_{x \in K} \langle a, x \rangle \qquad \text{où } K := \{x \in \mathbb{R}^n_+,\ Cx = d\}.$$

On peut montrer que cette structure particulière couvre un grand nombre de situations, puisque tout problème avec critère et contraintes affines peut se mettre sous cette forme.

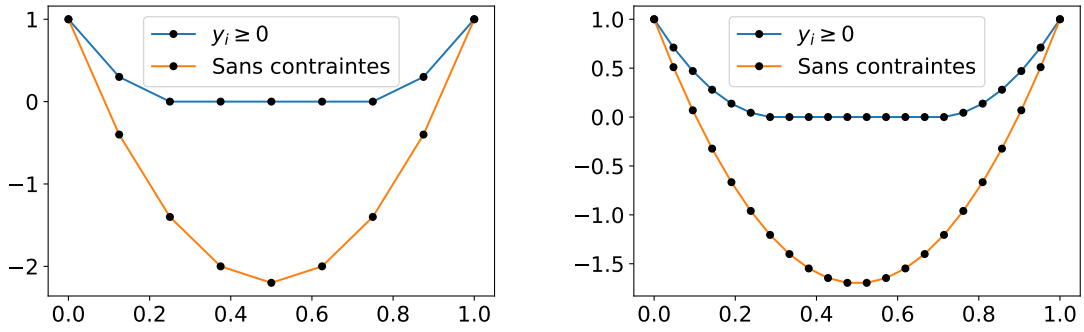

FIGURE II.3 – Solutions d'équilibre pour le système de masses reliées par des ressorts pour $N = 7$ et 20.

Définition 5.7. On dit qu'un point x de l'ensemble $K := \{x \in \mathbb{R}^n_+,\ Cx = d\}$ est un sommet de K (ou un point extrémal de K) si $x \neq 0$ et si, pour tout $x^1, x^2 \in K$ et $\lambda \in]0,1[$, si $x = \lambda x^1 + (1-\lambda)x^2$, alors $x^1 = x^2 = x$.

Par exemple, les sommets du simplexe

$$\Delta := \{x = (x_1, \ldots, x_n) \in \mathbb{R}^n_+,\ \sum_{i=1}^{n} x_i = 1\},$$

sont les vecteurs de la base canonique de $\mathbb{R}^n$.

Nous caractérisons maintenant les sommets de notre contrainte K. Pour un point $x = (x_1, \ldots, x_n) \in \mathbb{R}^n_+$, on définit $J^*(x) = \{j \in \{1, \ldots, n\},\ x_j > 0\}$. Notons aussi par C_j, pour $j \in \{1, \ldots, n\}$, la $j-$ième colonne de la matrice C. Notons que $Cx = \sum_{j=1}^n x_j C_j$.

Théorème 5.8. *Un point $x \in K$ est un sommet de K, si et seulement si, la famille $\{C_j\}_{j \in J^*(x)}$ est libre.*

Cela montre qu'il n'y a qu'un nombre fini de sommet, puisque, pour tout sous-ensemble de I_0 de $\{1, \ldots, n\}$ tel que la famille $(C_j)_{j \in J_0}$ est libre, il existe au plus un élément $x \in \mathbb{R}^n_+$ tel que $\sum_{j \in J_0} x_j C_j = d$.

Preuve. Supposons d'abord que la famille $\{C_j\}_{j \in J^*(x)}$ est libre. Soient $x^1, x^2 \in K$ et $\lambda \in]0,1[$ tels que $x = \lambda x^1 + (1-\lambda)x^2$. Comme les coefficients x^1_j et x^2_j sont positifs, $J^*(x^1) \subset J^*(x)$ et $J^*(x^2) \subset J^*(x)$. De plus,

$$d = \sum_{i \in J^*(x)} x_j C_j = \sum_{i \in J^*(x)} x^1_j C_j = \sum_{i \in J^*(x)} x^2_j C_j.$$

Comme la famille $\{C_j\}_{j \in J^*(x)}$ est libre, cela implique que $x^1_j = x^2_j = x_j$ pour tout j. Donc $x^1 = x^2$ et x est un sommet de K.

Inversement, supposons que la famille $\{C_j\}_{j \in J^*(x)}$ est liée et montrons que x ne peut pas être un sommet de K. Comme la famille $\{C_j\}_{j \in J^*(x)}$ est liée, il existe $j_0 \in J^*(x)$ et des coefficients

$(\alpha_j)_{j\in J^*(x)\setminus\{j_0\}}$ tels que $C_{j_0} = \sum_{j\in J^*(x)\setminus\{j_0\}} \alpha_j C_j$. Pour $\delta > 0$ à choisir plus loin, on définit alors $x^{\delta,+}$ et $x^{\delta,-}$ par

$$x_j^{\delta,\pm} := \begin{cases} 0 & \text{si } j \notin J^*(x) \\ x_j \pm \delta\alpha_j & \text{si } j \in J^*(x)\setminus\{j_0\} \\ x_{j_0} \mp \delta & \text{si } j = j_0 \end{cases}$$

Si $\delta > 0$ est suffisamment petit, on a $x^{\delta,\pm} \in \mathbb{R}^n_+$ et, par définition des coefficient (α_j) on a également que $Cx^{\delta,\pm} = d$, i.e., $x^{\delta,\pm} \in K$. De plus $x = (x^{\delta,+} + x^{\delta,-})/2$ ce qui montre que x n'est pas un sommet. □

Théorème 5.9. *Si le problème (P) admet un minimum, alors soit 0 soit un sommet de K est un point de minimum du problème (P).*

Il se peut que plusieurs sommets soient des minima. Attention, noter que le problème n'admet pas toujours de minimum. Une propriété particulière de la programmation linéaire est que dans ce cas, l'infimum est égal à $-\infty$. Cette propriété est une conséquence du lemme de Farkas et ne sera pas montrée ici.

Preuve. Soit x un point de minimum de (P) pour lequel le cardinal de $J^*(x)$ est minimal. Si $J^*(x) = 0$, alors $x = 0$ et on a fini. Sinon, supposons que x n'est pas un sommet de K. Alors la famille $\{C_j\}_{j\in J^*(x)}$ est liée. Il existe $j_0 \in J^*(x)$ et des coefficients $(\alpha_j)_{j\in J^*(x)\setminus\{j_0\}}$ tels que $C_{j_0} = \sum_{j\in J^*(x)\setminus\{j_0\}} \alpha_j C_j$. Pour $\delta > 0$ à choisir plus loin, on définit comme pour la preuve du théorème 5.8 $x^{\delta,+}$ et $x^{\delta,-}$ par

$$x_j^{\delta,\pm} := \begin{cases} 0 & \text{si } j \notin J^*(x) \\ x_j \pm \delta\alpha_j & \text{si } j \in J^*(x)\setminus\{j_0\} \\ x_{j_0} \mp \delta & \text{si } j = j_0 \end{cases}$$

Comme, pour $\delta > 0$ petit, $x^{\delta,+}$ et $x^{\delta,-}$ sont encore des éléments de K, et comme x est un minimum du problème (P), on doit avoir $\langle a, x^{\delta,+} - x\rangle = -\langle a, x^{\delta,-} - x\rangle = 0$. Donc $x^{\delta,+}$ et $x^{\delta,-}$ sont aussi des minima de P pour tout δ tel que $x^{\delta,+}$ et $x^{\delta,-}$ sont des éléments de K. Soit δ le plus grand réel pour lequel $x^{\delta,+}$ et $x^{\delta,-}$ sont tous deux dans K. Alors forcément il existe un indice $j \in J(x)$ tel que $x_j^{\delta,+} = 0$ ou $x_j^{\delta,-} = 0$, car sinon on pourrait encore augmenter δ. Supposons pour fixer les idées que cela arrive pour $x^{\delta,+}$. Alors $x^{\delta,+}$ est appartient à K, est un minimum du problème (P) et le cardinal de $J^*(x^{\delta,+})$ est strictement inférieur à celui de $J^*(x)$. Cela contredit la définition de x. □

L'intérêt du théorème 5.9 est de réduire la recherche du minimum aux sommets de l'ensemble K, c'est-à-dire à un nombre fini de points. L'algorithme du simplexe (de Dantzig) explique qu'il est possible d'effectuer une énumération intelligente de ces sommets.

Principe de l'algorithme du simplexe

On considère toujours le problème (P) défini dans la partie précédente et on suppose que la matrice C est de rang m. Ceci implique qu'il y a plus d'inconnues que de contraintes : $m \leq n$, ce qui est le plus souvent le cas. Nous supposerons également pour simplifier la discussion que tous les sommets de K sont *non dégénérés*, ce qui signifie que, pour tout sommet x de K, le cardinal de $J^*(x)$ est exactement m. Dans ce cas, la famille $(C_j)_{j\in J^*(x)}$ forme une base de $\mathbb{R}^m$ (et pas seulement une famille libre). On dit alors que $(C_j)_{j\in J^*(x)}$ est la *base* associée au sommet x.

Le principe de l'algorithme du simplexe est de parcourir des sommets de K (et donc des bases, de K) en faisant diminuer à chaque étape le critère.

Expliquons une étape de l'algorithme : soit x un sommet avec une base associée $(C_j)_{j\in J^*(x)}$. L'objectif est de remplacer un des vecteurs de la base par un vecteur hors de la base, ce qui définira le sommet suivant. Soit C_k un vecteur hors base (i.e., $k \notin J^*(x)$). Il existe un unique $m-$uplet de coefficients (α_j^k) tels que $C_k = \sum_{j\in J} \alpha_j^k C_j$. Pour $\delta > 0$, on considère alors les points $x^{k,\delta} = (x_j^{k,\delta})$ de la forme

$$x_j^{k,\delta} := \begin{cases} 0 & \text{si } j \notin J^*(x) \cup \{k\} \\ x_j - \delta\alpha_j^k & \text{si } j \in J^*(x) \\ \delta & \text{si } j = k \end{cases}$$

Notons que

$$\sum_j x_j^{k,\delta} C_j = \delta C_k + \sum_{j\in J^*(x)} (x_j - \delta\alpha_j^k) C_j = \sum_{j\in J} x_j C_j = d.$$

Donc, le point $x^{k,\delta}$ est encore dans K pourvu que ses coordonnées restent positives. Notons que c'est le cas pour $\delta > 0$ petit.

Calculons l'évolution du critère entre x et $x^{k,\delta}$:

$$\langle a, x^{k,\delta} - x\rangle = \delta(a_k - \sum_{j\in J^*(x)} \alpha_j^k a_j)$$

Comme δ est positif, le critère diminuera strictement si $a_k - \sum_{j\in J} \alpha_j^k a_j < 0$.

Proposition 5.10. *Une et une seule des trois alternatives suivantes a lieu :*

(A) Soit pour tout $k \notin J^*(x)$, $a_k - \sum_{j\in J^*(x)} \alpha_j^k a_j \geq 0$. Alors x est un minimum de (P).

(B) Soit il existe au moins un indice $k \notin J^*(x)$ tel que $a_k - \sum_{j\in J} \alpha_j^k a_j < 0$ et tel que $\alpha_j^k \leq 0$ pour tout $j \in J^*(x)$. Alors l'infimum du problème (P) est $-\infty$.

(C) Soit il existe au moins deux indices $k \notin J^*(x)$ et $j_0 \in J^*(x)$ tels que $a_k - \sum_{j\in J^*(x)} \alpha_j^k a_j < 0$ et $\alpha_{j_0}^k > 0$. Alors on considère le plus grand réel $\delta > 0$ tel que $x^{k,\delta}$ est encore dans K. Le point $x^{k,\delta}$ ainsi obtenu est un sommet et le critère en ce point est strictement inférieur à celui en x.

Sous les conditions énoncées ci-dessus, l'algorithme converge en temps fini puisqu'à chaque étape, le critère diminue strictement et qu'il n'y a un nombre fini de sommets.

Preuve de (A). Le seul point à vérifier dans l'analyse ci-dessus est le (A). Supposons que, pour tout $k \notin J$, $a_k - \sum_{j\in J} \alpha_j^k a_j \geq 0$ et montrons que x est un minimum du problème. Soit $y \in K$. Alors

$$Cy = \sum_k y_k C_k = \sum_{j\in J^*(x)} (\sum_k \alpha_j^k y_k) C_j = d = Cx = \sum_{j\in J^*(x)} x_j C_j.$$

Comme les $(C_j)_{j\in J^*(x)}$ forment une famille libre, cela implique que $\sum_k \alpha_j^k y_k = x_j$ pour tout $j \in J^*(x)$. Alors

$$\langle a, y\rangle - \langle a, x\rangle = \sum_k a_k y_k - \sum_{j\in J^*(x)} a_j x_j = \sum_k (a_k - \sum_{j\in J^*(x)} \alpha_j^k a_j) y_k \geq 0.$$

Donc x est un minimum. □

Partie III

Programmation dynamique

Cette partie est une courte introduction au contrôle optimal. La théorie du contrôle s'intéresse aux systèmes dynamiques dépendant d'un paramètre (appelé *contrôle* ou bien *commande*) sur lequel on peut agir pour, par exemple, amener la position du système d'un point à un autre. En contrôle optimal, on cherche à agir sur la commande du système dynamique de façon à optimiser un critère donné. Les systèmes dynamiques peuvent être de différentes natures (en temps discret ou continu, avec ou sans bruit, ...) et avoir différentes origines (mécaniques, électriques, chimiques, économiques,...). Par exemple en finance mathématique, on modélise fréquemment l'évolution d'un portefeuille comme un système dynamique stochastique sur lequel on agit (en temps discret ou continu) en vendant ou en achetant des actifs financiers. Un autre exemple d'application en économie est la théorie des anticipations rationnelles, qui fait largement appel au contrôle optimal (cf. la monographie de Lucas et Stokey [4] qui contient de très nombreux exemples économiques et traite également de la programmation markovienne, qui est hors du champ de ce cours).

Dans la partie précédente, nous avons expliqué comment écrire les conditions nécessaires d'optimalité avec contraintes sur l'état. Nous verrons qu'il est également possible d'écrire des conditions nécessaires d'optimalité dans le cadre du calcul des variations et du contrôle optimal : ce sont respectivement les conditions d'Euler et le principe du maximum de Pontryagin. Cependant, ces conditions nécessaires sont souvent difficiles à exploiter, et il vaut mieux mettre en place une méthode d'énumération intelligente. Le principe de programmation dynamique explique comment n'explorer qu'une partie de toutes les possibilités, tout en conservant l'optimalité.

De façon un peu caricaturale, le principe de programmation dynamique affirme qu'un chemin optimal entre deux points n'est constitué que de chemins optimaux. Autrement dit, si un chemin $(\mathcal{C})$ est optimal pour aller d'un point A à un point B, et si un point C appartient à $(\mathcal{C})$, alors les sous-chemins de $(\mathcal{C})$ allant de A à C et de C à B sont optimaux.

Nous explorerons ce principe en temps discret, puis en temps continu.

1 Problèmes en temps discret

On considère un système dynamique discret dont l'état à chaque instant est donné par

$$\begin{aligned} x_{n+1} &= f_n(x_n, u_n), \qquad n = 0,\ 1,\ 2, \ldots,\ N-1, \\ x_0 &= \bar{x}. \end{aligned}$$

L'instant final $N \in \mathbb{N}^*$ s'appelle *l'horizon* du problème. à chaque instant $n \in \{0, \ldots, N\}$, x_n est *l'état du système* et u_n est le *contrôle*, c'est-à-dire la décision prise à l'instant n. f_n est la dynamique du problème à l'instant n.

Nous supposerons que l'état du système vit dans un ensemble X fixé, c'est à dire que $x_n \in X$ pour tout $n \in \{0, 1, \ldots, N\}$. Le contrôle u_n vit dans un ensemble U_n et $f_n : X \times U_n \to X$. La condition initiale $\bar{x} \in X$ du système est fixée.

Pour la suite, il sera utile d'introduire la suite des contrôles

$$\mathbf{u} = (u_n)_{n=0}^{N-1} \quad \in \quad \mathbb{U} \coloneqq U_0 \times \cdots \times U_{N-1},$$

ainsi que des sous-parties de cette suite définies pour tout $p \in \{0, \dots, N-1\}$ comme

$$\mathbf{u}_p = (u_n)_{n=p}^{N-1} \quad \in \quad \mathbb{U}_p \coloneqq U_p \times \cdots \times U_{N-1}.$$

De cette façon, pour toute suite de contrôle $\mathbf{u} \in \mathbb{U}$, la suite des états $\mathbf{x} = (\bar{x}, x_1, \dots, x_N) \in X^N$ engendrée par le système dynamique est donc fonction de l'état initial $\bar{x}$ et de $\mathbf{u}$ et peut se noter $\mathbf{x} = \mathbf{x}(\bar{x}, \mathbf{u})$.

Dans un problème de contrôle optimal discret (ou d'horizon fini), étant donné une condition initiale $\bar{x}$, on définit pour tout contrôle $\mathbf{u} \in \mathbb{U}$ une fonction de coût

$$C(\bar{x}, \mathbf{u}) \coloneqq \sum_{n=0}^{N-1} L_n(x_n, u_n) + g(x_N), \quad \text{avec} \quad \mathbf{x} = \mathbf{x}(\bar{x}, \mathbf{u}). \tag{1.1}$$

La quantité $L_n(x_n, u_n)$ est appelée le *coût courant* à l'instant n, g étant le *coût terminal*. Ce sont des fonctions $L_n : X \times U_n \to \mathbb{R}$, $g : X \to \mathbb{R}$.

L'objectif est de minimiser le coût parmi tous les contrôles possibles $\mathbf{u} \in \mathbb{U}$, c'est à dire de calculer

$$\min_{\mathbf{u} \in \mathbb{U}} \sum_{n=0}^{N-1} L_n(x_n, u_n) + g(x_N) \tag{1.2}$$

On peut aussi étudier des problèmes en horizon infini avec $N = +\infty$. On considérera des problèmes avec critère escompté du type

$$\min_{\mathbf{u} \in \mathbb{U}} \sum_{n=0}^{+\infty} r^n L(x_n, u_n), \tag{1.3}$$

où, cette fois-ci, $\mathbb{U} = U^{\mathbb{N}}$ et $r \in]0, 1[$ est un facteur d'escompte

Dans les deux cas, les questions qui se posent sont l'existence de solutions à (1.2) et (1.3) (appelées aussi politiques optimales) et la caractérisation maniable de ces dernières, l'idéal étant d'obtenir une stratégie permettant de les calculer numériquement. Nous verrons qu'en exploitant la structure récursive de ces problèmes, on pourra atteindre ces objectifs en utilisant les idées intuitives et efficaces de la programmation dynamique, de la fonction valeur et des équations de Bellman.

Avant d'aller plus loin, considérons quelques exemples pour fixer les idées.

Remarque 1.1. En économie, on a plutôt tendance à maximiser un profit : on se ramène sans difficulté à ce cas en se rappelant que $\max(\dots) = -\min(-\dots)$.

1.1 Quelques exemples

Les exemples suivants sont tirés du polycopié de G. Carlier.

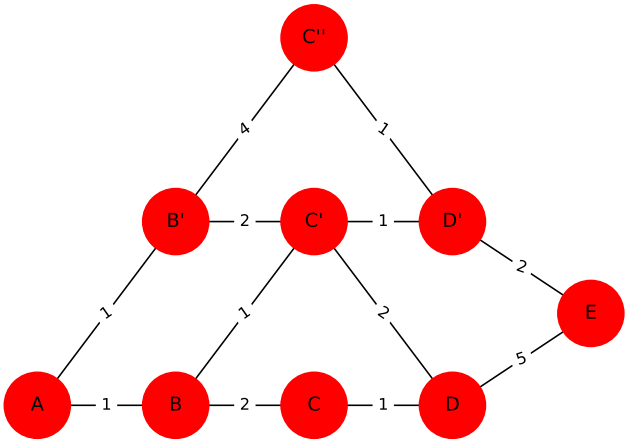

FIGURE III.1 – Graphe des villes visitées par le voyageur de commerce.

Le problème du voyageur de commerce Il s'agit ici d'un problème en horizon fini. Sa résolution illustrera de façon simple le principe de la programmation dynamique.

Considérons un voyageur de commerce qui doit se rendre de la ville A à la ville E en passant par plusieurs villes intermédiaires. Les chemins possibles sont modélisés pas un graphe ayant A et E pour sommets initial et final (voir figure III.1). Les autres sommets sont les villes étapes. Les arrêtes de ce graphe représentent les trajets intermédiaires. On notera $\Gamma(M)$ les successeurs de la ville M et pour $N \in \Gamma(M)$ on notera MN le temps du parcours MN. Le tableau suivant donne les valeurs $\Gamma(M)$ pour chaque ville et donne les temps de parcours entre étapes. Ainsi, on lit par exemple que $\Gamma(A) = \{B, B'\}$ et $AB = AB' = 1$. Nous pouvons observer que ce problème est bien du type (1.2). Ici, la fonction du coût courant est $L(M, N) = MN$ où M est la ville courante et $N \in \Gamma(M)$ est la prochaine ville étape qui constitue notre contrôle.

Pour déterminer le ou les plus courts chemins, on pourrait bien sû r tous les essayer mais il sera plus efficace d'utiliser le fait que si un chemin optimal de A à E passe par M, alors il est encore optimal de M à E. Cette observation est une expression simple du principe de programmation dynamique que nous formaliserons par la suite.

Introduisons la fonction valeur $V(M)$ égale au temps de parcours minimal entre M et E. Nous souhaitons calculer $V(A)$. Pour ce faire, on voit bien que le calcul doit partir de la fin en procédant par induction rétrograde. On a d'abord

$$V(D) = 5, \quad V(D') = 1.$$

On remonte ensuite aux villes précédentes. Le principe de programmation dynamique donne

$$V(C) = 6, \quad V(C') = \min\big(1 + V(D'), 2 + V(D)\big) = 1 + V(D') = 3, \quad V(C'') = 3.$$

En réitérant l'argument, il vient :

$$V(B) = \min\big(2 + V(C), 1 + V(C')\big) = 1 + V(C') = 4$$
$$V(B') = \min\big(2 + V(C'), 4 + V(C'')\big) = 5$$

et enfin

$$V(A) = \min\big(1 + V(B), 1 + V(B')\big) = 1 + V(B) = 5.$$

Le temps de parcours minimal est donc de 5 et correspond au seul parcours ABC'D'E.

Cet exemple élémentaire est instructif à plusieurs égards :

1. On voit aisément comment généraliser la stratégie précédente à des problèmes plus généraux de la forme (1.2) : on introduit les fonctions valeurs à différentes dates et on les calcule en partant de la fin par induction rétrograde, qui est exactement le principe de la programmation dynamique.
2. Dans l'exemple, on n'a pas exploré tous les chemins possibles mais seulement les chemins optimaux à partir de M. De ce fait, les raisonnements précédents montrent par exemple que si le voyageur de commerce s'égare en B' (ville qui n'est pas sur la trajectoire optimale), alors par la suite il sera optimal de passer par C'D'E.
3. Il peut paraître curieux alors qu'on s'est posé un seul problème issu du point A, de chercher à résoudre tous les problèmes issus des points intermédiaires. Donnons deux arguments pour lever cette objection : tout d'abord, la stratégie de résolution est robuste : si une erreur est commise à un moment donné et conduit à passer par la ville non optimale M, par la suite on peut se rattraper en suivant le chemin optimal à partir de M. La deuxième raison est que cette stratégie est naturelle (choisir la ville suivante en fonction de la ville où on se trouve maintenant plutôt que de suivre un plan exactement établi à l'avance) et permet de se rammener à une succession de problèmes statiques.

Un problème d'exploitation forestière : On considère une forêt qui initialement est de taille $x_0 = \bar{x}$ et x_n est sa taille à la date n (variable d'état). Un exploitant choisit à chaque période un niveau de coupe u_n (variable de contrôle) et l'évolution de la taille de la forêt est supposée être régie par la dynamique

$$x_{n+1} = H(x_n) - u_n.$$

En supposant que le prix du bois est constant et égal à 1 et que le coût de l'abattage est C, le profit actualisé de l'exploitant s'écrit

$$\sum_{n=0}^{\infty} \beta^n (u_n - C(u_n)).$$

En réécrivant ce profit en fonction de la variable d'état x_n et en imposant que $u_n \geq 0$ et $x_n \geq 0$, le programme de l'exploitant se réécrit sous la forme

$$\sup \left\{ \sum_{n=0}^{\infty} \beta^n [H(x_n) - x_{n+1} - C(H(x_n) - x_{n+1})] \right\}$$

sous les contraintes $x_0 = \bar{x}$ et $0 \leq x_{n+1} \leq H(x_n)$ pour tout $n \geq 0$.

1.2 Problèmes en horizon fini

Nous considérons ici le problème en horizon fini

$$\min_{\mathbf{u} \in \mathbb{U}} \sum_{n=0}^{N-1} L_n(x_n, u_n) + g(x_N), \tag{1.4}$$

avec $\mathbb{U} = U_0 \times \cdots \times U_N$. Le résultat principal à retenir est que ce problème de minimisation dans l'espace des suites $\mathbf{u} \in \mathbb{U}$ peut être traité par des techniques d'optimisation statique : c'est le principe de la *programmation dynamique*. Pour le mettre en oeuvre, il sera néanmoins nécessaire de résoudre un grand nombre de problèmes statiques balayant *toutes* les conditions initiales et commençant à *n'importe quel instant*.

Pour tout $x \in X$ et pour tout $p \in \{0, \dots, N-1\}$, nous définissons la *fonction valeur* du problème comme

$$V(p,x) \coloneqq \inf_{\mathbf{u}_p \in \mathbb{U}_p} \sum_{n=p}^{N-1} L_n(x_n, u_n) + g(x_N)$$

où l'état $\mathbf{x}_p = (x_n)_{n=p}^N$ est défini par récurrence par

$$\begin{cases} x_p = \bar{x} \\ x_{n+1} = f_n(x_n, u_n), \qquad n = p,\ p+1, \dots,\ N-1. \end{cases}$$

En d'autres termes, $\mathbf{x}_p = \mathbf{x}_p(\bar{x}, \mathbf{u}_p)$.

Notons que la quantité que nous cherchons à minimiser dans est $V(0, \bar{x})$. Le principe de la programmation dynamique repose sur le résultat suivant, où l'égalité (1.5) ci-dessous porte le nom *d'équation de Bellman*.

Théorème 1.2 (Programmation dynamique)**.** *Pour tout $x \in X$, on a*

$$\begin{cases} V(p,x) & = \inf_{u \in U_p} \left\{ L_p(x,u) + V(p+1, f_p(x,u)) \right\}, \quad \forall p \in \{0, \dots, N-1\} \\ V(N,x) & = g(x). \end{cases} \tag{1.5}$$

Preuve du théorème 1.2. Posons

$$W(p,x) \coloneqq \inf_{u \in U_p} \left\{ L_p(x,u) + V(p+1, f_p(x,u)) \right\}.$$

On veut montrer que $W = V$.

Soient $p \in \{0, \dots, N\}$, $x \in X$ et $\varepsilon > 0$ fixé. Soit $\mathbf{u}_p = (u_n)_{n \geq p}$ un contrôle $\varepsilon-$optimal pour $V(p,x)$ et $\mathbf{x}_p = \mathbf{x}_p(x, \mathbf{u}_p) = (x, x_{p+1}, \dots, x_N)$ l'état du système dynamique associé. Alors

$$\begin{array}{rcl} V(p,x) + \varepsilon & \geq & \sum_{n=p}^{N-1} L_n(x_n, u_n) + g(x_N) \ = \ L_p(x, u_p) + \sum_{n=p+1}^{N-1} L_n(x_n, u_n) + g(x_N) \\ & \geq & L_p(x, u_p) + V(p+1, f_p(x, u_p)) \ \geq \ W(p,x) \end{array}$$

puisque $x_{p+1} = f_p(x, u_p)$. Comme ε est arbitraire, cela montre que $V \geq W$.

Inversement, soit $u_p \in U_p$ un contrôle $\varepsilon-$optimal pour $W(p,x)$. Alors,

$$W(p,x) + \varepsilon \geq L_p(x, u_p) + V(p+1, f_p(x, u_p)).$$

De plus, soit également $\mathbf{u}_{p+1} = (u_n)_{n=p+1}^{N-1} \in \mathbb{U}_{p+1}$ une suite de contrôles $\varepsilon-$optimale pour $V(p+1, f_p(x, u_p))$. Alors,

$$V(p+1, f_p(x, u_p)) + \varepsilon \geq \sum_{n=p+1}^{N-1} L_n(x_n, u_n) + g(x_N).$$

où $\mathbf{x}_{p+1} = (x_{p+1} = f_p(x, u_p), \dots, x_N) = \mathbf{x}_{p+1}(x, \mathbf{u}_{p+1})$ est l'état du système dynamique associé à $\mathbf{u}_{p+1}$ partant de $f_p(x, u_p)$ au temps $p+1$.

Définissons alors le contrôle $\hat{\mathbf{u}}_p = (\hat{u}_p, \dots, \hat{u}_N)$ de composantes

$$\hat{u}_n := \begin{cases} u_p & \text{si } n = p \\ u_n & \text{si } n \geq p+1 \end{cases}$$

et notons $\hat{x}_p = \hat{x}_p(x, \hat{\mathbf{u}}_p)$ l'évolution de l'état à partir du temps p issu de x.

Notons que

$$\hat{x}_n = \begin{cases} x & \text{si } n = p \\ f_p(x, u_p) & \text{si } n = p+1 \\ x_n & \text{si } n \geq p+2. \end{cases} \tag{1.6}$$

Alors

$$\begin{aligned} V(p,x) \;&\leq\; \sum_{n=p}^{N-1} L_n(\hat{x}_n, \hat{u}_n) + g(\hat{x}_N) \;=\; L_p(x,u_p) + \sum_{n=p+1}^{N-1} L_n(x_n,u_n) + g(x_N) \\ &\leq\; L_p(x,u_p) + V(p+1, f_p(x,u_p)) + \varepsilon \;\leq\; W(p,x) + 2\varepsilon \end{aligned}$$

Comme ε est arbitraire, cela montre que $V \leq W$ et conclut la preuve. □

Le principe de la programmation dynamique montre que le problème initial peut s'exprimer de deux façons :

$$\begin{aligned} V(0,\bar{x}) &= \inf_{\mathbf{u}\in\mathbb{U}} \sum_{n=0}^{N-1} L_n(x_n,u_n) + g(x_N) \\ &= \inf_{u\in U_0} \left\{ L_0(\bar{x},u) + V(1, f_0(\bar{x},u)) \right\}. \end{aligned}$$

Le problème dans la deuxième égalité est en principe "plus simple" à résoudre que le problème initial, puisqu'il s'agit d'un problème de minimisation standard au lieu d'une minimisation sur une suite à valeurs dans $\mathbb{U}$. Pour calculer $V(0,\bar{x})$, on résout par induction rétrograde les problèmes :

$$\begin{aligned} V(N-1,x) &= \inf_{u\in U_{N-1}} \{ L_{N-1}(x,u) + g(f_{N-1}(x,u)) \} \qquad \forall x \in X, \\ V(N-2,x) &= \inf_{u\in U_{N-2}} \{ L_{N-2}(x,u) + V(N-1, f_{N-2}(x,u)) \} \qquad \forall x \in X, \\ &\vdots \\ V(0,x) &= \inf_{u\in U_0} \{ L_0(x,u) + V(1, f_0(x,u)) \} \qquad \forall x \in X. \end{aligned}$$

La fonction valeur permet le calcul des solutions optimales si l'infimum est bien atteint, c'est à dire, si pour tout $(n,x) \in \{0,\dots,N-1\} \times X$, il existe $u_n^*(x)$ un "feedback optimal" vérifiant

$$L_n(x,u_n^*(x)) + V(n+1, f_n(x,u_n^*(x))) = \inf_{u\in U_n} \left\{ L_n(x,u) + V(n+1, f_n(x,u)) \right\}.$$

Enonçons des conditions garantissant l'existence d'un tel feedback optimal. Supposons que :

1. les U_n et X sont des ensembles métriques et les U_n sont compacts,
2. les fonctions $f_n : X \times U_n \to X$, $L_n : X \times U_n \to \mathbb{R}$ et $g : X \to \mathbb{R}$ sont continues.

Nous vérifierons plus bas qu'alors $x \to V(n,x)$ est continue pour tout n. Dans ce cas, l'application $u \to L_n(x,u)+V(n+1,f_n(x,u))$ est continue sur U_n (pour tout $x \in X$) et, comme U_n est compact, a donc un minimum $u_n^*(x)$ sur U_n.

Expliquons maintenant la terminologie de "feedback optimal".

Proposition 1.3. *Soit $\bar{x} \in X$ une condition initiale fixée. Si on définit par récurrence les suites $(\bar{u}_n)$ et $(\bar{x}_n)$ par*

$$\bar{x}_0 = \bar{x}, \qquad \bar{u}_n = u_n^*(\bar{x}_n), \qquad \bar{x}_{n+1} = f_n(\bar{x}_n, \bar{u}_n),$$

alors la suite $(\bar{u}_n)$ est optimale pour le problème de contrôle discret, i.e.,

$$V(0,\bar{x}) = \sum_{n=0}^{N-1} L_n(\bar{x}_n, \bar{u}_n) + g(\bar{x}_N).$$

Preuve. Montrons par récurrence que pour tout $p \in \{0,\dots,N\}$,

$$V(0,\bar{x}) = \sum_{n=0}^{p-1} L_n(\bar{x}_n, \bar{u}_n) + V(p,\bar{x}_p).$$

Cette relation est clairement vraie pour $p = 0$. Maintenant, supposons-la vraie pour un certain $p \geq 0$. En utilisant d'abord la programmation dynamique puis le choix de u^*, on a

$$V(p,\bar{x}_p) = \inf_{u \in U_p} \{L_p(\bar{x}_p,u) + V(p+1, f_p(\bar{x}_p,u))\} = L_p(\bar{x}_p, u_p^*(\bar{x}_n)) + V(p+1, f_p(\bar{x}_p, u_p^*(x_p)))$$

où $u_p^*(x_p) = \bar{u}_p$ et $f_p(\bar{x}_p, u_p^*(x_p)) = \bar{x}_{p+1}$. On utilise alors l'hypothèse de récurrence pour obtenir :

$$\begin{aligned} V(0,\bar{x}) &= \sum_{n=0}^{p-1} L_n(\bar{x}_n,\bar{u}_n) + V(p,\bar{x}_p) = \sum_{n=0}^{p-1} L_n(\bar{x}_n,\bar{u}_n) + L_n(\bar{x}_p,\bar{u}_p) + V(p+1,\bar{x}_{p+1}) \\ &= \sum_{n=0}^{p} L_n(\bar{x}_n,\bar{u}_n) + V(p+1,\bar{x}_{p+1}), \end{aligned}$$

ce qui est la relation au rang $p+1$. Par récurrence, on en déduit le résultat pour tout $p \in \{0,\dots,N\}$. En particulier, pour $p = N$, on a $V(p,\bar{x}_p) = g(\bar{x}_N)$ et donc

$$V(0,\bar{x}) = \sum_{n=0}^{N-1} L_n(\bar{x}_n,\bar{u}_n) + g(\bar{x}_N),$$

ce qui prouve l'optimalité de $(\bar{u}_n)$. □

Nous finissons par la preuve de la continuité de V.

Proposition 1.4. *Supposons que les U_n et X sont des ensembles métriques, que les U_n sont compacts et que les fonctions $f_n : X \times U_n \to X$, $L_n : X \times U_n \to \mathbb{R}$ et $g : X \to \mathbb{R}$ sont continues. Alors $x \to V(n,x)$ est continue pour tout n.*

Preuve. On procède par récurrence descendante en utilisant le principe de programmation dynamique. La continuité pour $n = N$ est vraie par hypothèse, puisque $V(N, x) = g(x)$ avec g continue.

Supposons la continuité de $V(n+1, \cdot)$ et montrons celle de $V(n, \cdot)$. Par programmation dynamique, on a

$$V(n, x) = \inf_{u \in U_n} \{L_n(x, u) + V(n+1, f_n(x, u))\}.$$

Or l'application $(x, u) \to L_n(x, u) + V(n+1, f_n(x, u))$ est continue puisque la continuité de L_n et f_n figure dans nos hypothèses et que $V(n+1, \cdot)$ est continue par hypothèse de récurrence. L'ensemble U_n étant compact, cela implique la continuité de $V(n, \cdot)$ par le résultat classique évoqué ci-dessous et laissé en exercice. □

Proposition 1.5. *Soit X et U deux ensembles métriques et U un compact. On suppose que l'application $h : X \times U \to \mathbb{R}$ est continue. Alors l'application marginale*

$$\bar{h}(x) \coloneqq \min_{u \in U} h(x, u)$$

est continue sur X.

1.3 Problèmes en horizon infini

On suppose ici que l'ensemble de contrôle U est indépendant du temps, que le coût courant $L : X \times U \to \mathbb{R}$ est borné et indépendant du temps, et que le taux d'intérêt r vérifie : $r \in]0, 1[$. Pour tout $\bar{x} \in X$, on pose

$$V(\bar{x}) = \inf_{\mathbf{u} \in \mathbb{U}} \sum_{n=0}^{+\infty} r^n L(x_n, u_n),$$

où $\mathbb{U} = U^{\mathbb{N}}$ et où l'état $(x_n)_{n \in \mathbb{N}}$ est défini par récurrence par

$$\begin{cases} x_0 = \bar{x} \\ x_{n+1} = f(x_n, u_n), \qquad n \in \mathbb{N} \end{cases}$$

Noter que la somme $\displaystyle\sum_{n=0}^{+\infty} r^n L(x_n, u_n)$ est bien convergente car L est bornée et $r \in]0, 1[$.

Le principe de programmation dynamique dans le cas présent est donné par le résultat suivant, où l'équation (1.7) porte encore une fois le nom *d'équation de Bellman.*

Théorème 1.6 (Programmation dynamique)**.** *Pour tout $x \in X$, on a*

$$V(x) = \inf_{u \in U} \{L(x, u) + rV(f(x, u))\}. \tag{1.7}$$

Preuve du théorème 1.6. La démonstration est très proche de celle des problèmes à horizon fini. La seule différence repose sur la façon dont on élimine la variable temporelle. Posons

$$W(x) \coloneqq \inf_{u \in U} \{L(x, u) + rV(f(x, u))\}.$$

On veut prouver que $W = V$ pour tout $x \in X$.

Soit $\varepsilon > 0$ et (u_n) un contrôle $\varepsilon-$optimal pour $V(x)$. Alors

$$\begin{aligned} V(x)+\varepsilon \;\geq\; & \sum_{n=0}^{+\infty} r^n L(x_n,u_n) \;=\; L(x,u_0)+\sum_{n=1}^{+\infty} r^n L(x_n,u_n) \\ =\; & L(x,u_0)+r\sum_{n=0}^{+\infty} r^n L(x_{n+1},u_{n+1}) \;\geq\; L(x,u_0)+rV(f(x,u_0)) \;\geq\; W(x) \end{aligned}$$

(on a utilisé le fait que la trajectoire $(x_{n+1})_{n\geq 0}$ vérifie effectivement la relation de récurrence car f ne dépend pas de n). Cela prouve que $V(x) \geq W(x)$.

Inversement, soit $u \in U$ $\varepsilon-$optimal pour $W(x)$, (u_n) $\varepsilon-$optimal pour $V(f(x,u))$ et (x_n) la trajectoire associée issue de $f(x,u)$. On définit le contrôle $(\hat{u}_n)_{n\geq 0}$ par

$$\hat{u}_n \coloneqq \begin{cases} u & \text{si } n=0 \\ u_{n-1} & \text{si } n \geq 1 \end{cases}$$

et on note $(\hat{x}_n)_{n\geq 0}$ la trajectoire associée issue de x. Alors $\hat{x}_1 = f(\hat{x}_0,\hat{u}_0) = f(x,u)$ et, si $\hat{x}_n = x_{n-1}$, alors $\hat{x}_{n+1} = f(\hat{x}_n,\hat{u}_n) = f(x_{n-1},u_{n-1}) = x_n$ pour tout $n \geq 0$ (on utilise encore le fait que f ne dépend pas de n). Donc, par récurrence, $\hat{x}_n = x_{n-1}$ pour tout $n \geq 1$ et

$$\begin{aligned} W(x)+(1+r)\varepsilon \;\geq\; & L(x,u)+rV(f(x,u))+r\varepsilon \;\geq\; L(\hat{x}_0,\hat{u}_0)+r\sum_{n=0}^{+\infty} r^n L(x_n,u_n) \\ =\; & L(\hat{x}_0,\hat{u}_0)+\sum_{n=0}^{+\infty} r^{n+1} L(\hat{x}_{n+1},\hat{u}_{n+1}) \;=\; \sum_{n=0}^{+\infty} r^n L(\hat{x}_n,\hat{u}_n) \;\geq\; V(x). \end{aligned}$$

D'où $W(x) \geq V(x)$, ce qui conclut la preuve. □

Notons que, contrairement au cas de l'horizon fini, l'équation de Bellman (1.7) ne permet pas un calcul explicite direct de la fonction valeur, puisque V apparaît à gauche et à droite de l'égalité. Nous expliquons dans ce qui suit que, dans un certain cadre, V est l'unique solution de cette équation implicite, ce qui peut fournir des schémas de calcul numérique pour V.

Pour cela, posons

$$\|L\|_\infty \coloneqq \sup_{(x,u)\in X\times U} |L(x,u)|$$

et définissons $B(X)$ comme l'ensemble des applications bornées de X dans $\mathbb{R}$. On rappelle que $B(X)$, muni de la norme

$$\|h\|_\infty \coloneqq \sup_{x\in X} |h(x)| \qquad \forall h \in B(X)$$

est un espace de Banach. Définissons l'opérateur (non linéaire) $T : B(X) \to B(X)$ par

$$T(h)(x) \coloneqq \inf_{u\in U} \{L(x,u)+rh(f(x,u))\} \qquad \forall h \in B(X).$$

Notons qu'en effet $T(h) \in B(X)$ puisque, pour tout $x \in X$,

$$|T(h)(x)| \leq \inf_{u\in U} \{|L(x,u)|+r|h(f(x,u))|\} \leq \|L\|_\infty + r\|h\|_\infty.$$

Donc $\|T(h)\|_\infty \leq \|L\|_\infty + r\|h\|_\infty$ et $T(h) \in B(X)$.

Théorème 1.7. *L'opérateur T est contractant dans $B(X)$:*

$$\|T(h) - T(h')\|_\infty \leq r\|h' - h\|_\infty \qquad \forall h, h' \in B(X),$$

et la fonction valeur V est son unique point fixe dans $B(X)$.

Preuve. Remarquons d'abord que, lorsque L est bornée, V l'est aussi avec

$$\|V\|_\infty \leq \sum_{n=0}^{\infty} r^n \|L\|_\infty \leq \frac{\|L\|_\infty}{1-r}.$$

Le théorème 1.6 implique que V est un point fixe de T.

Il reste juste à vérifier que T est contractant, car alors il possède un unique point fixe. Soient $h, h' \in B(X)$ et $x \in X$. Pour tout $u \in U$ on a

$$L(x,u) + rh(f(x,u)) \leq L(x,u) + rh'(f(x,u)) + r\|h' - h\|_\infty.$$

En prenant l'inf par rapport à u à gauche et à droite, on obtient :

$$\begin{aligned} T(h)(x) &= \inf_{u \in U} \{L(x,u) + rh(f(x,u))\} \\ &\leq \inf_{u \in U} \left\{L(x,u) + rh'(f(x,u))\right\} + r\|h' - h\|_\infty = T(h')(x) + r\|h' - h\|_\infty. \end{aligned}$$

On en déduit que

$$T(h)(x) - T(h')(x) \leq r\|h' - h\|_\infty.$$

En inversant les rôles de h et h' on obtient de même

$$T(h')(x) - T(h)(x) \leq r\|h' - h\|_\infty.$$

D'où

$$|T(h)(x) - T(h')(x)| \leq r\|h' - h\|_\infty.$$

En prenant le sup en $x \in X$ on obtient finalement

$$\|T(h) - T(h')\|_\infty \leq r\|h' - h\|_\infty,$$

ce qui prouve que T est une contraction puisque $r \in]0,1[$. □

La caractérisation précédente fournit un algorithme pour calculer la fonction valeur. Pour une fonction $h_0 \in B(X)$ arbitraire, on définit par récurrence la suite de fonctions (h_k) par $h_{k+1} = T(h_k)$. Alors le théorème du point fixe affirme que la suite (h_k) converge dans $B(X)$ (i.e. uniformément) vers la fonction valeur V. Plus précisément

$$\|V - h_k\|_\infty \leq r^k \|V - h_0\|_\infty \qquad \forall k \in \mathbb{N}.$$

Comme pour les problèmes en horizon fini, la fonction valeur peut servir également à décrire les feedbacks optimaux. Supposons pour cela que, pour tout $x \in X$, il existe $u^*(x) \in U$ un "feedback optimal", i.e. vérifiant

$$L(x, u^*(x)) + rV(f(x, u^*(x))) = \inf_{u \in U} \{L(x,u) + rV(f(x,u))\}.$$

On peut montrer que, si U et X sont des ensembles métriques, si U est compact et si les fonctions $f : X \times U \to X$ et $L : X \times U \to \mathbb{R}$ et $g : X \to \mathbb{R}$ sont continues, alors un tel feedback existe : en effet la fonction valeur V est alors continue, et la fonction continue $u \to L(x,u) + rV(f(x,u)$ admet donc un minimum. La preuve de la continuité de V est dans ce cadre un peu plus délicate que dans le cas de l'horizon fini, et est omise.

Proposition 1.8. *Soit $\bar{x} \in X$ une condition initiale. Si on définit par récurrence les suites $(\bar{u}_n)$ et $(\bar{x}_n)$ par*

$$\bar{x}_0 = \bar{x}, \qquad \bar{u}_n = u^*(x_n), \qquad \bar{x}_{n+1} = f(\bar{x}_n, \bar{u}_n),$$

alors la suite $(\bar{u}_n)$ est optimale pour le problème de contrôle discret, i.e.,

$$V(\bar{x}) = \sum_{n=0}^{+\infty} r^n L(\bar{x}_n, \bar{u}_n).$$

Preuve. Montrons par récurrence que, pour tout $N \in \mathbb{N}$,

$$V(x) = \sum_{n=0}^{N-1} r^n L(\bar{x}_n, \bar{u}_n) + r^N V(\bar{x}_N). \tag{1.8}$$

C'est clairement vrai pour $N = 0$. Supposons la relation vraie à un rang N et montrons-la pour $N + 1$. Par programmation dynamique,

$$\begin{aligned} V(\bar{x}_N) &= \inf_{u \in U} \{L(\bar{x}_N, u) + rV(f(\bar{x}_N, u))\} = L(\bar{x}_N, u^*(\bar{x}_N)) + rV(f(\bar{x}_N, u^*(\bar{x}_N))) \\ &= L(\bar{x}_N, \bar{u}_N) + rV(f(\bar{x}_N, \bar{u}_N)). \end{aligned}$$

On utilise alors l'hypothèse de récurrence :

$$\begin{aligned} V(x) &= \sum_{n=0}^{N-1} r^n L(\bar{x}_n, \bar{u}_n) + r^N V(\bar{x}_N) = \sum_{n=0}^{N-1} r^n L(\bar{x}_n, \bar{u}_n) + r^N L(\bar{x}_N, \bar{u}_N) + r^{N+1} V(f(\bar{x}_N, \bar{u}_N)) \\ &= \sum_{n=0}^{N} r^n L(\bar{x}_n, \bar{u}_n) + r^{N+1} V(\bar{x}_{N+1}) \end{aligned}$$

Donc la relation est vraie au rang $N + 1$ et, par récurrence, pour tout N.

Faisons maintenant tendre N vers $+\infty$ dans la relation (1.8). Comme V est borné et r appartient à $]0, 1[$, le terme $(r^N V(\bar{x}_N))$ tend vers 0 et (1.8) devient

$$V(x) = \sum_{n=0}^{\infty} r^n L(\bar{x}_n, \bar{u}_n),$$

ce qui prouve l'optimalité de $(\bar{u}_n)$. □

2 Calcul des variations

Le calcul des variations s'intéresse aux problèmes de minimisation dans lesquels la variable à optimiser est une fonction. Pour simplifier, nous ne considérerons ici que des problèmes dans lesquels cette variable est une fonction définie sur un intervalle.

2.1 Quelques exemples de calcul des variations

Problème de la reine de Didon : C'est "historiquement" un des premiers problèmes de calcul des variations. Il s'agit de trouver la forme que doit prendre une bandelette souple (initialement en peau de chèvre), de longueur donnée L et attachée aux deux extrémités d'une plage, pour que la surface du domaine délimité d'un côté par la mer et de l'autre par la bandelette soit la plus grande possible.

Ce problème se formalise de la façon suivante : on suppose pour fixer les idées que la mer est l'ensemble $\{(x, y) \in \mathbb{R}^2 \mid y \leq 0\}$, et que la bandelette est attachée aux points $(0, 0)$ et $(1, 0)$. On suppose aussi (mais c'est une restriction) que la bandelette décrit le graphe d'une fonction $x : [0, 1] \to \mathbb{R}$.

Le problème revient alors à maximiser l'aire $f(x) \coloneqq \int_0^1 x(t)dt$ sous la contrainte de longueur $g(x) \coloneqq \int_0^1 \sqrt{1 + (x'(t))^2}dt = L$. L'inconnue est ici la courbe $x : [0, 1] \to \mathbb{R}$. Le critère à optimiser est $f(x)$ tandis que la contrainte est $g(x) = L$.

Le problème des géodésiques Etant donné une surface dans $\mathbb{R}^3$ (la surface de la terre par exemple) et deux points sur cette surface, on cherche le chemin le plus court sur cette surface reliant les deux points. Si on note S cette surface, il s'agit donc de minimiser la longueur d'une courbe $\gamma : [0, 1] \to S$ dont les extrémités sont les points donnés. Une telle courbe s'appelle une géodésique.

Le problème de résistance minimale de Newton Il s'agit de trouver quelle doit être la forme du nez d'un obus de fût cylindrique, pour que celui-ci présente le moins de résistance possible à l'air. La hauteur maximale du nez est fixée (car un nez de longeur "infini"—à la Pinocchio—serait optimal, car de résistance nulle, mais assez difficile à manipuler).

Newton a trouvé la solution à cette question en faisant deux hypothèses : l'une est que le nez doit être "convexe"', c'est-à-dire que, si on représente le nez comme une fonction de deux variables x et y au-dessus du fût dont la base est l'ensemble $\Omega = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$, la fonction $(x, y) \to z(x, y)$ doit être concave. Cette hypothèse, assez naturelle en pratique, permet le calcul de la résistance à l'air :

$$\int_\Omega \frac{dxdy}{1 + \|\nabla z(x, y)\|^2}$$

L'autre hypothèse qu'a fait Newton est aussi très naturelle : puisque le problème est à symétrie radiale, on peut penser que la solution est aussi radiale, i.e., $z = z(\sqrt{x^2 + y^2})$. En passant en coordonnées polaires, la résistance à l'air devient :

$$J(z) = \int_0^1 \frac{\rho d\rho}{1 + (z'(\rho))^2} \ .$$

Le problème consiste alors à minimiser $J(z)$ sur l'ensemble des fonctions concaves $z : [0, 1] \to [0, L]$. Newton a calculé explicitement la solution et a montré—ce qui est assez surprenant—que cette solution a "le bout du nez plat", i.e., que la fonction z optimale doit être constante au voisinage de l'origine.

Tout aussi surprenant, la solution de Newton est en fait fausse : en effet, l'hypothèse que la solution est radiale est erronnée, et on peut trouver des formes (non radiales, bien sûr) qui ont une résistance strictement inférieure à celle donnée par la solution de Newton. La forme de la solution optimale est cependant un problème toujours ouvert...

2.2 Conditions nécessaires d'optimalité

Avant de commencer à parler de conditions nécessaires d'optimalité, nous devons rappeler comment on dérive dans un espace de fonctions.

Différentiabilité

Soit X un espace vectoriel normé et $f : X \to \mathbb{R}$ une application. On rappelle que f est (Fréchet) différentiable en x_0 s'il existe une forme linéaire continue $df(x_0) : X \to \mathbb{R}$ telle que

$$f(x) = f(x_0) + df(x_0)(x - x_0) + \|x - x_0\|\varepsilon(x)$$

où l'application $\varepsilon : X \to \mathbb{R}$ vérifie $\lim_{x\to x_0} \varepsilon(x) = 0$. De plus on dit que f est de classe $\mathcal{C}^1$ sur X si f est différentiable en tout point $x_0 \in X$ et si l'application $x_0 \to df(x_0)$ est continue de X dans X^*, où X^* est l'ensemble des formes linéaires continues de X dans $\mathbb{R}$. On rappelle que X^* est muni de la norme

$$|||L|||_{X^*} = \sup_{x\in X,\ \|x\|\leq 1} |L(x)| \qquad \forall L \in X^* .$$

Le principal exemple que nous considérerons est le cas de l'espace vectoriel $X = \mathcal{C}^1([0,1];\mathbb{R}^N)$ des applications $x : [0,1] \to \mathbb{R}^N$ de classe $\mathcal{C}^1$, muni de la norme

$$\|x\| = \max_{t\in[0,1]} |x(t)| + \max_{t\in[0,1]} |x'(t)| \qquad \forall x \in X ,$$

où $|y|$ désigne la norme euclidienne dans $\mathbb{R}^N$. Rappelons que X, muni de cette norme, est un espace de Banach, c'est-à-dire un espace vectoriel normé complet.

Soit $L : [0,1] \times \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ une application continue et

$$f(x) = \int_0^1 L(t, x(t), x'(t))dt \qquad \forall x \in X .$$

Proposition 2.1. *On suppose que $L = L(t,x,p)$ est de classe $\mathcal{C}^1$ sur $[0,1] \times \mathbb{R}^N \times \mathbb{R}^N$. Alors f est de classe $\mathcal{C}^1$ sur X et*

$$df(x_0)(v) = \int_0^1 \langle \frac{\partial L}{\partial x}(t, x_0(t), x_0'(t)), v(t)\rangle + \langle \frac{\partial L}{\partial p}(t, x_0(t), x_0'(t)), v'(t)\rangle \, dt \qquad \forall v \in X,\ \forall x_0 \in X . \tag{2.1}$$

Preuve. Posons

$$\mathcal{L}(v) = \int_0^1 \langle \frac{\partial L}{\partial x}(t, x_0(t), x_0'(t)), v(t)\rangle + \langle \frac{\partial L}{\partial p}(t, x_0(t), x_0'(t)), v'(t)\rangle \, dt \qquad \forall v \in X .$$

Alors $\mathcal{L}$ est une forme linéaire continue sur X. Nous devons montrer que l'application

$$\varepsilon(x) = \frac{1}{\|x - x_0\|} \left(f(x) - f(x_0) - \mathcal{L}(x - x_0)\right)$$

vérifie $\lim_{x\to x_0} \varepsilon(x) = 0$.

Pour cela, fixons $\delta > 0$ petit et montrons qu'il existe $\eta > 0$ tel que, si $\|x - x_0\| \leq \eta$, on a $|\varepsilon(x)| \leq \delta$. Notons d'abord que l'ensemble $K = \{(t, x_0(t), x_0'(t)),\ t \in [0,1]\}$ est compact dans

$[0,1] \times \mathbb{R}^N \times \mathbb{R}^N$ car x est de classe $\mathcal{C}^1$ sur $[0,1]$. Comme L est de classe $\mathcal{C}^1$, sa différentielle est uniformément continue dans un voisinage de K et donc il existe $\eta > 0$ tel que, pour tout $(y,z) \in \mathbb{R}^N \times \mathbb{R}^N$, et pour tout $t \in [0,1]$ tel que $|y - x_0(t)| + |z - x_0'(t)| \leq \eta$, on a :

$$\Big| L(t,y,z) - L(t,x_0(t),x_0'(t)) - \langle \tfrac{\partial L}{\partial x}(t,x_0(t),x_0'(t)), y - x_0(t)\rangle - \langle \tfrac{\partial L}{\partial p}(t,x_0(t),x_0'(t)), z - x_0'(t)\rangle \Big| \\ \leq \delta(|y - x_0(t)| + |z - x_0'(t)|) \; .$$

En particulier, si $x \in X$ est tel que $\|x - x_0\| \leq \eta$, on a, pour tout $t \in [0,1]$, $|x(t) - x_0(t)| + |x'(t) - x_0'(t)| \leq \eta$, et donc

$$\Big| L(t,x(t),x'(t)) - L(t,x_0(t),x_0'(t)) - \langle \tfrac{\partial L}{\partial x}, x(t) - x_0(t)\rangle + \langle \tfrac{\partial L}{\partial p}, x'(t) - x_0'(t)\rangle \Big| \\ \leq \delta(|x(t) - x_0(t)| + |x'(t) - x_0'(t)|) \leq \delta \|x - x_0\| \; ,$$

où, pour simplifier la formule, on a écrit $\frac{\partial L}{\partial x}$ à la place de $\frac{\partial L}{\partial x}(t,x_0(t),x_0'(t))$ et $\frac{\partial L}{\partial p}$ à la place de $\frac{\partial L}{\partial p}(t,x_0(t),x_0'(t))$. On déduit de l'inégalité triangulaire que

$$\begin{aligned} &|f(x) - f(x_0) - \mathcal{L}(x - x_0)| \\ &\quad \leq \textstyle\int_0^1 \Big| L(t,x(t),x'(t)) - L(t,x_0(t),x_0'(t)) - \langle \frac{\partial L}{\partial x}, x(t) - x_0(t)\rangle + \langle \frac{\partial L}{\partial p}, x'(t) - x_0'(t)\rangle \Big| \, dt \\ &\quad \leq \delta \|x - x_0\| \; , \end{aligned}$$

i.e., $|\varepsilon(x)| \leq \delta$. Nous avons donc montré que $\lim_{x \to x_0} \varepsilon(x) = 0$. Donc f est différentiable en tout point x_0 de X, et $df(x_0)$ est donnée par (2.1).

Reste à prouver que f est de classe $\mathcal{C}^1$ sur X. Soit (x_n) une suite d'éléments de X qui tend vers $x \in X$ pour la norme $\|\cdot\|$. Alors la suite de fonctions continues (x_n) converge uniformément vers x tandis que la suite de fonctions continues (x_n') converge uniformément vers x'. On a, pour tout $v \in X$ avec $\|v\| \leq 1$:

$$\begin{aligned} &|df(x_n)(v) - df(x)(v)| \\ &\leq \textstyle\int_0^1 \big| \frac{\partial L}{\partial x}(t,x_n(t),x_n'(t)) - \frac{\partial L}{\partial x}(t,x(t),x'(t)) \big| \, |v(t)| \\ &\qquad\qquad + \Big| \tfrac{\partial L}{\partial p}(t,x_n(t),x_n'(t)) - \tfrac{\partial L}{\partial p}(t,x(t),x'(t)) \Big| \, |v'(t)| \, dt \\ &\leq \sup_{t \in [0,1]} \Big\{ \big| \tfrac{\partial L}{\partial x}(t,x_n(t),x_n'(t)) - \tfrac{\partial L}{\partial x}(t,x(t),x'(t)) \big| + \Big| \tfrac{\partial L}{\partial p}(t,x_n(t),x_n'(t)) - \tfrac{\partial L}{\partial p}(t,x(t),x'(t)) \Big| \Big\} \end{aligned}$$

ce qui tend vers 0 lorsque $n \to +\infty$ puisque L est de classe $\mathcal{C}^1$ et que (x_n) et (x_n') convergent uniformément vers x et x' respectivement. □

Problèmes sans contrainte

Soient $A, B \in \mathbb{R}^N$. On considère le problème de minimisation sur l'espace $X = \mathcal{C}(0,1;\mathbb{R}^N)$:

$$(\mathcal{P}) \qquad \inf_{x \in X, \; x(0)=A, \; x(1)=B} \int_0^1 L(t,x(t),x'(t))dt \; .$$

Nous ne discuterons pratiquement pas (sauf dans le cas convexe) la question de l'existence d'un minimum : celle-ci est assez délicate. Commençons par étudier les conditions nécessaires d'optimalité.

Théorème 2.2 (Equation d'Euler). *Si $L = L(t,x,p)$ est de classe $\mathcal{C}^1$ sur $[0,1]\times\mathbb{R}^N\times\mathbb{R}^N$ et si la fonction $x\in X$ est un minimum du problème $(\mathcal{P})$, alors la fonction $t\to\frac{\partial L}{\partial p}(t,x(t),x'(t))$ est de classe $\mathcal{C}^1$ sur $[0,1]$ et*

$$\frac{d}{dt}\,\frac{\partial L}{\partial p}(t,x(t),x'(t)) = \frac{\partial L}{\partial x}(t,x(t),x'(t)) \qquad \forall t\in[0,1]\ . \tag{2.2}$$

La relation (2.2) s'appelle l'équation d'Euler (ou d'Euler-Lagrange) du problème. On appelle extrémale de $(\mathcal{P})$ toute application $x:]a,b[\to\mathbb{R}$ vérifiant l'équation (2.2) sur un intervalle (non vide) $]a,b[$.

La preuve du théorème repose sur le lemme suivant :

Lemme 2.3 (Dubois-Raymond). *Soit $\phi:[0,1]\to\mathbb{R}^N$ une application continue. On suppose que, pour tout $v\in\mathcal{C}^1([0,1];\mathbb{R}^N)$ tel que $v(0)=v(1)=0$, on a $\int_0^1\langle\phi(t),v'(t)\rangle dt=0$. Alors la fonction ϕ est constante sur $[0,1]$.*

Remarque : La réciproque est évidente : si ϕ est une constante, alors

$$\int_0^1\langle\phi,v'(t)\rangle dt = \langle\phi,\int_0^1 v'(t)dt\rangle = \langle\phi,v(1)-v(0)\rangle = 0\ .$$

Preuve du Lemme 2.3. Posons $c=\int_0^1\phi(s)ds$ et $v(t)=\int_0^t\phi(s)ds-ct$. Alors v est de classe $\mathcal{C}^1$ sur $[0,1]$, $v(0)=v(1)=0$ et $v'(t)=\phi(t)-c$. Donc

$$\int_0^1|\phi(t)-c|^2dt=\int_0^1\langle\phi(t)-c,v'(t)\rangle dt=\int_0^1\langle\phi(t),v'(t)\rangle dt-\langle c,\int_0^1 v'(t)\rangle dt=0\ ,$$

la première intégrale étant nulle par hypothèse, la seconde d'après la remarque ci-dessus. Comme $t\to\phi(t)-c$ est continue, on en déduit que $\phi(t)=c$ pour tout $t\in[0,1]$. □

Preuve du théorème 2.2. Rappelons que $X=\mathcal{C}^1([0,1];\mathbb{R}^N)$ est muni de la norme

$$\|x\|=\max_{t\in[0,1]}|x(t)|+\max_{t\in[0,1]}|x'(t)| \qquad \forall x\in X\ ,$$

où $|y|$ désigne la norme euclidienne dans $\mathbb{R}^N$.

Soit $v:[0,1]\to\mathbb{R}^N$ de classe $\mathcal{C}^1$ sur $[0,1]$, tel que $v(0)=v(1)=0$. Alors, pour tout $\lambda\in\mathbb{R}$, $x_\lambda=x+\lambda v$ appartient à X et vérifie $x_\lambda(0)=A$ et $x_\lambda(1)=B$. Par définition de x, l'application $\lambda\to f(x_\lambda)$ a un minimum en $\lambda=0$, et donc a une dérivée nulle en $\lambda=0$. Puisque f est différentiable (cf. la proposition 2.1) et que l'application $\lambda\to x_\lambda$ est dérivable, de dérivée v, on déduit du théorème de dérivation des fonctions composées que

$$\frac{d}{d\lambda}f(x_\lambda)_{|_{\lambda=0}}=df(x)(v)=0\ .$$

D'après la Proposition 2.1 on a

$$df(x)(v)=\int_0^1\langle\frac{\partial L}{\partial x}(t,x(t),x'(t)),v(t)\rangle+\langle\frac{\partial L}{\partial p}(t,x(t),x'(t)),v'(t)\rangle\ dt\ .$$

Posons $\phi_1(t) = \int_0^t \frac{\partial L}{\partial x}(s, x(s), x'(s))ds$ et $\phi(t) = -\phi_1(t) + \frac{\partial L}{\partial p}(t, x(t), x'(t))$. Notons que ϕ_1 est de classe $\mathcal{C}^1$ tandis que ϕ est continue. En faisant une intégration par partie, on obtient

$$\int_0^1 \langle \frac{\partial L}{\partial x}(t, x(t), x'(t)), v(t)\rangle dt = [\langle \phi_1(t), v(t)\rangle]_0^1 - \int_0^1 \langle \phi_1(t), v'(t)\rangle dt = -\int_0^1 \langle \phi_1(t), v'(t)\rangle dt \ .$$

Donc

$$0 = df(x)(v) = \int_0^1 \langle -\phi_1(t), v'(t)\rangle + \langle \frac{\partial L}{\partial p}(t, x(t), x'(t)), v'(t)\rangle \ dt = \int_0^1 \langle \phi(t), v'(t)\rangle \ dt \ .$$

Comme cette égalité est vraie pour tout v de classe $\mathcal{C}^1$ tel que $v(0) = v(1) = 0$, on déduit du Lemme de Dubois-Raymond que ϕ est constante.

Donc $\frac{\partial L}{\partial p}(t, x(t), x'(t)) = \phi + \phi_1(t)$ est de classe $\mathcal{C}^1$ et

$$\frac{d}{dt}\frac{\partial L}{\partial p}(t, x(t), x'(t)) = \phi_1'(t) = \frac{\partial L}{\partial x}(t, x(t), x'(t)) \ .$$

□

Lorsque le critère est convexe, l'équation d'Euler devient une condition suffisante d'optimalité :

Proposition 2.4. *On suppose que L est de classe $\mathcal{C}^1$ et que, pour tout $t \in [0,1]$, la fonction $(x,p) \to L(t,x,p)$ est convexe. Si $x \in X$ vérifie l'équation d'Euler (2.2), ainsi que les conditions au bord $x(0) = A$ et $x(1) = B$, alors x est un minimum du problème $(\mathcal{P})$.*

Preuve. Fixons $y \in X$ tel que $y(0) = A$ et $y(1) = B$ et vérifions que $f(y) \geq f(x)$, où

$$f(z) = \int_0^1 L(t, z(t), z'(t))dt \qquad \forall z \in X \ .$$

Notons d'abord que f est convexe sur X puisque L l'est. Par conséquent l'application $\phi : \mathbb{R} \to \mathbb{R}$ définie par $\phi(t) = f((1-t)x + ty)$ est aussi convexe. Comme f est différentiable en tout point de X, ϕ est dérivable sur $\mathbb{R}$. Pour montrer que $f(y) \geq f(x)$, il suffit de montrer que 0 est un minimum de ϕ et donc, par convexité, de vérifier que $\phi'(0) = 0$: par théorème des fonctions composées, on a

$$\begin{array}{rcl} \phi'(0) = df(x)(y-x) & = & \int_0^1 \langle \frac{\partial L}{\partial x}, (y-x)(t)\rangle + \langle \frac{\partial L}{\partial p}, (y-x)'(t)\rangle dt \\ & = & \int_0^1 \langle \frac{\partial L}{\partial x}, (y-x)(t)\rangle + \langle -\frac{d}{dt}\frac{\partial L}{\partial p}, (y-x)(t)\rangle \ dt \ = \ 0 \end{array}$$

après une intégration par parties. □

Comme nous allons le voir, l'équation d'Euler conduit le plus souvent à la résolution d'une équation différentielle d'ordre 2.

Example 2.5 (Mouvement d'une particule soumise à gravitation)**.** Le principe de Hamilton affirme que le mouvement d'une particule de masse $m > 0$, de poids mg est régi par l'équation d'Euler pour l'énergie

$$\frac{m}{2}\int_0^1 |x'(t)|^2 - 2gx(t) \ dt \ ,$$

où $x(t) \in \mathbb{R}$ est la hauteur de la particule et x' la vitesse de son déplacement vertical. L'équation d'Euler s'écrit

$$mx''(t) = -mg \qquad \forall t \in [0,1] \ .$$

La solution obtenue est donc une parabole. Notons que la fonction $L = L(x,p) = \frac{m}{2}(|p|^2 - 2gx)$ est convexe, et donc toute solution de l'équation ci-dessus est un minimum de $(\mathcal{P})$ pour $A = x(0)$ et $B = x(1)$.

Application aux géodésiques sur une surface de $\mathbb{R}^3$

On suppose qu'une surface S de $\mathbb{R}^3$ est donnée par une paramétrisation $\Phi : U \to \mathbb{R}^3$, où U est un ouvert : $S = \Phi(U)$. L'application Φ est supposée régulière (disons $\mathcal{C}^\infty$), avec $d\Phi(x)$ est de rang 2 sur U. Par exemple,

— la sphère de centre 0 et de rayon 1 est donnée (en coordonnées sphériques) par

$$\Phi(\varphi, \theta) = (\cos(\varphi)\sin(\theta), \sin(\varphi)\sin(\theta), \cos(\theta)) \qquad (\varphi, \theta) \in \mathbb{R}^2 \ .$$

— Le cylindre (en coordonnées cylindriques)

$$\Phi(\varphi, z) = (\cos(\varphi), \sin(\varphi), z) \qquad (\varphi, z) \in \mathbb{R}^2$$

Si $x : [0,1] \to U$ est une fonction différentiable, alors $\gamma(t) = \Phi(x(t))$ est une courbe sur la surface, de longueur

$$J(\gamma) = \int_0^1 |\gamma'(t)|dt = \int_0^1 |(\Phi \circ x)'(t)|dt \ .$$

La distance géodésique entre deux point $\Phi(A)$ et $\Phi(B)$ de la variété (où $A, B \in U$) s'exprime alors par le problème de minimisation suivant :

$$\min \left\{ \int_0^1 |(\Phi \circ x)'(t)|dt \mid x \in \mathcal{C}^1([0,1]; U),\ x(0) = A,\ x(1) = B \right\} \tag{2.3}$$

Ce problème a un très grand nombre de solutions puisque, si $\theta : [0,1] \to [0,1]$ est $\mathcal{C}^1$, avec $\theta' > 0$ et $\theta(0) = 0$, $\theta(1) = 1$, et si on note $\gamma = \Phi \circ x$, alors

$$J(\gamma \circ \theta) = \int_0^1 |\gamma'(\theta(t))|\theta'(t)dt = J(\gamma)$$

On a donc intérêt à trouver une formulation réduisant ce nombre de solutions. Soit

$$\min \left\{ \int_0^1 |(\Phi \circ x)'(t)|^2 dt \mid x \in \mathcal{C}^1([0,1]; U),\ x(0) = A,\ x(1) = B \right\} \tag{2.4}$$

Nous allons montrer que les problèmes (2.3) et (2.4) sont intimement liés.

Lemme 2.6. *On a*

$$\begin{aligned} &\inf \left\{ \textstyle\int_0^1 |(\Phi \circ x)'(t)|dt \mid x \in \mathcal{C}^1([0,1]; U),\ x(0) = A,\ x(1) = B \right\} \\ &\quad = \inf \left\{ \left(\textstyle\int_0^1 |(\Phi \circ x)'(t)|^2 dt \right)^{\frac{1}{2}} \mid x \in \mathcal{C}^1([0,1]; U),\ x(0) = A,\ x(1) = B \right\} \end{aligned}$$

et, si x est un minimum du problème (2.4), alors x est un minimum du problème (2.3). De plus, si x est une extrémale pour (2.4), alors x est une extrémale pour (2.3), x est de classe $\mathcal{C}^2$ sur $[0,1]$ et

$$\frac{d}{dt}|(\Phi \circ x)'(t)|^2 = 0 \ .$$

Remarque : Inversement, il n'est pas vrai que, si x est un minimum pour le problème (2.3), alors x est un minimum pour (2.4). Cela est vrai après renormalisation (cf. la preuve du Lemme).

On appelle géodésique toute extrémale de (2.4), i.e., toute solution de l'équation d'Euler associée à $L(x,p) = |D\Phi(x)p|^2$.

Preuve. D'après Cauchy-Schwarz on a

$$\int_0^1 |(\Phi \circ x)'(t)|dt \leq \left(\int_0^1 |(\Phi \circ x)'(t)|^2 dt \right)^{\frac{1}{2}} \tag{2.5}$$

pour tout $x \in \mathcal{C}^1([0,1];U)$. Inversement, soit $x \in \mathcal{C}^1([0,1];U)$, avec $x(0) = A$ et $x(0) = B$, et supposons que $x'(t) \neq 0$. En fait cette hypothèse n'est pas restrictive car on peut montrer (mais nous ne le ferons pas) que l'ensemble des x possédant cette propriété est dense dans $\mathcal{C}^1([0,1];U)$. Posons $\gamma = \Phi \circ x$ et $s(t) = \int_0^t |\gamma'(\tau)|d\tau$. Alors par hypothèse s est une bijection de $[0,1]$ dans $[0,a]$ (où $a = \int_0^1 |\gamma'(\tau)|d\tau > 0$) d'inverse θ_1 qui est de classe $\mathcal{C}^1$. Posons maintenant $\theta(t) = \theta_1(t/a)$, qui est défini sur $[0,1]$. Alors $x_1 = x \circ \theta$ vérifie $x_1(0) = x(0) = A$ et $x_1(1) = x(1) = B$ et

$$|(\Phi \circ x_1)'| = |(\gamma \circ \theta)'| = |\gamma'(\theta)|\theta' = \frac{1}{a} .$$

Donc

$$\int_0^1 |(\Phi \circ x)'(t)|dt = \int_0^1 |(\Phi \circ x_1)'(t)|dt = \left(\int_0^1 |(\Phi \circ x_1)'(t)|^2 dt \right)^{\frac{1}{2}}$$

Cela montre l'égalité entre les infima.

De plus, (2.5) implique que, si x est un minimum de (2.4), alors x est un minimum de (2.3) puisque les infima coïncident.

Supposons maintenant que x est extrémal pour (2.4). Posons $L(x,p) = |D\Phi(x)p|^2$. Notons que L est homogène de degré 2 en p, i.e., $L(x,\lambda p) = \lambda^2 L(x,p)$ pour tout $\lambda > 0$. En dérivant cette égalité par rapport à λ en $\lambda = 1$, on obtient

$$\langle \frac{\partial L}{\partial p}(x,p), p \rangle = 2L(x,p) \qquad \forall (x,p) \in \mathbb{R}^N \times \mathbb{R}^N .$$

Notons ensuite que $\frac{\partial L}{\partial p}(x,p) = 2(D\Phi(x))^T D\Phi(x)p$, où la matrice $(D\Phi(x))^T D\Phi(x)$ est inversible puisque $D\Phi(x)$ est de rang 2. Comme $t \to \frac{\partial L}{\partial p}(x(t), x'(t))$ est de classe $\mathcal{C}^1$, l'application

$$t \to x'(t) = \frac{1}{2}[(D\Phi(x))^T D\Phi(x)]^{-1} \frac{\partial L}{\partial p}(x(t), x'(t))$$

l'est aussi, i.e., x est de classe $\mathcal{C}^2$. Calculons maintenant

$$\begin{array}{rcll} \frac{d}{dt} L(x(t), x'(t)) & = & \langle \frac{\partial L}{\partial x}, x' \rangle + \langle \frac{\partial L}{\partial p}, x'' \rangle & \\ & = & \langle \frac{d}{dt} \frac{\partial L}{\partial p}, x' \rangle + \langle \frac{\partial L}{\partial p}, x'' \rangle & \text{(d'après l'équation d'Euler)} \\ & = & \frac{d}{dt} \langle \frac{\partial L}{\partial p}, x' \rangle = \frac{d}{dt} 2L(x(t), x'(t)) & \text{(par homogénéité de } L(x,\cdot)) \end{array}$$

D'où $\frac{d}{dt} L(x(t), x'(t)) = 0$. Enfin, comme x est extrémal pour (2.4), on a

$$\frac{d}{dt} 2(D\Phi(x))^T x' = \frac{\partial L}{\partial x}(x, x')$$

et, comme $t \to L(x(t), x'(t))$ est constante sur $[0,1]$,

$$\frac{d}{dt} \frac{\partial \sqrt{L}}{\partial p}(x,x') = \frac{1}{\sqrt{L(x,x')}} \frac{d}{dt} (D\Phi(x))^T x' = \frac{1}{2\sqrt{L(x,x')}} \frac{\partial L}{\partial x}(x,x') = \frac{\partial \sqrt{L}}{\partial x}(x,x') .$$

Donc x est extrémale pour (2.3). □

Le cas du cylindrique est particulièrement simple à étudier : dans ce cas, si $x(t) = (\varphi(t), z(t))$, alors

$$|(\Phi \circ x)'(t)|^2 = \big|(-\sin(\varphi(t))\varphi'(t), \cos(\varphi(t))\varphi'(t), z'(t))\big|^2 = (\varphi'(t))^2 + (z'(t))^2 \ .$$

L'équation d'Euler est donc

$$\frac{d}{dt}\varphi'(t) = 0 \qquad \text{et} \qquad \frac{d}{dt}z'(t) = 0 \ ,$$

c'est-à-dire que les géodésiques du cylindre sont les mouvements à vitesse constante en φ et en z.

Dans le cas de la sphère, on a

$$L(\varphi, \theta, p_\varphi, p_\theta) = (\sin(\theta))^2 (p_\varphi)^2 + (p_\theta)^2 \ .$$

Considèrons deux points $A = (\varphi_A, \theta_A)$ et $B = (\varphi_B, \theta_B)$. Quitte à effectuer une rotation, on peut supposer que $\varphi_A = \varphi_B$. L'équation d'Euler pour le problème (2.4) dit que, si $x(t) = (\varphi(t), \theta(t))$ est une extrémale du problème, alors

$$\left\{ \begin{array}{ll} (i) & \frac{d}{dt}\ \varphi'(t)(\sin(\theta(t)))^2 = 0 \\ (ii) & \frac{d}{dt}\ \theta'(t) = (\varphi'(t))^2 \sin(\theta(t))\cos(\theta(t)) \end{array} \right.$$

De (i), on tire que $\varphi'(t)(\sin(\theta))^2 = c$. Donc φ est monotone et, comme $\varphi(0) = \varphi_A = \varphi_B = \varphi(1)$, on a donc φ constant. Mais alors, (ii) dit que θ a une vitesse constante. Les géodésiques sur la sphère sont donc des portions de cercles.

Problèmes avec une contrainte d'égalité

On travaille toujours dans l'ensemble $X = \mathcal{C}^1(0, 1; \mathbb{R}^N)$. Soit deux fonctions : $f : X \to \mathbb{R}$ (le critère) et $g : X \to \mathbb{R}$ (la contrainte) de la forme

$$f(x) = \int_0^1 L(t, x(t), x'(t))dt \qquad \text{et} \qquad g(x) = \int_0^1 M(t, x(t), x'(t))dt \ ,$$

où $L, M : [0, 1] \times \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ sont continues. On étudie le problème de minimisation sous contrainte :

$$(\mathcal{C}) \qquad \min \{f(x) \mid x \in X,\ x(0) = A,\ x(1) = B,\ g(x) = 0\}$$

Théorème 2.7. *On suppose que $L, M : [0, 1] \times \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ sont de classe $\mathcal{C}^1$ et que x est un minimum du problème $(\mathcal{C})$. S'il existe $z \in X$ tel que*

$$z(0) = z(1) = 0 \qquad \text{et} \qquad dg(x)(z) \neq 0 \ , \tag{2.6}$$

alors il existe un multiplicateur $\lambda \in \mathbb{R}$ tel que l'application $t \to \frac{\partial(L+\lambda M)}{\partial p}(t, x(t), x'(t))$ est de classe $\mathcal{C}^1$ sur $[0, 1]$ avec

$$\frac{d}{dt}\ \frac{\partial(L + \lambda M)}{\partial p}(t, x(t), x'(t)) = \frac{\partial(L + \lambda M)}{\partial x}(t, x(t), x'(t)) \qquad \forall t \in [0, 1] \ . \tag{2.7}$$

Remarque : la condition (2.6) n'est rien d'autre qu'une condition de qualification de la contrainte $K = \{x \in X \mid x \in X,\ x(0) = A,\ x(1) = B,\ g(x) = 0\}$.

Preuve du théorème. Soit $v \in X$ tel que $v(0) = v(1) = 0$. Introduisons l'application $\Phi : \mathbb{R}^2 \to \mathbb{R}^2$ définie par

$$\Phi(s,t) = (f(x+sv+tz), g(x+sv+tz)) \qquad \forall (s,t) \in \mathbb{R}^2 ,$$

où z est défini par (2.6). Alors de la proposition 2.1 et du théorème de dérivation des fonctions composée on déduit que Φ est de classe $\mathcal{C}^1$ sur $\mathbb{R}^2$. Nous affirmons que le jacobien de Φ est nul en $(0,0)$. En effet, sinon, le théorème d'inversion local affirme qu'il existe deux ouverts $\mathcal{O}$ et $\mathcal{O}'$ de $\mathbb{R}^2$ contenant respectivement $(0,0)$ et $(f(x),0)$, tel que Φ est un difféomorphisme de $\mathcal{O}$ dans $\mathcal{O}'$. En particulier, pour $\epsilon > 0$ petit, le point $(f(x)-\epsilon, 0)$ appartient à $\mathcal{O}'$, et donc il existe $(s,t) \in \mathcal{O}$ tel que $\Phi(s,t) = (f(x)-\epsilon, 0)$. Mais alors la fonction $y = x + sv + tz$ vérifie les contraintes $y(0) = A$ et $y(1) = B$ et $g(y) = 0$ et est telle que $f(y) = f(x) - \epsilon < f(x)$, ce qui contredit le fait que x est un minimum de $(\mathcal{C})$. On en déduit que le jacobien de Φ doit être nul en $(0,0)$, i.e.,

$$df(x)(v)dg(x)(z) - df(x)(z)dg(x)v) = 0 .$$

En posant $\lambda = -\frac{df(x)(z)}{dg(x)(z)}$ (ce qui est possible puisque $dg(x)(z) \neq 0$ par hypothèse (2.6)), on a donc :

$$d(f-\lambda g)(x)(v) = 0 \qquad \forall v \in X,\ v(0) = v(1) = 0 .$$

On peut alors conclure la démonstration comme pour le théorème 2.2. □

Example 2.8 (Problème de la reine de Didon)**.** Il s'agit de maximiser l'aire enclose sous le graphe de la fonction $x : [0,1] \to \mathbb{R}$ sous la contrainte que ce graphe soit de longueur L et que $x(0) = x(1) = 0$:

$$\max \left\{ \int_0^1 x(t)dt \mid x(0) = x(1) = 0 \text{ et } \int_0^1 \sqrt{1 + (x'(t))^2}dt = L \right\} .$$

Si on pose $L(x,p) = -x$ et $M(x,p) = \sqrt{1+p^2}$, alors le problème est de la forme $(\mathcal{C})$. Une extrémale x de ce problème doit vérifier, pour un certain $\lambda \in \mathbb{R}$,

$$\frac{d}{dt} \frac{\lambda x'(t)}{\sqrt{1 + (x'(t))^2}} = -1 \qquad \forall t \in [0,1] .$$

Notons que $\lambda \neq 0$. Après un petit calcul, on en déduit qu'il existe une constance $c \in \mathbb{R}$ telle que

$$x'(t) = \frac{(c-t)/\lambda}{\sqrt{1 - (c-t)^2/\lambda^2}} \qquad \forall t \in [0,1] ,$$

soit

$$x(t) = \lambda \sqrt{1 - (c-t)^2/\lambda^2} + cte \tag{2.8}$$

Comme $x(0) = x(1) = 0$, on doit avoir $c = 1/2$. De plus, comme on cherche à maximiser $\int_0^1 x(t)dt$,on a intérêt à avoir $x(t) \geq 0$ pour tout $t \in [0,1]$, et donc $x'(0) \geq 0$. D'où $\lambda > 0$. Le graphe de x est donc une portion de cercle de rayon λ, centrée en $(1/2, -\sqrt{\lambda^2 - 1/4})$. Comme on veut que le graphe de x soit de longueur L, on a $2\lambda \sin(L/2\lambda) = 1$, où l'angle $L/2\lambda$ doit être inférieur à $\pi/2$. Ce problème possède donc une extrémale seulement si $1 < L < \pi/2$, extrémale donnée par (2.8) pour $c = 1/2$ et λ l'unique solution de $2\lambda \sin(L/2\lambda) = 1$ telle que $L/2\lambda < \pi/2$.

3 Contrôle optimal

Le contrôle optimal est une généralisation du calcul des variations tel que vu dans la partie précédente, mais avec en plus une contrainte sur la dérivée de la fonction sur laquelle on minimise. Plus précisément, on cherche à minimiser une quantité de la forme

$$\int_0^T L(t, x(t), u(t))\ dt + g(x(T))$$

sous la contrainte que le couple $(x(\cdot), u(\cdot))$ vérifie l'équation différentielle ordinaire (EDO) :

$$\begin{cases} \dot{x}(t) = f(t, x(t), u(t)), & t \in [0, T] \\ x(0) = x_0 \end{cases}$$

Dans toute cette partie, nous supposons que U est un espace métrique compact, que $f : [0, T] \times \mathbb{R}^N \times U \to \mathbb{R}^N$ est globalement continue et uniformément Lipschitzienne par rapport à la variable d'espace : $\exists K > 0$ tel que

$$\|f(t, x, u) - f(t, y, u)\| \leq K\|x - y\| \qquad \forall (t, x, y, u) \in [0, T] \times \mathbb{R}^N \times \mathbb{R}^N \times U.$$

Nous supposons également que $L : [0, T] \times \mathbb{R}^N \times U \to \mathbb{R}$ et $g : \mathbb{R}^N \to \mathbb{R}$ sont continues.

3.1 Le théorème de Cauchy-Lipschitz

En théorie du contrôle, il est rare que le contrôle $u(\cdot)$ soit continu en temps (il est souvent "bang-bang", c'est-à-dire constant par morceaux). C'est pourquoi on doit développer un résultat spécifique d'existence et d'unicité de solution pour l'équation différentielle ordinaire (EDO).

Définissons pour cela le cadre fonctionnel. Notons $W^{1,\infty}([0, T], \mathbb{R}^d)$ l'ensemble des primitives de fonction L^∞ à valeurs dans $\mathbb{R}^d$. Nous admettrons que cet espace coïncide avec l'ensemble des fonctions lipschitziennes de $[0, T]$ dans $\mathbb{R}^d$. Nous admettrons également qu'une fonction lipschitienne est dérivable presque partout et est la primitive de sa dérivée (théorème de Rademacher).

Fixons $t_0 \in [0, T]$. Un contrôle en temps initial $t_0 \in [0, T]$ est une application mesurable $u : [t_0, T] \to U$. Pour une position initiale $(t_0, x_0) \in [0, T] \times \mathbb{R}^N$ donnée on appelle solution de l'équation différentielle ordinaire (EDO)

$$\begin{cases} \dot{x}(t) = f(t, x(t), u(t)), & t \in [t_0, T] \\ x(0) = x_0 \end{cases} \tag{3.1}$$

une application $x \in W^{1,\infty}([t_0, T], \mathbb{R}^N)$ qui vérifie la relation $\dot{x}(t) = f(t, x(t), u(t))$ presque partout sur $[t_0, T]$ et la condition initiale $x(t_0) = x_0$.

Théorème 3.1 (Cauchy-Lipschitz)**.** *Pour toute position initiale $(t_0, x_0) \in [0, T] \times \mathbb{R}^N$ et pour tout contrôle $u : [t_0, T] \to U$, il existe une unique solution de l'EDO* (3.1)*.*

Idée de la preuve. C'est la même que pour le théorème de Cauchy-Lipschitz classique. Elle consiste à montrer que l'application $\Phi : C^0([t_0, T], \mathbb{R}^N) \to C^0([t_0, T], \mathbb{R}^N)$ définie par

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

est contractante (pour la norme $\|\cdot\|_\infty$), et donc possède un unique point fixe (puisque l'espace $C^0([t_0,T],\mathbb{R}^N)$ muni de la norme $\|\cdot\|_\infty$ est un espace complet). Cela est vrai pourvu que $T-t_0$ soit suffisamment petit. Notons que le point fixe x de Φ est bien dans $W^{1,\infty}$ puisque primitive de la fonction mesurable et bornée $s \to f(s,x(s),u(s))$ (mais pas forcément continue, puisque $u(\cdot)$ ne l'est pas).

Pour $T-t_0$ grand, on "recolle" les bouts de solutions comme pour le théorème de Cauchy-Lipschitz usuel. □

Rappelons le théorème de Cauchy-Peano, dont nous aurons besoin plus bas :

Théorème 3.2 (de Cauchy-Peano)**.** *Soit $F : [0,T]\times\mathbb{R}^N \to \mathbb{R}$ une application continue à croissance au plus linéaire : $\exists M > 0$,*

$$\|F(t,x)\| \leq M(1+\|x\|) \qquad \forall (t,x) \in [0,T]\times\mathbb{R}^N.$$

Alors l'EDO

$$\begin{cases} \dot{x}(t) = F(t,x(t)), & t \in [t_0,T] \\ x(0) = x_0 \end{cases}$$

possède au moins une solution (qui est alors de classe C^1).

Remarque : il n'y a pas unicité de la solution en général.

3.2 Le principe du maximum de Pontryagin

Soit $x_0 \in \mathbb{R}^N$ une condition initiale fixée. On considère le problème de contrôle optimal :

$$\inf_{u(\cdot)} \int_0^T L(t,x(t),u(t))\ dt + g(x(T))$$

sous la contrainte que $u : [0,T] \to U$ est mesurable et que $x(\cdot)$ est l'unique solution de l'EDO

$$\begin{cases} \dot{x}(t) = f(t,x(t),u(t)), & t \in [0,T] \\ x(0) = x_0 \end{cases}$$

On définit le *Hamiltonien du système* $H : [0,T]\times\mathbb{R}^N\times\mathbb{R}^N \to \mathbb{R}$ par

$$H(t,x,p) \coloneqq \sup_{u\in U} \left\{-\langle p, f(t,x,u)\rangle - L(t,x,u)\right\}.$$

On suppose que H est de classe C^1.

Théorème 3.3 (Principe du maximum)**.** *Si (x^*,u^*) est optimal dans le problème ci-dessus, alors il existe une application $p^* : [0,T] \to \mathbb{R}^N$ de classe C^1 telle que le couple (x^*,p^*) vérifie le système*

$$\begin{cases} \dot{x^*}(t) = -\dfrac{\partial H}{\partial p}(t,x^*(t),p^*(t)), & t \in [0,T] \\ \dot{p^*}(t) = \dfrac{\partial H}{\partial x}(t,x^*(t),p^*(t)), & t \in [0,T] \\ x(0) = x_0,\ p^*(T) = \dfrac{\partial g}{\partial x}(x^*(T)). \end{cases}$$

De plus

$$-\langle p^*(t), f(t,x^*(t),u^*(t))\rangle - L(t,x^*(t),u^*(t)) = H(t,x^*(t),p^*(t)) \qquad t \in [0,T].$$

La preuve de ce résultat est assez délicate : nous renvoyons le lecteur aux monographies de Cesari et de Fleming-Rishel par exemple.

3.3 Le principe de programmation dynamique

Définissons la fonction valeur $V : [0,T] \times \mathbb{R}^N \to \mathbb{R}$ par

$$V(t_0, x_0) \coloneqq \inf_{u(\cdot)} \int_{t_0}^{T} L(t, x(t), u(t))\ dt + g(x(T) \tag{3.2}$$

sous la contrainte que $u : [0,T] \to U$ est mesurable et que $x(\cdot)$ est l'unique solution de l'EDO

$$\begin{cases} \dot{x}(t) = f(t, x(t), u(t)), & t \in [t_0, T] \\ x(t_0) = x_0 \end{cases}$$

Théorème 3.4. *On a, pour tout $0 \leq t_0 < t_1 \leq T$,*

$$V(t_0, x_0) = \inf_{u(\cdot)} \int_{t_0}^{t_1} L(t, x(t), u(t))\ dt + V(t_1, x(t_1))$$

sous la contrainte que le couple $(x(\cdot), u(\cdot))$ vérifie l'équation différentielle

$$\begin{cases} \dot{x}(t) = f(t, x(t), u(t)), & t \in [t_0, t_1] \\ x(t_0) = x_0 \end{cases}$$

Preuve. Appelons $W(t_0, x_0)$ la quantité

$$W(t_0, x_0) \coloneqq \inf_{u(\cdot)} \int_{t_0}^{t_1} L(t, x(t), u(t))\ dt + V(t_1, x(t_1))$$

On veut montrer l'égalité $W = V$.

Soit $\varepsilon > 0$ et $u(\cdot)$ $\varepsilon-$optimal pour $V(t_0, x_0)$. Alors

$$\begin{aligned} V(t_0, x_0) - \varepsilon \ &\geq \ \int_{t_0}^{t_1} L(t, x(t), u(t))\ dt + \int_{t_1}^{T} L(t, x(t), u(t))\ dt + g(x(T)) \\ &\geq \ \int_{t_0}^{t_1} L(t, x(t), u(t))\ dt + V(t_1, x(t_1)) \end{aligned}$$

puisque $x(\cdot)_{|_{[t_1,T]}}$ est l'unique solution de l'EDO avec contrôle $u(\cdot)_{|_{[t_1,T]}}$ et donnée initiale $(t_1, x(t_1))$. Donc

$$V(t_0, x_0) - \varepsilon \ \geq \int_{t_0}^{t_1} L(t, x(t), u(t))\ dt + V(t_1, x(t_1)) \ \geq \ W(t_0, x_0),$$

ce qui prouve l'inégalité $V \geq W$ car ε est arbitraire.

Inversement, soit $\varepsilon > 0$ et (x_1, u_1) $\varepsilon-$optimal pour $W(t_0, x_0)$. Choisissons également (x_2, u_2) $\varepsilon-$optimal pour $V(t_1, x_1(t_1))$ et posons

$$(x(t), u(t)) = \begin{cases} (x_1(t), u_1(t)) & \text{si } t \in [t_0, t_1] \\ (x_2(t), u_2(t)) & \text{si } t \in]t_1, T] \end{cases}$$

Alors le couple $x(\cdot)$ est l'unique solution de l'EDO avec contrôle u et donnée initiale (t_0, x_0). Donc

$$\begin{aligned} V(t_0, x_0) \ &\leq \ \int_{t_0}^{T} L(t, x(t), u(t))\ dt + g(x(T)) \\ &= \ \int_{t_0}^{t_1} L(t, x_1(t), u_1(t))\ dt + \int_{t_1}^{T} L(t, x_2(t), u_2(t))\ dt + g(x_2(T)) \\ &\leq \ \int_{t_0}^{t_1} L(t, x_1(t), u_1(t))\ dt + V(t_1, x_1(t_1)) + \varepsilon \qquad \text{(par définition de } (x_2, u_2)) \\ &\leq \ W(t_0, x_0) + 2\varepsilon \qquad \text{(par définition de } (x_1, u_1)) \end{aligned}$$

Cela montre que $V(t_0, x_0) \leq W(t_0, x_0)$ car ε est arbitraire. D'où l'égalité $V = W$. □

3.4 Lien avec les équations de Hamilton-Jacobi

Du fait du caractère continu du temps, le principe de programmation dynamique peut sembler peu utile : il n'est en effet plus question de résoudre la fonction valeur par induction rétrograde comme nous l'avons fait en temps discret. L'intérêt de la programmation dynamique est qu'elle permet de montrer que la fonction valeur vérifie une équation aux dérivées partielles : l'équation de *Hamilton-Jacobi*. Pour cela, rappelons que le *Hamiltonien du système* $H : [0,T] \times \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ est défini par

$$H(t,x,p) \coloneqq \sup_{u \in U} \{-\langle p, f(t,x,u)\rangle - L(t,x,u)\}.$$

Vu nos hypothèses de continuité sur f et L, l'application H est continue.

Théorème 3.5. *On suppose que la fonction valeur V définie par (3.2) est de classe C^1 sur $]0,T[\times\mathbb{R}^N$ et C^0 sur $[0,T]\times\mathbb{R}^N$. Alors V satisfait l'équation de Hamilton-Jacobi (HJ)*

$$\begin{cases} -\dfrac{\partial V}{\partial t}(t,x) + H(t,x,\dfrac{\partial V}{\partial x}(t,x)) = 0 \qquad \forall (t,x) \in]0,T[\times\mathbb{R}^N \\ V(T,x) = g(x) \qquad \forall x \in \mathbb{R}^N \end{cases}$$

Le théorème ci-dessus a une portée limitée, puisqu'il repose sur l'hypothèse *a priori* que la fonction valeur est assez régulière (C^1), ce qui est rarement le cas. Il possède cependant un double intérêt. D'abord il reste vrai dans lorsque V est continue (ce qui est correct sous des conditions standards sur les données), à condition de recourir à une notion de solution généralisée pour l'équation de HJ (la notion de solution de viscosité, cf. la monographie de Barles sur le sujet). Ensuite il montre le rôle majeur joué par l'équation de HJ dans le problème, rôle confirmé par le théorème de vérification énoncé ci-dessous.

Preuve. Soit $u_0 \in U$ (que l'on regarde comme un contrôle constant) et considérons la solution (x, u_0) de l'EDO avec donnée initiale x_0 en temps t_0. On a alors par programmation dynamique, avec $t_1 = t_0 + h$ (où $h > 0$ est petit)

$$\begin{aligned} V(t_0,x_0) &\leq \int_{t_0}^{t_0+h} L(t,x(t),u_0)\ dt + V(t_0+h, x(t_0+h)) \\ &= V(t_0,x_0) + \int_{t_0}^{t_0+h} L(t,x(t),u_0) + \frac{\partial V}{\partial t}(t,x(t)) + \langle \frac{\partial V}{\partial x}(t,x(t)), \dot{x}(t)\rangle\ dt \\ &= V(t_0,x_0) + \int_{t_0}^{t_0+h} L(t,x(t),u_0) + \frac{\partial V}{\partial t}(t,x(t)) + \langle \frac{\partial V}{\partial x}(t,x(t)), f(t,x(t),u_0)\rangle\ dt \\ &= V(t_0,x_0) + h(L(t_0,x_0,u_0) + \frac{\partial V}{\partial t}(t_0,x_0) + \langle \frac{\partial V}{\partial x}(t_0,x_0), f(t_0,x_0,u_0)\rangle) + o(h) \end{aligned}$$

où la dernière égalité vient de la continuité des fonctions. On simplifie à gauche et à droite de l'expression par $V(t_0,x_0)$, on divise par h et on fait tendre $h \to 0$: cela donne

$$0 \leq L(t_0,x_0,u_0) + \frac{\partial V}{\partial t}(t_0,x_0) + \langle \frac{\partial V}{\partial x}(t_0,x_0), f(t_0,x_0,u_0)\rangle,$$

ce qui se réécrit

$$-\frac{\partial V}{\partial t}(t_0,x_0)+\left(-\langle\frac{\partial V}{\partial x}(t_0,x_0),f(t_0,x_0,u_0)\rangle-L(t_0,x_0,u_0)\right)\leq 0$$

et donc, comme u_0 est arbitraire,

$$\begin{aligned}&-\frac{\partial V}{\partial t}(t_0,x_0)+H(t,x,\frac{\partial V}{\partial x}(t_0,x_0))\\&=-\frac{\partial V}{\partial t}(t_0,x_0)+\sup_{u\in U}\left(-\langle\frac{\partial V}{\partial x}(t_0,x_0),f(t_0,x_0,u)\rangle-L(t_0,x_0,u)\right)\leq 0.\end{aligned}$$

L'inégalité inverse se montre par l'absurde. Supposons le résultat faux, c'est-à-dire qu'il existe $\varepsilon>0$ et (t_0,x_0) tels que

$$-\frac{\partial V}{\partial t}(t_0,x_0)+H(t_0,x_0,\frac{\partial V}{\partial x}(t_0,x_0))\ <\ -2\varepsilon.$$

Alors, par continuité, l'inégalité reste vraie pour (t,x) appartenant à un voisinage $B_\eta(t_0,x_0)$ (où $\eta>0$) de (t_0,x_0) :

$$-\frac{\partial V}{\partial t}(t,x)+\sup_{u\in U}\left\{-\langle\frac{\partial V}{\partial x}(t,x),f(t,x,u)\rangle-L(t,x,u)\right\}\ <\ -\varepsilon\qquad\forall(t,x)\in B_\eta(t_0,x_0),$$

ce qui se réécrit

$$\frac{\partial V}{\partial t}(t,x)+\langle\frac{\partial V}{\partial x}(t,x),f(t,x,u)\rangle+L(t,x,u)\ >\ \varepsilon\qquad\forall(t,x)\in B_\eta(t_0,x_0),\ \forall u\in U.\tag{3.3}$$

Soit $h>0$ petit, (x_h,u_h) εh^2-optimal pour la programmation dynamique sur l'intervalle $[t_0,t_0+h]$:

$$V(t_0,x_0)+\varepsilon h^2\geq\int_{t_0}^{t_0+h}L(t,x_h(t),u_h(t))\ dt+V(t_0+h,x_h(t_0+h)).$$

Alors

$$\begin{aligned}&V(t_0,x_0)-\varepsilon h^2\\&\quad\geq V(t_0,x_0)+\int_{t_0}^{t_0+h}L(t,x_h(t),u_h(t))+\frac{\partial V}{\partial t}(t,x_h(t))+\langle\frac{\partial V}{\partial x}(t,x_h(t)),f(t,x_h(t),u_h(t))\rangle\ dt\end{aligned}$$

Lorsque h est assez petit, le couple $(t,x_h(t))$ reste dans $B_\eta(t_0,x_0)$ pour $t\in[t_0,t_0+h]$. Donc, par (3.3),

$$V(t_0,x_0)-\varepsilon h^2\ \geq\ V(t_0,x_0)+\int_{t_0}^{t_0+h}\varepsilon\ dt\ =\ V(t_0,x_0)+\varepsilon h,$$

ce qui est impossible pour $h>0$ petit. On a donc une contradiction, ce qui conclut la preuve du théorème. □

Comme dans le cas discret, un des intérêts de l'équation de Hamilton-Jacobi est de fournir un "feedback optimal". En temps continu, un feedback est une application continue $\tilde{u} : [0,T]\times\mathbb{R}^N \to U$. A partir d'un feedback, on peut construire des contrôles de la façon suivante : si $\tilde{u}$ est un feedback et (t_0, x_0) est une donnée initiale, on résout l'EDO

$$\begin{cases} \dot{x}(t) = f(t, x(t), \tilde{u}(t, x(t))), & t \in [0,T] \\ x(0) = x_0 \end{cases}$$

(dont la solution existe d'après le théorème de Cauchy-Péano), puis on pose $u(t) = \tilde{u}(t, x(t))$. Le "feedback" est optimal si le contrôle $(u(t))$ est optimal pour le problème. Le résultat suivant est un exemple de théorème de vérification, qui affirme, sous des hypothèses de régularité assez fortes, que l'équation de Hamilton-Jacobi possède une seule solution—qui est la fonction valeur—et fournit également un feedback optimal.

Théorème 3.6 (de vérification)**.** *Supposons que*

1. *$W : [0,T]\times\mathbb{R}^N \to \mathbb{R}$ est de classe C^1 sur $]0,T[\times\mathbb{R}^N$ et C^0 sur $[0,T]\times\mathbb{R}^N$,*
2. *W satisfait l'équation de Hamilton-Jacobi*

$$\begin{cases} -\dfrac{\partial W}{\partial t}(t,x) + H(t,x,\dfrac{\partial W}{\partial x}(t,x)) = 0 & \forall (t,x) \in (0,T)\times\mathbb{R}^N \\ W(T,x) = g(x) & \forall x \in \mathbb{R}^N \end{cases}$$

3. *il existe une application continue $\tilde{u}^* : (0,T)\times\mathbb{R}^N \to U$ telle que, pour tout $(t,x) \in]0,T[\times\mathbb{R}^N$,*

$$-\langle \frac{\partial W}{\partial x}(t,x), f(t,x,\tilde{u}^*(t,x))\rangle - L(t,x,\tilde{u}^*(t,x)) = H(t,x,\frac{\partial W}{\partial x}(t,x)).$$

Alors $W = V$ et un feedback optimal est donné par $\tilde{u}^*$.

Autant le résultat est lourd à énoncer, autant la démonstration est simple et directe. C'est donc plus la démonstration qu'il faut retenir que le théorème lui-même.

Preuve. Fixons un contrôle arbitraire $u : [0,T] \to U$ et notons $x(\cdot)$ la solution de l'EDO associée. On a

$$\begin{aligned} \frac{d}{dt}W(t,x(t)) &= \frac{\partial W}{\partial t}(t,x(t)) + \langle \frac{\partial W}{\partial x}(t,x(t)), \dot{x}(t)\rangle \\ &= \frac{\partial W}{\partial t}(t,x(t)) + \langle \frac{\partial W}{\partial x}(t,x(t)), f(t,x(t),u(t))\rangle. \end{aligned}$$

Or, d'après l'équation de HJ satisfaite par W, on a, pour tout $(x,u) \in \mathbb{R}^N \times U$,

$$\begin{aligned} 0 &= -\frac{\partial W}{\partial t}(t,x) + \sup_{u\in U}\left\{ -\langle \frac{\partial W}{\partial x}(t,x), f(t,x,u)\rangle - L(t,x,u)\right\} \\ &\geq -\frac{\partial W}{\partial t}(t,x) - \langle \frac{\partial W}{\partial x}(t,x), f(t,x,u)\rangle - L(t,x,u) \end{aligned}$$

Donc

$$\begin{aligned} \frac{d}{dt}W(t,x(t)) &= \frac{\partial W}{\partial t}(t,x(t)) + \langle \frac{\partial W}{\partial x}(t,x(t)), f(t,x(t),u(t))\rangle \\ &\geq -L(t,x(t),u(t)) \end{aligned}$$

Intégrons l'inégalité ci-dessus entre t_0 et T, en tenant compte du fait que $x(t_0) = x_0$ et que $W(T, \cdot) = g$:

$$g(x(T)) - W(t_0, x_0) = \int_{t_0}^{T} \frac{d}{dt} W(t, x(t))\ dt \ \geq\ -\int_{t_0}^{T} L(t, x(t), u(t))\ dt\ ,$$

ce qui donne

$$W(t_0, x_0) \leq \int_{t_0}^{T} L(t, x(t), u(t))\ dt + g(x(T)).$$

Comme ceci est vrai pour tout contrôle u, on en déduit une première inégalité :

$$W(t_0, x_0) \leq V(t_0, x_0).$$

Pour obtenir l'inégalité inverse (et l'optimalité de $\tilde{u}^*$), considérons x^* une solution de l'EDO

$$\begin{cases} \dot{x^*}(t) = f(t, x^*(t), \tilde{u}^*(t, x^*(t))), & t \in [t_0, T] \\ x^*(t_0) = x_0 \end{cases}$$

Reprenons maintenant la dérivée de W le long de cette solution particulière :

$$\begin{aligned} \frac{d}{dt} W(t, x^*(t)) &= \frac{\partial W}{\partial t}(t, x^*(t)) + \langle \frac{\partial W}{\partial x}(t, x^*(t)), \dot{x^*}(t) \rangle \\ &= \frac{\partial W}{\partial t}(t, x^*(t)) + \langle \frac{\partial W}{\partial x}(t, x^*(t)), f(t, x^*(t), \tilde{u}^*(t, x^*(t))) \rangle \\ &= -L(t, x^*(t), \tilde{u}^*(t, x(t))) \end{aligned}$$

où la dernière égalité vient de l'équation de HJ satisfaite par W et de la définition de $\tilde{u}^*$. Intégrons l'inégalité ci-dessus entre t_0 et T, en utilisant le fait que $x(t_0) = x_0$ et que $W(T, \cdot) = g$:

$$g(x^*(T)) - W(t_0, x_0) = \int_{t_0}^{T} \frac{d}{dt} W(t, x^*(t))\ dt \ =\ -\int_{t_0}^{T} L(t, x^*(t), \tilde{u}^*(t, x^*(t)))\ dt\ ,$$

ce qui donne

$$W(t_0, x_0) = \int_{t_0}^{T} L(t, x^*(t), \tilde{u}^*(t, x^*(t)))\ dt + g(x^*(T)) \ \geq\ V(t_0, x_0),$$

puisque $\tilde{u}^*(\cdot, x^*(\cdot))$ est un contrôle particulier. Comme on avait déjà prouvé que $W(t_0, x_0) \leq V(t_0, x_0)$, il y a égalité dans l'inégalité ci-dessus, et $\tilde{u}^*$ est optimal pour le problème. □

Partie IV

Eléments de bibliographie

Pré-requis : ce cours fait suite au cours d'optimisation de L3, au cours de systèmes différentiels pour la partie contrôle optimal ainsi qu'au cours d'analyse fonctionnelle de M1. Pour les notes de cours, voir :

— Carlier G., Calcul différentiel et optimisation, Dauphine, disponible en ligne, 2008-2009.
— Glass O. Analyse fonctionnelle et équations aux dérivées partielles, 2014-2015.
— Vialard F.-X., Optimisation Numérique, Dauphine, disponible en ligne, 2013.

Pour aller plus loin :

— Barles G., "Solutions de viscosité des équations de Hamilton-Jacobi", Springer-Verlag (1994).
— Bertsekas D. M. "Dynamic Programming and Stochastic Control", cours en ligne du MIT, 2011
— Carlier G. "Programmation dynamique", notes de cours de l'ENSAE, 2007.
— Ciarlet P.G., "Introduction à l'analyse matricielle et à l'optimisation", Collection Mathématiques appliquées pour la maîtrise, Masson.
— Cesari L. (1983). "Optimization—theory and applications". Springer.
— Culioli J.-C., "Introduction à l'optimisation", Ellipse, 1994.
— Fleming W. H. et Rishel R. W., "Deterministic and Stochastic Optimal Control", Springer, 1975
— Lucas Jr R. E. et Stokey N., "Recursive Methods in Economics Dynamics", Harvard University Press (1989).
— Hiriart-Urruty J.-B., "Optimisation et analyse convexe", Mathématiques, PUF, 1998.
— Hiriart-Urruty J.-B., "L'optimisation", Que sais-je ?, PUF, 1996.
— Minoux M., "Programmation mathématique, théorie et algorithmes (tomes 1 et 2)", Dunod, 1983.
— Trélat E. "Contrôle optimal : théorie et applications", polycopié en ligne de Paris 6, 2013.

Quelques "Toolboxes". Impossible de rendre compte de tous les programmes accessibles. Voici deux exemples :

— Le logiciel libre SciPy (ensemble de bibliothèques Python) possède un package dédié à l'optimisation : scipy.optimize. Quelques fonctionnalités :
  — Optimisation sans contrainte,
  — Moindres carrés non-linéaires
  — Optimisation dans des boites
— Matlab possède une toolbox pour résoudre entre autres
  — des problèmes d'optimisation non linéaire avec ou sans contrainte
  — des problèmes de programmation linéaire et quadratique
  — des problèmes de moindres carrés non linéaires, ajustement de données et équations non linéaires