Université Paris Dauphine

Optimisation et programmation dynamique

Master mention Mathématiques appliquées 1ère année

2015-2016

Pierre Cardaliaguet

Introduction

L'objet de ce cours est de présenter quelques notions sur deux types de problèmes d'optimisation : l'optimisation avec 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 recourent à des concepts similaires (conditions d'optimalité, fonction valeur, etc...).

Un mot d'avertissement : comme souvent en mathématiques, nous présentons ci-après un cadre d'étude de ces problèmes, en aucun cas des "recettes" pour les "résoudre". Trois raisons à cela :

- D'abord parce que l'expression "résoudre" est ambigüe. Dans la plupart des cas pratiques, la solution explicite est inaccessible, et il faut recourir à des méthodes numériques—celles-ci s'appuyant fortement sur l'analyse mathématique du problème (existence de solution, conditions nécessaires d'optimalité, dualité, programmation dynamique,...).
- Bien garder en tête qu'assez fréquemment les problèmes rencontrés en pratique sortent du cadre des hypothèses de l'analyse du cours. C'est notamment le cas en contrôle optimal, où je ne connais pas un exemple utilisé en économie qui entre parfaitement dans le cadre étudié : il faut alors comprendre au cas par cas ce qui reste correct et ce qui ne l'est plus, en adaptant les techniques développées dans ce cours au problème étudié.
- Enfin, les problèmes rencontrés en pratique ont très souvent une structure spéciale, qu'il faut bien sûr utiliser pour gagner en efficacité. Le cours se contente de traiter de situations assez générales, et fait l'impasse sur plusieurs cas particuliers importants. En optimisation, nous ne mentionnerons que rapidement l'algorithme du simplexe (dans le cas où les contraintes et le critère sont affines) qui est de loin l'algorithme le plus utilisé mais qui est traité dans d'autres UE. En contrôle optimal, on rencontre fréquemment des problèmes avec dynamiques linéaires et coûts quadratiques, dont la résolution fait l'objet de techniques algébriques efficaces, mais qui sort du cadre du cours.

Pré-requis : le cours utilise largement (et 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 figurent en appendice.

Certaines parties du polycopié sont reprises d'années antérieures et ne figurent pas au programme du cours 2014-2015. *Sont hors programme* la démonstration du théorème de Kuhn & Tucker (section 1.3) ainsi que les conditions du second ordre (section 1.6) ; la partie sur l'optimisation de portefeuille financier (section 1.7) est également hors programme, mais pourra être traitée en exercice.

Table des matières

Chapitre 1

Optimisation sous contraintes

1.1 Existence d'un minimum

Dans cette partie, très brève, on rappelle quelques définitions élémentaires, et on énonce des conditions suffisantes d'existence d'un minimum.

1.1.1 Vocabulaire

Infimum et minimum

Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction et K un sous-ensemble non vide de $\mathbb{R}^n$.

Définition 1.1.1. *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 :

- 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

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

- Si f n'est pas minorée, alors l'infimum de f est $-\infty$.
- 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 1.1.2. *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 une solution $\bar{x}$.

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

Fonctions coercives

Définition 1.1.3. *Une fonction $f : \mathbb{R}^n \to \mathbb{R}$ est coercive si*

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

Remarque : Peu importe la norme $\|\cdot\|$ que l'on utilise puisque, sur $\mathbb{R}^n$, toutes les normes sont équivalentes. En pratique, on choisit la norme la plus adaptée à la fonction f étudiée.

Exemples à connaitre :

1. Soit A une matrice symétrique, carrée, d'ordre N, b un vecteur de $\mathbb{R}^n$, et c un réel. Alors la fonction $f : \mathbb{R}^n \to \mathbb{R}$ définie par
$$f(x) = x^T Ax + b^T x + c$$
est coercive, si et seulement si, A est une matrice définie positive. Rappelons que toute matrice symétrique est diagonalisable. Une matrice est positive, si et seulement si, toutes ses valeurs propres sont positives. Elle est définie positive, si et seulement si, toutes ses valeurs propres sont strictement positives. Si A est symétrique, on a les inégalités suivantes :
$$\forall x \in \mathbb{R}^n, \ \ \lambda_{min}\|x\|^2 \leq \langle Ax, x\rangle \leq \lambda_{max}\|x\|^2$$
où λ_{min} et λ_{max} sont repectivement la plus petite et la plus grande valeur propre de A. En particlier, si A est définie positive, alors $\lambda_{min} > 0$.
2. Toute fonction minorée par une fonction coercive est coercive.

Proposition 1.1.1. *On suppose que la fonction $f : \mathbb{R}^n \to \mathbb{R}$ est 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 $m = \max_i |m_i|$. Soit M fixé. Il existe une constante $R_i > 0$ telle que

$$\forall |x_i| \geq R_i, \ f_i(x_i) \geq M + nm \ .$$

Posons $R = \max_i R_i$. Alors pour tout $x \in \mathbb{R}^n$ avec $\|x\|_\infty \geq R$, il existe $i \in \{1, \ldots, n\}$ tel que $|x_i| \geq R \geq R_i$. Donc $f_i(x_i) \geq M + nm$. Comme

$$\forall j \in \{1, \ldots, n\}, \ f_j(x_j) \geq m_j \geq -m$$

on en conclut que

$$f(x) \geq M + \sum_{j\neq i}(m_j + m) \geq M$$

Donc on a 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,

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

Par conséquent la fonction f est coercive. □

1.1.2 Condition d'existence d'un minimum sous contraintes

Voici le résultat le plus important du chapitre :

Théorème 1.1.1. *Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction continue et K un sous-ensemble non vide de $\mathbb{R}^n$. Le problème*

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

a une solution si **l'une des deux** *conditions suivantes est satisfaite*

1. *la contrainte K est compacte (théorème de Weierstrass)*
2. *la fonction f est coercive et la contrainte K est fermée,*

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

Preuve. Le premier cas étant très classique, on ne montre que le second. Supposons que f soit coercive et K fermé. Notons $m = \inf_{x\in K} f(x)$ et K_1 l'ensemble

$$K_1 = \{x \in K \mid f(x) \leq m + 1\} \text{ si } m \in \mathbb{R} \text{ et } K_1 = \{x \in K \mid f(x) \leq 0\} \text{ sinon .}$$

Comme f est coercive, l'ensemble K_1 est compact. De plus, K_1 est non vide par définition de l'infimum. Donc, d'après la première partie du théorème, f atteint son minimum sur K_1 en un point $\bar{x} \in K_1$. Montrons que $\bar{x}$ est un minimum de f sur K. En effet, pour tout $x \in K$,

i) soit x appartient à K_1, et alors $f(\bar{x}) \leq f(x)$ car f atteint son minimum sur K_1 en $\bar{x}$,

ii) soit $x \notin K_1$, auquel cas on a : si $m \in \mathbb{R}$, $f(\bar{x}) \leq m+1 < f(x)$ (car $\bar{x} \in K_1$) et si $m = -\infty$, $f(\bar{x}) \leq 0 < f(x)$.

Ceci montre donc que $\bar{x}$ est un minimum de f sur K. □

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

Proposition 1.1.2. *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. En effet, comme l'ensemble $\overline{K}$ est compact, la fonction continue f admet un minimum sur $\overline{K}$, c'est-à-dire qu'il existe un élément $\bar{x}$ de $\overline{K}$ tel que

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

Montrons par l'absurde que $\bar{x}$ appartient en fait à K. En effet, sinon, $\bar{x}$ appartient au bord de K, car K est un ouvert. Donc $f(x_0) < f(\bar{x})$ par hypothèse. Mais cette inégalité est en contradiction avec le fait que $\bar{x}$ est un minimum de f sur $\overline{K}$. Donc $\bar{x}$ appartient à K. Comme $\bar{x}$ est un minimum de f sur $\overline{K}$, que $K \subset \overline{K}$ et que $\bar{x}$ appartient à K, on en déduit que $\bar{x}$ est aussi un minimum de f sur K. □

1.2 Conditions nécessaires d'optimalité

Soit K un sous-ensemble de $\mathbb{R}^n$ et f une application de $\mathbb{R}^n$ dans $\mathbb{R}$. On cherche le ou les minima du problème $(\mathcal{P})$

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

Dans ce chapitre, nous cherchons *des conditions nécessaires d'optimalité*, c'est-à-dire des conditions, portant sur la dérivée de f, satisfaites par le ou les minima du problème. Ces conditions portent le nom de conditions de Kuhn & Tucker, ou de Karush, Kuhn & Tucker, ou encore KKT.

1.2.1 Condition nécessaire d'optimalité dans un ouvert

On suppose ici que K est un ouvert de $\mathbb{R}^n$ et que f une application de $\mathbb{R}^n$ dans $\mathbb{R}$ de classe $\mathcal{C}^1$.

Théorème 1.2.1 (Condition d'Euler)**.** *Si K est un ouvert de $\mathbb{R}^n$ et que f une application de $\mathbb{R}^n$ dans $\mathbb{R}$ de classe $\mathcal{C}^1$, et si x^* est un minimum du problème $\mathcal{P}$, alors*

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

Remarques :

1. Rappelons qu'il existe une condition du second ordre, pour les applications de classe $\mathcal{C}^2$: la matrice symétrique $Hess_f(x^*)$ est une matrice positive, c'est à dire que ses valeurs propres sont toutes positives ou nulles.

2. La preuve de ce résultat étant le prototype des preuves en optimisation, il faut absolument la connaître.

Preuve. Soit v 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 $x^* + hv$ appartient à K. Or x^* étant un minimum du problème, on a

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

Comme

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

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

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

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

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

c'est-à-dire que $\nabla f(x^*) = 0$. □

1.2.2 Le théorème de Kuhn & Tucker

Le théorème de Kuhn-Tucker avec lagrangien généralisé

Dans le cadre général du théorème de Kuhn & Tucker, la contrainte K est de la forme

$$K = \{x \in \mathbb{R}^n,\ g_i(x) \leq 0,\ i \in I,\ h_j(x) = 0,\ j \in J\}$$

où $I = \{1, \ldots, l\}$ indexe les contraintes d'inégalité et $J = \{1, \ldots, m\}$ indexe les contraintes d'égalité. Les fonctions g_i et h_i sont toutes supposées de classe $\mathcal{C}^1$ de $\mathbb{R}^n$ dans $\mathbb{R}$. Pour tout $x \in K$, on appelle contraintes saturées les indices $i \in \{1, \ldots, l\}$ tels que $g_i(x) = 0$:

$$I(x) = \{i \in \{1, \ldots, l\} \mid g_i(x) = 0\}$$

Théorème 1.2.2. *Si un point x^* est un minimum du problème $(\mathcal{P})$, alors il existe $p_0 \in \mathbb{R}_+$, $p \in \mathbb{R}^l_+$ et $q \in \mathbb{R}^m$ avec*

$$\begin{cases} i) & \sum_i p_i g_i(x^*) = 0 & \text{(condition d'exclusion)} \\ ii) & (p_0, p, q) \neq 0 & \\ iii) & p_0 \nabla f(x^*) + \sum_i p_i \nabla g_i(x^*) + \sum_j q_j \nabla h_j(x^*) = 0 & \text{(condition nécessaire)} \end{cases}$$

Remarque :

1. Le vecteur (p_0, p, q) est appelé **le multiplicateur généralisé** associé à la solution x^* ("généralisé", car, comme nous le verrons plus loin, on peut en général prendre $p_0 = 1$).
2. La condition d'exclusion signifie que, si $i \notin I(x^*)$, alors $p_i = 0$.

3. Il est tout à fait possible que $p_0 = 0$ dans l'expression précédente. Ce cas est cependant assez "pathologique", au sens où il correspond à une contrainte peu "régulière". Le "vrai" théorème de Kuhn & Tucker affirme que, si la contrainte est "qualifiée", on peut prendre $p_0 = 1$ (c.f. théorème 1.2.3 ci-dessous).
4. La partie importante du théorème est, bien sur, la condition nécessaire. Nous verrons plus loin une interprétation géométrique de cette condition.
5. On appelle **Lagrangien généralisé** la fonction

$$L(x, p_0, p, q) = p_0 f(x) + \sum_i p_i g_i(x) + \sum_j q_j h_j(x)$$

La condition nécessaire d'optimalité s'écrit aussi

$$\frac{\partial L}{\partial x}(x^*, p_0, p, q) = 0 \ .$$

Le théorème de Kuhn & Tucker pour les contraintes qualifiées

Définition 1.2.1 (Qualification)**.** *On dit que la contrainte* **non linéaire** K *est qualifiée en un point* $x^* \in K$ *si, pour tout* $\lambda \in \mathbb{R}^l_+$ *et* $\mu \in \mathbb{R}^m$ *avec*

$$\begin{cases} a) & \sum_i \lambda_i g_i(x^*) = 0 \qquad \text{(condition d'exclusion)} \\ b) & \sum_i \lambda_i \nabla g_i(x^*) + \sum_j \mu_j \nabla h_j(x^*) = 0 \end{cases}$$

on a nécessairement $\lambda = 0$ *et* $\mu = 0$.

Remarque : La notion de qualification n'est pas du tout géométrique. Deux ensembles de contraintes peuvent définir le même ensemble, l'un étant qualifié, l'autre non. La définition précédente n'est valable que pour les contraintes non linéaires.

Si la contrainte K est qualifiée, le théorème de Kuhn & Tucker peut se reformuler de la façon suivante :

Théorème 1.2.3. *Soit* K *la contrainte fermée définie par*

$$K = \{x \in R^n,\ g_i(x) \le 0,\ i \in I,\ h_j(x) = 0,\ j \in J\}$$

Si un point x^* *est un minimum du problème* $\mathcal{P}$ *et si* K *est qualifiée en* x^**, alors il existe* $\lambda \in \mathbb{R}^l_+$ *et* $\mu \in \mathbb{R}^m$ *avec*

$$\begin{cases} i) & \sum_i \lambda_i g(x^*) = 0 \qquad \text{(condition d'exclusion)} \\ ii) & \nabla f(x^*) + \sum_i \lambda_i \nabla g_i(x^*) + \sum_j \mu_j \nabla h_j(x^*) = 0 \quad \text{(condition nécessaire)} \end{cases}$$

En d'autres termes, on peut prendre $p_0 = 1$ dans le théorème 1.2.2. *Nous admettrons sans démonstration pour l'instant* que le résultat reste valable sans hypothèse de qualification pour des contraintes affines, c'est-à-dire lorsque g_i et h_i sont des fonctions affines.

Preuve. Soient p_0, p et q donnés par le théorème 1.2.2. Montrons d'abord que $p_0 \neq 0$. En effet, sinon, les conditions (i) et (iii) du théorème 1.2.2 s'écrivent

$$\sum_i p_i g_i(x^*) = 0 \text{ et } \sum_i p_i \nabla g_i(x^*) + \sum_j q_j \nabla h_j(x^*) = 0 \ .$$

Or la contrainte est qualifiée en x^*, donc $p = 0$ et $q = 0$. Mais alors $(p_0, p, q) = 0$, ce qui est en contradiction avec la condition (ii) du théorème. Par conséquent, p_0 est non nul et, plus précisément $p_0 > 0$. Posons maintenant $\lambda = p/p_0$ et $\mu = q/p_0$. Comme $p_0 > 0$, $\lambda \in \mathbb{R}^l_+$ et la condition (i) du théorème 1.2.2 s'écrit alors $\sum_i \lambda_i g_i(x^*) = 0$, tandis que la condition nécessaire (iii) donne

$$\nabla f(x^*) + \sum_i \lambda_i \nabla g_i(x^*) + \sum_j \mu_j \nabla h_j(x^*) = 0$$

Donc le théorème est démontré. □

La définition de qualification n'a de sens que si l'on peut donner des critères de qualification. C'est l'objet de la proposition suivante :

Proposition 1.2.1. *On suppose que les deux conditions suivantes sont satisfaites en $x \in K$:*

1. *la famille $\{\nabla h_1(x), \ldots, \nabla h_m(x)\}$ est libre,*
2. *il existe un vecteur $v \in \mathbb{R}^n$ tel que*

$$\forall j \in \{1, \ldots, m\} \ , \ \langle \nabla h_j(x), v \rangle = 0$$

et

$$\forall i \in I(x) \ , \ \langle \nabla g_i(x), v \rangle \ < \ 0 \ .$$

Alors la contrainte K est qualifiée en x.

Remarques :

- On verra au chapitre suivant que cette condition s'écrit de façon plus sympathique lorsque la contrainte est convexe.
- Le vecteur (p, q) est appelé **le multiplicateur de Lagrange du problème** associé à la solution x^*.
- On appelle **Lagrangien du problème** la fonction

$$L(x, \lambda, \mu) = f(x) + \sum_i \lambda_i g_i(x) + \sum_j \mu_j h_j(x)$$

La condition nécessaire d'optimalité s'écrit aussi, sous les hypothèses du théorème 1.2.3

$$\frac{\partial L}{\partial x}(x^*, \lambda, \mu) = 0 \ .$$

Preuve. Soient λ et μ comme dans la définition de la qualification. Montrons que $\lambda = 0$ et $\mu = 0$. On sait déjà que $\lambda_i = 0$ si $i \notin I(x)$, d'après (i). Soit v comme dans la proposition. Comme

$$\sum_{i \in I(x)} \lambda_i \nabla g_i(x^*) + \sum_j \mu_j \nabla h_j(x^*) = 0$$

on en déduit que

$$\sum_{i \in I(x)} \lambda_i \langle \nabla g_i(x^*), v \rangle + \sum_j \mu_j \langle \nabla h_j(x^*), v \rangle = 0$$

Or $\langle \nabla h_j(x^*), v \rangle = 0$, et $\langle \nabla g_i(x^*), v \rangle < 0$ pour tout $i \in I(x)$. Comme, pour tout i, $\lambda_i \geq 0$, l'égalité

$$\sum_{i \in I(x)} \lambda_i \langle \nabla g_i(x^*), v \rangle = 0$$

implique que, pour tout i, $\lambda_i = 0$. Alors

$$\sum_j \mu_j \nabla h_j(x^*) = 0$$

Comme la famille $\{\nabla h_1(x), \ldots, \nabla h_m(x)\}$ est libre, on doit avoir $\mu_j = 0$ pour tout j. Donc finalement, λ et μ sont nuls. La contrainte est donc qualifiée. □

Exemples

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

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

La contrainte $K = \{(x, y) \mid x^2 + y^2 = 1\}$ est compacte, et la fonction $f(x, y) = 2x + y$ est continue, donc le problème admet une solution (x^*, y^*). Posons $h(x, y) = x^2 + y^2 - 1$. Montrons que la contrainte est qualifiée

en tout point. Remarquons d'abord qu'il y a une seule contrainte d'égalité ($J = \{1\}$) et qu'il n'y a pas de contrainte d'inégalité ($I = \emptyset$). Soit (x, y) un point de K. Il suffit de montrer que $\nabla h(x, y) \neq 0$. Or

$$\nabla h(x, y) = \begin{pmatrix} 2x \\ 2y \end{pmatrix}$$

qui n'est nul que si $(x, y) = (0, 0)$. Comme $(0, 0)$ ne vérifie pas la contrainte, on peut conclure que celle-ci est qualifiée en tout point. Le théorème de Kuhn & Tucker affirme alors qu'il existe $\mu \in \mathbb{R}$ (pas de contrainte d'inégalité et une contrainte d'égalité) tel que

$$\nabla f(x^*, y^*) + \mu \nabla h(x^*, y^*) = 0$$

Il faut donc résoudre le système :

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

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

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

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 minimum : c'est le point $(-2\sqrt{5}/5, -\sqrt{5}/5))$.

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

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

Le critère est $f(x, y) = xy$, il n'y a pas de contrainte d'égalité ($J = \emptyset$), et il y a une contrainte d'égalité $g(x, y) = x^2 + y^2 - 1$. On remarque qu'il y a bien un minimum, car la contrainte est compacte. Montrons que la contrainte est qualifiée en tout point. Soit (x, y) un point de K. Si $g(x, y) < 0$, il n'y a rien à vérifier. Sinon, on doit trouver un vecteur $v \in \mathbb{R}^2$ tel que

$$\langle \nabla g(x, y), v \rangle \ < \ 0 \ .$$

Remarquons pour cela qu'il suffit que $\nabla g(x, y)$ soit non nul car alors on peut prendre $v = -\nabla g(x, y)$, ce qui donne

$$\langle \nabla g(x, y), v \rangle = -\|\nabla g(x, y)\|^2 \ .$$

Mais nous avons déjà démontré, dans l'exercice précédent, que si $g(x, y) = 0$ alors $\nabla g(x, y) \neq 0$ si $g(x, y) = x^2 + y^2 - 1$. Ceci permet d'affirmer que la contrainte est qualifiée. Le théorème de Kuhn & Tucker affirme alors que, si (x, y) est un minimum du problème, il existe $\lambda \geq 0$ tel que

$$\nabla f(x, y) + \lambda g(x, y) = 0$$

c'est-à-dire que

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

On montre (ce n'est pas très agréable) que ce système a cinq solutions possibles : $(\lambda = 0,\ x = y = 0)$, $(\lambda = 1,\ x = -y = \sqrt{2}/2)$, $(\lambda = 1,\ x = -y = -\sqrt{2}/2)$, $(\lambda = -1,\ x = y = \sqrt{2}/2)$, $(\lambda = -1,\ x = y = -\sqrt{2}/2)$. 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 :

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

Donc il y a deux solutions : $(\sqrt{2}/2, -\sqrt{2}/2)$ et $(-\sqrt{2}/2, \sqrt{2}/2)$.

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

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

La contrainte $K=\{(x,y,z)\in\mathbb{R}^3 \mid x^2+y^2=1,\ y^2+z^2=4\}$ est compacte, car fermée et contenue dans $[-1,1]\times[-1,1]\times[-2,2]$. La fonction $f(x,y)=x+z$ est continue, donc le problème admet une solution (x^*,y^*,z^*). Posons $h_1(x,y,z)=x^2+y^2-1$ et $h_2(x,y,z)=y^2+z^2-4$. Cherchons les points où la contrainte est qualifiée. Ce sont les points $(x,y,z)\in K$ tels que 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 ne forment pas un système libre, si et seulement si, $x=z=0$. Or ce cas est impossible car il imposerait $y^2=1$ et $y^2=4$. Donc le système de vecteurs est toujours libre, et la contrainte est qualifiée. Soit maintenant $(x^*,y^*,z^*)\in K$ un minimum du problème. Les conditions nécessaires d'optimalité s'écrivent : il existe μ_1 et μ_2 deux réels tels que

$$\begin{cases} 1+2\mu_1x^*=0 \\ 2\mu_1y^*+2\mu_2y^*=0 \\ 1+2\mu_2z^*=0 \\ (x^*)^2+(y^*)^2=1 \\ (y^*)^2+(z^*)^2=4 \end{cases}$$

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

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

et on aboutit à une contradiction. Donc $y^*=0$, ce qui impose $x^*=\pm1$ et $z^*=\pm2$. On voit alors facilement que le minimum du problème est $(-1,0,-2)$.

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

$$\begin{array}{c} \min \quad x+y+z \\ x^2+y^2+z^2\le 1 \\ x\ge 0 \end{array}$$

La contrainte est compacte et la fonction $f(x,y,z)=x+y+z$ continue, donc le problème admet une solution (x^*,y^*,z^*). La contrainte est constituée de deux inégalités $g_1(x,y,z)=x^2+y^2+z^2-1\le 0$ et $g_2(x,y,z)=-x\le 0$. Montrons que la contrainte est qualifiée. 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 $\lambda_1\ge 0$ et $\lambda_2\ge 0$ tels que

$$\begin{cases} 1+2\lambda_1x^*-\lambda_2=0 \\ 1+2\lambda_1y^*=0 \\ 1+2\lambda_1z^*=0 \end{cases}$$

avec la condition d'exclusion

$$\lambda_1((x^*)^2+(y^*)^2+(z^*)^2-1)+\lambda_2(-x^*)=0$$

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 $\lambda_1=\sqrt{2}$. D'où $x^*=0$ et $y^*=z^*=\sqrt{2}/4$. 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 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 $g(x,y,z)=x+y+z\leq 1$. Vérifions que la contrainte est qualifiée. La condition (1) de la Proposition 1.2.1 est satisfaite car $\nabla h(x,y,z)=0$ si et seulement si $x=y=z=0$, ce qui est impossible car $h(x,y,z)=1$. 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

$$\left\{\begin{array}{l} 1+\lambda+2\mu x^*=0 \\ 2+\lambda+2\mu y^*=0 \\ 3+\lambda+2\mu z^*=0 \\ x^2+y^2+z^2=1 \\ \lambda(x^*+y^*+z^*)=0 \end{array}\right.$$

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

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

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

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

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 $(-\sqrt{14}/14,-\sqrt{14}/7,-3\sqrt{14}/14)$.

Preuve du théorème de Kuhn & Tucker par pénalisation

Cette preuve se fait en plusieurs étapes.

1. On considère la pénalisation suivante

$$f_N(x)=f(x)+\|x-x^*\|^2+\frac{N}{2}\left[\sum_{i\in I}\max(0,g_i(x))^2+\sum_{j\in J}h_j(x)^2\right]$$

Remarquons que f_N est une fonction de classe $\mathcal{C}^1$, et que $f_N(x^*)=f(x^*)$.

2. Nous montrons maintenant qu'il existe $\epsilon_0>0$ tel que, pour tout $\epsilon\in]0,\epsilon_0[$, il existe $N_\epsilon>0$ tel que, pour tout x tel que $\|x-x^*\|=\epsilon$, on a $f_{N_\epsilon}(x)>f_{N_\epsilon}(x^*)$. *Preuve.* Comme x^* est un minimum local de f sur la contrainte K, il existe $\epsilon_0>0$ tel que, pour tout $x\in K$ et $\|x-x^*\|\leq\epsilon_0$, on a $f(x)\geq f(x^*)$. Fixons $\epsilon\in]0,\epsilon_0[$ arbitraire. Nous raisonnons maintenant par l'absurde. Supposons que, pour tout N, il existe x_N tel que $\|x_N-x^*\|=\epsilon$ et $f_N(x_N)\leq f(x^*)$. Comme x_N est une suite bornée, x_N converge, à une sous-suite près (sous-suite encore notée (x_n)) vers un point $\bar{x}$. Montrons que $\bar{x}$ appartient à K, vérifie $\|\bar{x}-x^*\|=\epsilon$ et $f(\bar{x})<f(x^*)$. On aura ainsi trouvé une contradiction. Le fait que $\|\bar{x}-x^*\|=\epsilon$ est clair car $\|x_N-x^*\|=\epsilon$. D'autre part, on a

$$f(x_N)+\|x_N-x^*\|^2+\frac{N}{2}\left[\sum_{i\in I}\max(0,g_i(x_N))^2+\sum_{j\in J}h_j(x_N)^2\right]\leq f(x^*) \tag{1.1}$$

Donc

$$0\leq\left[\sum_{i\in I}\max(0,g_i(x_N))^2+\sum_{j\in J}h_j(x_N)^2\right]\leq\frac{2}{N}\left[f(x^*)-f(x_N)-\epsilon^2\right]$$

Lorsque $N \to +\infty$, on obtient

$$0 \leq \left[\sum_{i \in I} \max(0, g_i(\bar{x}))^2 + \sum_{j \in J} h_j(\bar{x})^2 \right] \leq 0$$

c'est-à-dire que $\bar{x}$ appartient à K. Enfin, d'après l'inégalité (1.1) et le fait que $\|x_N - x^*\|^2 = \epsilon^2$, on a aussi

$$f(x_N) + \epsilon^2 \leq f(x^*)$$

Donc, en passant à la limite, on obtient $f(\bar{x}) \leq f(x^*) - \epsilon^2 < f(x^*)$. Or, par définition de x_0, il est impossible d'avoir à la fois $\bar{x} \in K$, $\|\bar{x} - x_0\| \leq \epsilon_0$ et $f(\bar{x}) < f(x_0)$. On a donc obtenu une contradiction avec l'hypothèse. On en déduit que, pour tout $\epsilon \in]0, \epsilon_0[$, il existe $N_\epsilon > 0$ tel que, pour tout x tel que $\|x - x^*\| = \epsilon$, on a $f_{N_\epsilon}(x) > f_{N_\epsilon}(x^*)$.

3. Fixons maintenant $\epsilon \in]0, \epsilon_0[$. Comme, pour tout x tel que $\|x - x^*\| = \epsilon$, on a $f_{N_\epsilon}(x) > f_{N_\epsilon}(x^*)$, la fonction f_{N_ϵ} admet un minimum local sur l'ouvert $\{x,\ \|x - x^*\| < \epsilon\}$ en un point noté x_ϵ (c.f. lemme 1.1.2.) En ce point, la condition nécessaire d'optimalité dans un ouvert s'applique (c.f. théorème 1.2.1), et on obtient

$$\nabla f_{N_\epsilon}(x_\epsilon) = 0 \ ,$$

c'est-à-dire

$$\nabla f(x_\epsilon) + 2(x_\epsilon - x^*) + N \left[\sum_{i \in I} \max(0, g_i(x_\epsilon)) \nabla g_i(x_\epsilon) + \sum_{j \in J} h_j(x_\epsilon) \nabla h_j(x_\epsilon) \right] = 0$$

4. Posons

$$\rho^\epsilon = \left(1 + N^2 \sum_{i \in I} max(0, g_i(x_\epsilon))^2 + N^2 \sum_{j \in J} h_j(x_\epsilon)^2 \right)^{\frac{1}{2}}$$

et

$$p_0^\epsilon = \frac{1}{\rho^\epsilon},\ p_i^\epsilon = N p_0^\epsilon \max(0, g_i(x_\epsilon)),\ q_j^\epsilon = N p_0^\epsilon h_j(x_\epsilon)$$

Lorsque ϵ tend vers 0, x_ϵ tend vers x^*, et le vecteur $(p_0^\epsilon, p^\epsilon, q^\epsilon)$ qui est de norme 1 dans R^{1+l+m} tend, à une sous-suite près, vers un vecteur de norme 1 noté (p_0, p, q). En divisant la condition nécessaire obtenue ci-dessus par ρ^ϵ, et en faisant tendre ϵ vers 0, on obtiend finalement l'égalité

$$p_0 \nabla f(x^*) + \left[\sum_{i \in I} p_i \nabla g_i(x^*) + \sum_{j \in J} p_j \nabla h_j(x^*) \right] = 0$$

5. Remarquons enfin que la condition d'exclusion est satisfaite. En effet, si $g_i(x^*) < 0$ pour un certain i, alors on a $g_i(x^\epsilon) < 0$ pour $\epsilon > 0$ suffisamment petit. Donc $p_i^\epsilon = 0$ par définition. En passant à la limite, on obtient $p_i = 0$.

1.3 Démonstration géométrique du théorème de Kuhn & Tuker

1.3.1 Condition d'Euler abstraite

Soit K un sous-ensemble de $\mathbb{R}^n$ et x un point de K.

Définition 1.3.1 (cône tangent)**.** *On appelle cône tangent à K en x l'ensemble des vecteurs v de $\mathbb{R}^n$ pour lesquels il existe une suite de réels $s_k > 0$ et une suite de vecteurs $v_k \in \mathbb{R}^n$ tels que*

i) $s_k \to 0^+$ et $v_k \to v$,

ii) pour tout k, $x + s_k v_k$ appartient à K.

Notations : On note le cône tangent à K en x : $T_K(x)$.

Proposition 1.3.1. *Le cône tangent est bien un cône, c'est-à-dire que*

$$\forall v \in T_K(x),\ \forall \lambda \geq 0,\ \lambda v \in T_K(x) \ .$$

De plus, c'est un ensemble fermé.

Preuve. Remarquons d'abord que le vecteur nul appartient à $T_K(x)$ par définition de $T_K(x)$.

Soit $v \in T_K(x)$ et $\lambda \geq 0$. Si $v = 0$ ou $\lambda = 0$, alors $\lambda v = 0 \in T_K(x)$. On suppose donc que v et λ sont non nuls.

Soit $v_k \to v$ et $s_k \to 0^+$ tels que $x + s_k v_k$ appartient à K. Alors, en posant $s'_k = s_k/\lambda$ et $v'_k = \lambda v_k$, on a bien que $v'_k \to \lambda v_k$, que $s'_k \to 0^+$ et que $x + s'_k v'_k = x + s_k v_k$ appartient à K. Donc λv appartient bien à $T_K(x)$.

On veut montrer maintenant que $T_K(x)$ est fermé. Pour cela, on considère une suite (v_p) d'éléments de $T_K(x)$, et on suppose que cette suite (v_p) converge vers un vecteur v. On doit montrer que v appartient à $T_K(x)$. Par définition de $T_K(x)$, pour tout p il existe une suite $v_k^p \to v_p$ et une suite $s_k^p \to 0^+$ tels que $x + s_k^p v_k^p$ appartient à K.

Pour tout $l > 0$, on considère p_l et k_l tel que $\|v_{p_l} - v\| \leq 1/(2k)$, $\|v_{p_l} - v_{k_l}^{p_l}\| \leq 1/(2k)$ et $s_{k_l}^{p_l} \leq 1/k$. Alors

$$\|v_{k_l}^{p_l} - v\| \leq \|v_{k_l}^{p_l} - v_{k_l}\| + \|v_{p_l} - v\| \leq 1/(2k) + 1/(2k) \leq 1/k$$

Par conséquent, la suite $w_l = v_{k_l}^{p_l}$ converge vers v et la suite $s_l = s_{k_l}^{p_l}$ tend vers 0^+. De plus, $x + s_l w_l$ appartient à K car

$$x + s_l w_l = x + s_{k_l}^{p_l} v_{k_l}^{p_l} \in K$$

Donc v appartient bien à $T_K(x)$. □

Exercice 1.1. *Soit K le carré de $\mathbb{R}^2$. Calculer $T_K(x, y)$ pour tout $(x, y) \in K$.*

Exercice 1.2. *Soit K la spirale de $\mathbb{R}^2$.*

$$K = \{(x, y) \in \mathbb{R}^2 \mid (x, y) = (0, 0) \text{ ou } \exists t > 0 \text{ avec } (x, y) = (t\cos(1/t), t\sin(1/t))\}$$

Montrer que K est fermé. Calculer $T_K(0, 0)$.

Théorème 1.3.1 (Condition nécessaire d'Euler)**.** *On suppose que la fonction $f : \mathbb{R}^n \to \mathbb{R}$ est de classe $\mathcal{C}^1$. Si x^* est une solution du problème*

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

alors

$$\forall v \in T_K(x^*) \ , \ \langle \nabla f(x^*), v \rangle \ \geq \ 0 \ .$$

Remarques : Pour pouvoir exploiter ce résultat, il faut être capable de calculer l'ensemble $T_K(x^*)$. Cela n'est possible en général que sous des hypothèses de *qualification* de la contrainte.

Preuve. Soit v un vecteur de $T_K(x^*)$. Par définition, il existe s_k tendant vers 0^+ et v_k tendant vers v tels que $x^* + s_k v_k$ appartient à K pour tout k.

Comme x^* est un minimum du problème, on a

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

Or

$$f(x^* + s_k v_k) = f(x^*) + s_k \langle \nabla f(x^*), v_k \rangle + \|s_k v_k\| \epsilon(s_k v_k)$$

où $\epsilon = \epsilon(x)$ est une fonction tendant vers 0 lorsque x tend vers 0. Par conséquent,

$$s_k \langle \nabla f(x^*), v_k \rangle + \|s_k v_k\| \epsilon(s_k v_k) \geq 0 \ .$$

On divise cette inegalité par $s_k > 0$, et on fait tendre k vers $+\infty$. Alors s_k tend vers 0, v_k tend vers v et $s_k v_k$ tend vers 0. On obtient à la limite

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

□

On suppose que les contraintes sont décrites par

$$K = \{x \in \mathbb{R}^n \mid \forall i \in I, \ g_i(x) \leq 0 \text{ et } , \ \forall j \in J, \ h_j(x) = 0\}$$

où $I = \{1, \ldots, l\}$ et $J = \{1, \ldots, m\}$.

Proposition 1.3.2. *Si les applications g_i et h_i sont de classe $\mathcal{C}^1$, alors pour tout $x \in K$,*

$$T_K(x) \subset \{v \in \mathbb{R}^n \mid \forall i \in I(x),\ \langle \nabla g_i(x), v\rangle \leq 0 \ \text{ et }, \ \forall j \in J,\ \ \langle \nabla h_j(x), v\rangle = 0\}$$

où $I(x)$ désigne les indices des contraintes saturées, c'est-à-dire

$$I(x) = \{i \in \{1, \dots, l\} \mid g_i(x) = 0\} \ .$$

Malheureusement, l'inclusion inverse n'est pas toujours vraie. Cela conduit à la définition suivante

Définition 1.3.2. *On dit que la contrainte K est* **qualifiée** *en $x \in K$ si*

$$T_K(x) = \{v \in \mathbb{R}^n \mid \forall i \in I(x),\ \langle \nabla g_i(x), v\rangle \leq 0 \ \text{ et }, \ \forall j \in J,\ \ \langle \nabla h_j(x), v\rangle = 0\}$$

Nous verrons plus loin que cette définition coïncide bien avec la définition de qualification de la première partie lorsque les contraintes sont "non linéaires".

La notion de qualification dépend en fait de la description de la contrainte (c'est-à-dire des fonctions g_i et h_j). Ce n'est pas une description géométrique. Voici des exemples de contraintes non qualifiées :

Exercice 1.3. *Soit*

$$K = \{(x, y) \in \mathbb{R}^2 \mid y \geq x^2,\ y \leq 0\}$$

Montrer que la contrainte K n'est pas qualifiée au point $(0, 0)$. Même question, toujours au point $(0, 0)$, pour la contrainte

$$K = \{(x, y) \in \mathbb{R}^2 \mid y \leq x^3,\ \ y \geq 0\}$$

Preuve de la proposition. Soit $v \in T_K(x)$. Alors il existe $s_k \to 0^+$ et $v_k \to v$ tels que

$$\forall k \geq 0,\ x + s_k v_k \in K \ .$$

Cela signifie que

$$\forall k \geq 0,\ \forall i \in I,\ g_i(x + s_k v_k) \leq 0 \text{ et }, \ \forall j \in J,\ h_j(x + s_k v_k) = 0 \ .$$

Si $i \in I(x)$, c'est-à-dire $g_i(x) = 0$, alors

$$0 \geq g_i(x + s_k v_k) = s_k \langle \nabla g_i(x), v_k\rangle + \|s_k v_k\| \epsilon(s_k v_k)$$

ce qui implique, après avoir divisé par s_k et fait tendre k vers $+\infty$ que

$$\langle \nabla g_i(x), v\rangle \leq 0 \ .$$

Cette inégalité est donc vraie pour tout $i \in I(x)$. D'autre part, pour tout $j \in J$, on a

$$0 = h_j(x + s_k v_k) = s_k \langle \nabla h_j(x), v_k\rangle + \|s_k v_k\| \epsilon(s_k v_k)$$

ce qui implique, après avoir divisé par s_k et fait tendre k vers $+\infty$ que

$$\langle \nabla h_j(x), v\rangle = 0 \ .$$

Cette égalité est vraie pour tout $j \in J$. On en déduit le résultat annoncé. □

1.3.2 Le lemme de Farkas

L'objet de cette partie est de retrouver le théorème de Kuhn & Tucker, en utilisant la condition nécessaire d'Euler.

Lemme 1.3.1 (Lemme de Farkas)**.** *Soient c et c_i pour $i = 1, \dots, k$ des vecteurs de $\mathbb{R}^n$. On fait l'hypothèse suivante :*

$$\forall v \in \mathbb{R}^n, \quad \text{si } \forall i = 1, \dots, k,\ \langle c_i, v\rangle \geq 0, \quad \text{alors } \langle c, v\rangle \geq 0 \ .$$

Dans ce cas, il existe des réels $\lambda_i \geq 0$ pour $i = 1, \dots, l$ tels que

$$c = \sum_{i=1}^{k} \lambda_i c_i \ .$$

La preuve du lemme de Farkas est assez longue. Elle repose sur la remarque fondamentale suivante :

Proposition 1.3.3. *Soit $\{a_1, \dots, a_k\}$ une famille de $\mathbb{R}^N$. L'ensemble C défini par*

$$C = \{x \in \mathbb{R}^N \mid \exists \lambda_1 \geq 0, \dots, \exists \lambda_k \geq 0,\ x = \lambda_1 a_1 + \cdots + \lambda_k a_k\}$$

est un cône convexe fermé de $\mathbb{R}^N$.

Preuve de la proposition. L'ensemble C est clairement un cône convexe. La véritable difficulté est le fait que C est fermé. Nous montrons ce résultat par récurrence sur k.

Pour $k = 1$, il est clair que l'ensemble C est fermé.

On suppose le résultat démontré pour $k-1$. Montrons-le pour k. Soit $\mathcal{B} = \{a_1, \dots, a_k\}$ une famille de k vecteurs de $\mathbb{R}^N$. Nous devons prouver que l'ensemble

$$C = \{x \in \mathbb{R}^N \mid \exists \lambda_1 \geq 0, \dots, \exists \lambda_k \geq 0,\ x = \lambda_1 a_1 + \cdots + \lambda_k a_k\}$$

est fermé On considère deux cas :

Cas 1 : la famille $\mathcal{B} = \{a_1, \dots, a_k\}$ est libre. Soit F le sous-espace vectoriel de $\mathbb{R}^N$ engendré par cette famille. Notons que $\{a_1, \dots, a_k\}$ est une base de F. De plus, F étant de dimension finie, F est fermé. Par conséquent, si une suite de vecteurs x_n de C converge vers un vecteur x de $\mathbb{R}^N$, alors x appartient nécessairement à F. Soient $(\lambda_1^n, \dots, \lambda_k^n)$ les coordonnées de x_n dans la base $\mathcal{B}$ et $(\lambda_1, \dots, \lambda_k)$ celles de x dans cette base. La convergence de x_n vers x entraîne que les suites (λ_i^n) (pour $i = 1, \dots, k$) convergent vers λ_i. Comme les x_n appartiennent à C, on a

$$\forall n \geq 0,\ \forall i = 1, \dots, k,\ \lambda_i^n \geq 0\ .$$

En passant à la limite, on obtient que $\lambda_i \geq 0$ pour $i = 1, \dots, k$. Donc x appartient à C.

Cas 2 : la famille $\mathcal{B}$ est liée. Notons

$$C_i = \{x \in \mathbb{R}^N,\ \forall j \neq i,\ \exists \lambda_j \geq 0 \text{ tel que } x = \sum_{j \neq i} \lambda_j x_j\}$$

Par hypothèse de récurrence, les ensembles C_j sont fermés (car les familles $\{a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_k\}$ possèdent $k-1$ éléments). Nous allons montrer que

$$C = \bigcup_{i=1}^{k} C_i\ , \tag{1.2}$$

ce qui prouvera clairement que C est fermé. Notons d'abord que l'inclusion $\bigcup_{i=1}^k C_i \subset C$ est évidente. Montrons l'inclusion inverse. Soit x appartenant à C. Soient $\lambda_1, \dots, \lambda_k$ des réels positifs ou nuls tels que

$$x = \lambda_1 a_1 + \cdots + \lambda_k a_k\ .$$

Comme la famille $\mathcal{B}$ est liée, il existe une famille $\alpha_1, \dots, \alpha_k$, avec les α_i non tous nuls, telle que

$$\alpha_1 a_1 + \dots \alpha_k a_k = 0\ .$$

Quitte à multiplier l'égalité précédente par -1, on peut supposer qu'au moins un des coeffients α_i est strictement négatif. Notons

$$I = \{i \in \{1, \dots, k\}\ ,\ \alpha_i < 0\}\ .$$

Posons

$$t = \min_{i \in I} \lambda_i / |\alpha_i|\ .$$

Alors, pour tout $j \in \{1, \dots, k\}$, on a : $\lambda_j + t\alpha_j \geq 0$, et pour au moins un $i \in I$ (celui pour lequel on a $t = \lambda_i/|\alpha_i|$), on a l'égalité : $\lambda_i + t\alpha_i = 0$. Par conséquent :

$$x = (\lambda_1 + t\alpha_i)a_1 + \dots (\lambda_k + t\alpha_k)a_k$$

appartient à C_i. Ceci établit l'égalité (1.2) et conclut la preuve de la proposition. □

Avant d'aborder la démonstration du lemme de Farkas, rappelons l'énoncé du théorème de séparation :

Théorème 1.3.2. *Soient C_1 et C_2 deux sous-ensembles convexes de $\mathbb{R}^N$. On suppose que C_1 est fermé et C_2 compact. Alors il existe un vecteur $q \in \mathbb{R}^N$ et une constante α tels que*

$$\max_{x\in C_2}\langle q, x\rangle \quad < \quad \alpha \quad \leq \quad \inf_{y\in C_1}\langle q, y\rangle \ .$$

Preuve du lemme de Farkas. On raisonne par l'absurde, en supposant que le vecteur c n'appartient pas à l'ensemble

$$C = \{x \in \mathbb{R}^N \mid \exists \lambda_1 \geq 0, \ldots, \exists \lambda_k \geq 0,\ x = \lambda_1 c_1 + \cdots + \lambda_k c_k\} \ .$$

Comme l'ensemble $C_1 = C$ est convexe fermé et que l'ensemble $C_2 = \{y\}$ est convexe compact, le théorème de séparation affirme qu'il existe un vecteur $q \in \mathbb{R}^N$ et une constante α tels que

$$\langle q, c\rangle \ < \ \alpha \ \leq \ \inf_{y\in C}\langle q, y\rangle \ .$$

Comme C est un cône, on voit facilement que

$$\inf_{y\in C}\langle q, y\rangle = 0 \ ,$$

et en particulier que $\alpha \leq 0$. Comme les vecteurs $c_1, \ldots, c_k$ appartiennent à C, on a

$$\forall i \in \{1, \ldots, k\},\ \langle q, c_i\rangle \geq 0 \ .$$

De plus, $\langle q, c\rangle < \alpha \leq 0$. Ceci est en contradiction avec l'hypothèse du lemme de Farkas. □

Exercice 1.4. *Démontrer la réciproque du lemme de Farkas.*

Exercice 1.5. *Soient c, c_i pour $i = 1, \ldots, l$ et d_j pour $j = 1, \ldots, m$ des vecteurs de $\mathbb{R}^n$. On fait l'hypothèse suivante :*

$$\forall v \in \mathbb{R}^n,\ [\text{si } \forall i = 1, \ldots, l,\ \langle c_i, v\rangle \geq 0 \text{ et },\ \forall j = 1, \ldots, m,\ \langle d_j, v\rangle = 0]\,,\ \text{alors } \langle c, v\rangle \geq 0 \ .$$

Montrer que, dans ce cas, il existe des réels $\lambda_i \geq 0$ pour $i = 1, \ldots, l$ et $\mu_j \in \mathbb{R}$ pour $j = 1, \ldots, m$ tels que

$$c = \sum_{i=1}^{l} \lambda_i c_i + \sum_{j=1}^{m} \mu_j d_j \ .$$

On suppose toujours que la contrainte est de la forme

$$K = \{x \in \mathbb{R}^n \mid \forall i \in I,\ g_i(x) \leq 0 \text{ et },\ \forall j \in J,\ h_j(x) = 0\}$$

où $I = \{1, \ldots, l\}$ et $J = \{1, \ldots, m\}$ et où les applications g_i et h_j sont de classe $\mathcal{C}^1$.

Exercice 1.6. *Redémontrer le théorème de Kuhn & Tucker : Si x^* est un minimum local du problème $(\mathcal{P})$, et si les contraintes sont qualifiées au point x^*, alors il existe $\lambda_i \geq 0$ (pour $i \in I$) et $\mu_j \in \mathbb{R}$ (pour $j \in J$) tels que*

$$\begin{cases} i) & \nabla f(x^*) + \sum_{i=1}^{l} \lambda_i \nabla g_i(x^*) + \sum_{j=1}^{m} \mu_j \nabla h_j(x^*) = 0 \\ ii) & \sum_{i=1}^{l} \lambda_i g_i(x^*) = 0 \end{cases}$$

Dans le reste de ce chapitre, nous énonçons diverses hypothèses sur les fonctions g_i et h_j pour que les contraintes soient qualifiées. Nous procédons par ordre de difficulté croissante

1. pour les contraintes affines
2. pour les contraintes d'inégalités seules
3. pour les contraintes d'égalités seules
4. dans le cas général

1.3.3 Les contraintes affines

Rappelons qu'une fonction $\phi : \mathbb{R}^n \to \mathbb{R}$ est dite affine s'il existe un vecteur $a \in \mathbb{R}^n$ et un réel b tels que

$$\forall x \in \mathbb{R}^n,\ \phi(x) = \langle a, x\rangle + b\ .$$

Nous montrons maintenant que les contraintes affines sont qualifiées.

Lemme 1.3.2. *Soit $\phi : \mathbb{R}^n \to \mathbb{R}$ une fonction affine. Alors*

$$\forall x, y \in \mathbb{R}^n,\ \phi(y) = \phi(x) + \langle \nabla\phi(x), (y - x)\rangle\ .$$

Preuve. Si ϕ est affine, il existe $a \in \mathbb{R}^n$ et $b \in \mathbb{R}$ tels que

$$\forall x \in \mathbb{R}^n,\ \phi(x) = \langle a, x\rangle + b\ .$$

Alors $\nabla\phi(x) = a$, et

$$\phi(y) = \langle a, y\rangle + b = \langle a, x\rangle + b + \langle a, y - x\rangle = \phi(x) + \langle \nabla\phi(x), (y - x)\rangle\ .$$

□

Proposition 1.3.4. *Si les fonctions g_i et h_j sont affines, alors la contrainte K est qualifiée en tout point.*

En particulier, on peut appliquer le théorème de Kuhn & Tucker *sans autre hypothèse que le fait que les fonctions sont affines.*

Preuve. On a déjà vu (Proposition 1.3.2) que

$$T_K(x) \subset \{v \in \mathbb{R}^n \mid \forall i \in I(x),\ \langle \nabla g_i(x), v\rangle \leq 0 \ \text{ et }\ , \forall j \in J,\ \ \langle \nabla h_j(x), v\rangle = 0\}$$

où

$$I(x) = \{i \in \{1, \ldots, l\} \mid g_i(x) = 0\}\ .$$

Montrons l'inclusion inverse. Soit $v \in \mathbb{R}^n$ tel que

$$\forall i \in I(x),\ \langle \nabla g_i(x), v\rangle \leq 0 \text{ et } \forall j \in J,\ \langle \nabla h_j(x), v\rangle = 0\ .$$

Il faut montrer que $v \in T_K(x)$. Soit $s > 0$. Alors, pour tout $i \in I(x)$,

$$g_i(x + sv) = g_i(x) + s\langle \nabla g_i(x), v\rangle\ \leq\ 0$$

et pour tout $j \in I$,

$$h_j(x + sv) = h_j(x) + s\langle \nabla h_i(x), v\rangle\ =\ 0$$

car les fonctions g_i et h_j sont affines. Reste à voir ce qui se passe pour les contraintes d'inégalité g_i si $i \notin I(x)$. On sait que, dans ce cas, $g_i(x) < 0$. Donc, par continuité de g_i, il existe $\overline{s_i} > 0$ tel que

$$\forall s \in [0, \overline{s_i}],\ g_i(x + sv) < 0\ .$$

Prenons $\bar{s} = \min_{i \notin I(x)} \overline{s_i}$. Alors $\bar{s} > 0$ et

$$\forall s \in [0, \bar{s}],\ \forall i \in I,\ g_i(x + sv) \leq 0 \text{ et }\ , \forall j \in J,\ h_j(x + sv) = 0\ .$$

Si on prend $s_k = \bar{s}/k$ et $v_k = v$, on a montré que

$$\forall k > 0,\ x + s_k v_k \in K\ ,$$

ce qui prouve que v appartient à $T_K(x)$. □

Exercice 1.7. *On suppose que les fonctions g_i sont concaves et de classe $\mathcal{C}^1$, et que les fonctions h_j sont affines. Montrer qu'alors la contrainte K est qualifiée en tout point.*

1.3.4 Les contraintes d'inégalités

On suppose dans cette partie qu'il n'y a pas de contraintes d'égalités, c'est-à-dire que

$$K = \{x \in \mathbb{R}^n \mid \forall i \in I,\ g_i(x) \leq 0\}$$

où g_i sont applications de classe $\mathcal{C}^1$ de $\mathbb{R}^n$ dans $\mathbb{R}$.

Proposition 1.3.5. *Soit x un point de K. La contrainte K est qualifiée en x si la condition suivante est satisfaite :*
il existe un vecteur $\bar{v} \in \mathbb{R}^n$ tel que, pour tout $i \in I(x)$, $\langle \nabla g_i(x), \bar{v} \rangle < 0$.

On rappelle que $I(x)$ est l'ensemble des indices des contraintes saturées au point x, c'est-à-dire

$$I(x) = \{i \in I \mid g_i(x) = 0\}\ .$$

Remarquons que si $I(x) = \emptyset$, la contrainte est qualifiee en x.

Preuve. 1. On sait déjà (Proposition 1.3.2) que l'inclusion suivante est vérifiée :

$$T_K(x) \subset \{v \in \mathbb{R}^n \mid \forall i \in I(x),\ \langle \nabla g_i(x), v \rangle \leq 0\}\ .$$

Pour montrer que la contrainte est qualifiée, il faut démontrer l'inégalité inverse.

2. Pour cela, on commence par montrer que

$$\{v \in \mathbb{R}^n \mid \forall i \in I(x),\ \langle \nabla g_i(x), v \rangle < 0\} \subset T_K(x)$$

Preuve : Soit v dans l'ensemble de gauche. Comme les fonctions g_i sont de classe $\mathcal{C}^1$, on a, pour tout $i \in I(x)$, et pour tout $s > 0$,

$$g_i(x + sv) = g_i(x) + s\langle \nabla g_i(x), v \rangle + \|sv\|\epsilon(sv) = g_i(x) + s(\langle \nabla g_i(x), v \rangle + \|v\|\epsilon(sv))$$

où ϵ tend vers 0 lorsque sv tend vers 0. Comme $g_i(x) = 0$ et $\langle \nabla g_i(x), v \rangle < 0$, il existe $\overline{s_i} > 0$ tel que

$$\forall s \in [0, \overline{s_i}],\ \langle \nabla g_i(x), v \rangle + \|v\|\epsilon(sv) < 0$$

Donc, pour tout $i \in I(x)$ et pour tout $s \in [0, \overline{s_i}]$, on a $g_i(x + sv) < 0$.
D'autre part, si $i \notin I(x)$, c'est-à-dire $g_i(x) < 0$, un argument de continuité nous permet d'affirmer qu'il existe $\overline{s_i} > 0$ tel que

$$\forall s \in [0, \overline{s_i}],\ g_i(x + sv) < 0$$

Posons $\bar{s} = \min_{i \in I} s_i$. On a montré que

$$\forall s \in [0, \bar{s}],\ \forall i \in I,\ g_i(x + sv) < 0$$

ce qui prouve que

$$\forall s \in [0, \bar{s}],\ x + sv \in K\ .$$

Si on choisit $s_k = \bar{s}/k$ et $v_k = v$, on a montré que

$$\forall k > 0\ ,\ x + s_k v_k \in K\ ,$$

ce qui implique que v appartient à $T_K(x)$.

3. On montre maintenant que, pour tout v tel que

$$\forall i \in I(x),\ \langle \nabla g_i(x), v \rangle \leq 0\ ,$$

il existe une suite v_k de limite v telle que

$$\forall i \in I(x),\ \langle \nabla g_i(x), v_k \rangle < 0\ ,$$

Preuve : Pour ce faire, on va utiliser l'hypothèse de la proposition : il existe un vecteur $\bar{v} \in \mathbb{R}^n$ tel que, pour tout $i \in I(x)$, $\langle \nabla g_i(x), \bar{v} \rangle < 0$.
Alors, en prenant $v_k = v + \frac{1}{k}\bar{v}$, on montre facilement le résultat.

4. On affirme finalement que l'inégalité désirée

$$\{v \in \mathbb{R}^n \mid \forall i \in I(x),\ \langle \nabla g_i(x), v\rangle \ \leq\ 0\} \subset T_K(x)$$

est vérifiée.
Preuve : Soit v tel que

$$\forall i \in I(x),\ \langle \nabla g_i(x), v\rangle \ \leq\ 0\ .$$

D'après l'étape précédente, il existe v_k suite convergeant vers v telle que

$$\forall i \in I(x),\ \langle \nabla g_i(x), v_k\rangle \ <\ 0\ .$$

D'après l'étape 2 ces inégalité impliquent que $v_k \in T_K(x)$. Or l'ensemble $T_K(x)$ est fermé (voir Proposition 1.3.1). Donc v appartient aussi à $T_K(x)$.

□

1.3.5 Le cas des contraintes d'égalités

On suppose dans cette partie qu'il n'y a pas de contraintes d'inégalités, c'est-à-dire que

$$K = \{x \in \mathbb{R}^n \mid \forall j \in J,\ h_j(x) = 0\}$$

où h_j sont applications de classe $\mathcal{C}^1$ de $\mathbb{R}^n$ dans $\mathbb{R}$.

Proposition 1.3.6. *Soit x un point de K. La contrainte K est qualifiée en x si la condition suivante est satisfaite :*
la famille $\{\nabla h_1(x), \ldots, \nabla h_m(x)\}$ est libre.

Preuve. 1. On sait déjà (Proposition 1.3.2) que l'inclusion suivante est vérifiée :

$$T_K(x) \subset \{v \in \mathbb{R}^n \mid \forall j \in J,\ \langle \nabla h_j(x), v\rangle = 0\}\ .$$

Pour montrer que la contrainte est qualifiée, il faut démontrer l'inégalité inverse.

2. On commence par des préliminaires. Montrons d'abord qu'il existe un voisinage U de x et une constante $\alpha > 0$ tels que, pour tout $x' \in U$, on a

$$\min_{y\in\mathbb{R}^m,\ \|y\|=1} \left\| \sum_{j=1}^{m} y_j \nabla h_j(x') \right\| \geq \alpha > 0$$

Preuve : On raisonne par l'absurde. Si l'assertion est fausse, il existe une suite (x_n) de limite x et une suite (y^n), de norme 1, telles que

$$\lim_n \left\| \sum_{j=1}^{m} y_j^n \nabla h_j(x_n) \right\| = 0$$

Il existe une sous-suite $(y^{n'})$ de (y^n) qui converge vers un y de norme 1. En passant à la limite dans l'égalité ci-dessus, on a

$$\left\| \sum_{j=1}^{m} y_j \nabla h_j(x) \right\| = 0$$

ce qui impose que

$$\sum_{j=1}^{m} y_j \nabla h_j(x) = 0$$

Or la famille $\{\nabla h_1(x), \ldots, \nabla h_m(x)\}$ est libre, et donc $y_1 = \cdots = y_m = 0$. Ceci est en contradiction avec $\|y\| = 1$.

3. Posons

$$\theta_h = \left(\sum_{j=1}^{m} (h_j(x + hv))^2 \right)^{\frac{1}{2}}$$

On affirme que

$$\lim_{h\to 0^+} \frac{\theta_h}{h} = 0\ .$$

Preuve : En effet,

$$h_j(x+hv) = h_j(x) + h\langle \nabla h_j(x), v\rangle + h\|v\|\epsilon(hv) = h\|v\|\epsilon_j(hv)$$

car $h_j(x) = h\langle \nabla h_j(x), v\rangle = 0$. Donc

$$\theta_h = \left(\sum_{j=1}^m h^2\|v\|^2(\epsilon_j(hv))^2\right)^{\frac{1}{2}} = h\|v\|\left(\sum_{j=1}^m(\epsilon_j(hv))^2\right)^{\frac{1}{2}}$$

et

$$\theta_h/h = \left(\sum_{j=1}^m(\epsilon_j(hv))^2\right)^{\frac{1}{2}} \to 0 \text{ si } h \to 0^+ .$$

4. Soit $v \in \mathbb{R}^n$ tel que

$$\forall j \in J,\ \langle \nabla h_j(x), v\rangle = 0 .$$

On considère maintenant pour tout $h > 0$ la fonction

$$\phi_h(w) = \left(\sum_{j=1}^m(h_j(x+hv+w))^2\right)^{\frac{1}{2}} + \frac{1}{2h}\|w\|^2$$

On remarque que

$$\phi_h(0) = \theta_h$$

et

$$\forall w \in \mathbb{R}^n, \text{ avec } \|w\| = (2h\theta_h)^{\frac{1}{2}},\ \phi(w) \geq \theta_h .$$

Donc, d'après la proposition 1.1 de la première partie, ϕ_h admet un minimum w_h dans la boule ouverte de centre 0 et de rayon $(2h\theta_h)^{\frac{1}{2}}$.

Supposons un instant que $\left(\sum_{j=1}^m(h_j(x+h_kv+w_{h_k}))^2\right)^{\frac{1}{2}}$ soit nul pour une suite h_k tendant vers 0^+. Alors, en posant $v_k = v + w_{h_k}/h_k$, le point $x + h_kv_k$ appartient à K. De plus, v_k tend vers v lorsque $k \to +\infty$ car

$$\|w_{h_k}/h_k\| \leq (2h_k\theta_{h_k})^{\frac{1}{2}}/h_k \to 0$$

d'après l'étape 3. Par conséquent, on a montré que v appartient à $T_K(x)$, ce qui était le résultat demandé.

5. Reste à montrer que $\left(\sum_{j=1}^m(h_j(x+h_kv+w_{h_k}))^2\right)^{\frac{1}{2}}$ est nul pour au moins une suite $h_k \to 0^+$. Pour cela, on raisonne par l'absurde, en supposant que la quantité

$$\left(\sum_{j=1}^m(h_j(x+hv+w_h))^2\right)^{\frac{1}{2}}$$

ce n'est pas nulle pour h suffisament petit. Dans ce cas, on peut écrire les conditions nécessaires d'optimalité pour w_h :

$$2\frac{\sum_{j=1}^m h_j\nabla h_j}{\left(\sum_{j=1}^m(h_j)^2\right)^{\frac{1}{2}}} + \frac{1}{h}w_h = 0$$

où on a omis la dépendance en $x + hw_h$ dans les h_j. Posons $y_j^h = \frac{h_j}{\left(\sum_{j=1}^m(h_j)^2\right)^{\frac{1}{2}}}$. Alors (y_i^h) est un vecteur de $\mathbb{R}^m$ de norme 1. Par conséquent, d'après la partie 2,

$$2\alpha \leq \left\| 2\frac{\sum_{j=1}^m h_j\nabla h_j}{\left(\sum_{j=1}^m(h_j)^2\right)^{\frac{1}{2}}}\right\| = \frac{1}{h}\|w_h\| \leq \frac{1}{h}(2h\theta_h)^{\frac{1}{2}}/h \to 0 \text{ quand } h \to 0^+ .$$

On a donc une contradiction.

□

1.3.6 Le cas général

On suppose maintenant que l'on est dans le cas général, c'est-à-dire que

$$K = \{x \in \mathbb{R}^n \mid \forall i \in I,\ g_i(x) \leq 0 \text{ et },\ \forall j \in J,\ h_j(x) = 0\}$$

où $I = \{1, \ldots, l\}$ et $J = \{1, \ldots, m\}$, et où g_i et h_j sont applications de classe $\mathcal{C}^1$ de $\mathbb{R}^n$ dans $\mathbb{R}$.

Proposition 1.3.7. *Soit x un point de K. La contrainte K est qualifiée en x si les conditions suivantes sont satisfaites :*
(i) la famille $\{\nabla h_1(x), \ldots, \nabla h_m(x)\}$ est libre.
(ii) Il existe un vecteur $\bar{v} \in \mathbb{R}^n$ tel que

$$\forall j \in J,\ \langle \nabla h_j(x), \bar{v}\rangle = 0 \text{ et } \forall i \in I(x),\ \langle \nabla g_i(x), \bar{v}\rangle\ <\ 0\ .$$

Preuve. 1. On sait déjà (Proposition 1.3.2) que l'inclusion suivante est vérifiée :

$$T_K(x) \subset \{v \in \mathbb{R}^n \mid \forall i \in I(x),\ \langle \nabla g_i(x), v\rangle \leq 0\ \text{ et },\ \forall j \in J,\ \langle \nabla h_j(x), v\rangle = 0\}\ .$$

Comme d'habitude, il faut démontrer l'inégalité inverse.

2. On va d'abord montrer que

$$\{v \in \mathbb{R}^n \mid \forall i \in I(x),\ \langle \nabla g_i(x), v\rangle\ <\ 0\ \text{ et },\ \forall j \in J,\ \langle \nabla h_j(x), v\rangle = 0\} \subset T_K(x)\ .$$

Appelons C l'ensemble de gauche. Considérons $\widetilde{K}$ la contrainte constituée uniquement des conditions d'égalité :

$$\widetilde{K} = \{x \in \mathbb{R}^n \mid \forall j \in J,\ h_j(x) = 0\}\ .$$

D'après l'hypothèse (i) et la proposition 1.3.6, on sait que la contrainte $\widetilde{K}$ est qualifiée en x. Donc

$$T_{\widetilde{K}}(x) = \{v \in \mathbb{R}^n \mid \forall j \in J,\ \langle \nabla h_j(x), v\rangle = 0\}\ .$$

Par conséquent, pour tout v appartenant à C, il existe une suite $s_k \to 0^+$ et une suite $v_k \to v$ telle que $x + s_k v_k \in \widetilde{K}$. Montrons qu'en fait, $x + s_k v_k \in K$ dès que k est suffisament grand.
Soit $i \in I(x)$. Alors $g_i(x) = 0$ et $\langle \nabla g_i(x), v\rangle\ <\ 0$. Donc

$$g_i(x + s_k v_k) = s_k \langle \nabla g_i(x), v_k\rangle + \|s_k v_k\|\epsilon(s_k v_k) = s_k(\langle \nabla g_i(x), v_k\rangle + \|v_k\|\epsilon(s_k v_k)) \leq 0$$

dès que $k \geq k_i$ pour un certain k_i.
Par les arguments habituels, si $i \notin I(x)$ alors il existe k_i tel que pour tout $k \geq k_i$, on a $g_i(x + s_k v_k) \leq 0$. Donc si on choisit $\bar{k} = \max_i k_i$, il est clair que

$$\forall k \geq \bar{k},\ g_i(x + s_k v_k) \leq 0$$

Comme on savait déjà que $x + s_k v_k \in \widetilde{K}$, c'est-à-dire que $\forall j \in J$, $h_j(x + s_k v_k) = 0$, on en déduit que

$$\forall k \geq \bar{k},\ x + s_k v_k \in K\ .$$

Ceci implique que v appartient à $T_K(x)$.

3. Pour conclure la démonstration, il suffit de montrer que

$$\overline{C} = \{v \in \mathbb{R}^n \mid \forall i \in I(x),\ \langle \nabla g_i(x), v\rangle\ \leq\ 0\ \text{ et },\ \forall j \in J,\ \langle \nabla h_j(x), v\rangle = 0\}$$

où $\overline{C}$ est la fermeture de C. En effet, comme $C \subset T_K(x)$ et que $T_K(x)$ est fermé, on aura le résultat désiré. Remarquons d'abord que l'inclusion

$$\overline{C} \subset \{v \in \mathbb{R}^n \mid \forall i \in I(x),\ \langle \nabla g_i(x), v\rangle\ \leq\ 0\ \text{ et },\ \forall j \in J,\ \langle \nabla h_j(x), v\rangle = 0\}$$

est évidente. Il suffit donc de montrer que, si $v \in \mathbb{R}^n$ est tel que

$$\forall i \in I(x),\ \langle \nabla g_i(x), v\rangle\ \leq\ 0\ \text{ et },\ \forall j \in J,\ \langle \nabla h_j(x), v\rangle = 0$$

alors $v \in \overline{C}$. On considère $v_k = v + \frac{1}{k}\overline{v}$, où $\bar{v}$ est donné par les hypothèses. Alors

$$\forall j \in J,\ \langle \nabla h_j(x), v_k\rangle = \langle \nabla h_j(x), v\rangle + \frac{1}{k}\langle \nabla h_j(x), \bar{v} >= 0$$

et

$$\forall i \in I(x),\ \langle \nabla g_i(x), v_k\rangle = \langle \nabla g_i(x), v\rangle + \frac{1}{k}\langle \nabla g_i(x), \bar{v}\rangle\ <\ 0\ .$$

Donc v_k appartient à C. On en déduit que v appartient à $\overline{C}$.

□

1.3.7 Méthode de résolution d'un problème de minimisation

On résout un problème de minimisation sous contraintes en quatre étapes

1. On montre *a priori* que le problème admet une solution.
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 où on ne peut pas appliquer la proposition 1.3.7.
3. On cherche ensuite les points satisfaisant les conditions nécessaires de Kuhn & Tucker. On note cette 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 : Lorsque le problème est convexe, il suffit de trouver des points vérifiant les conditions nécessaires d'optimalité pour pouvoir conclure à l'existence d'un minimum.

1.4 Problèmes convexes et dualité

On considère le problème $(\mathcal{P})$

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

où $K = \{x \in \mathbb{R}^n \mid g_1(x) \leq 0, \ldots, g_l(x) \leq 0\}$. On suppose que les applications f, $g_1, \ldots, g_l$ sont convexes et de classe $\mathcal{C}^1$.

1.4.1 Une condition de qualification de la contrainte

Proposition 1.4.1. *On suppose que l'intérieur de K est non vide. Alors la contrainte K est qualifiée.*

Preuve. Fixons x_0 un point à l'intérieur de K. Soit x un point du bord de K, et posons $v = (x_0 - x)$. 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 $i \in I(x)$,

$$\langle \nabla g_i(x), v \rangle \ \leq \ g_i(x_0) - g_i(x) \ < \ 0 \text{ car } g_i(x) = 0 \ .$$

Donc la contrainte est qualifiée d'après la proposition 1.2.1. □

1.4.2 Le théorème de Kuhn & Tucker exprimé en termes de Lagrangien

Il est habituel d'introduire le lagrangien du problème. Le lagrangien est une application L de $\mathbb{R}^n \times \mathbb{R}^l$ dans $\mathbb{R}$ définie par

$$L(x, \lambda) = f(x) + \lambda^T g(x) = f(x) + \sum_{i=1}^{l} \lambda_i g_i(x) \ .$$

Le théorème de Kuhn & Tucker s'écrit alors

Théorème 1.4.1. *Si un point x^* est un minimum du problème $\mathcal{P}$ et si K est qualifiée en x^*, alors il existe $\lambda^* \in \mathbb{R}^l_+$ avec*

$$\begin{cases} i) & (\lambda^*)^T g(x^*) = 0 \quad \text{condition d'exclusion} \\ ii) & \nabla_x L(x^*, \lambda^*) = 0 \quad \text{condition nécessaire} \end{cases}$$

où

$$\nabla_x L(x^*, \lambda^*) = \nabla f(x^*) + (Jg(x^*))^T \lambda^* = \nabla f(x^*) + \sum_{i=1}^{l} \lambda_i^* \nabla g_i(x^*)$$

1.4.3 Les conditions nécessaires sont suffisantes

Lemme 1.4.1. *Soit $F : \mathbb{R}^n \to \mathbb{R}$ une application convexe de classe $\mathcal{C}^1$. Le point x^* est un minimum de F sur $\mathbb{R}^n$, si et seulement si, $\nabla F(x^*) = 0$.*

Preuve. Dans un sens, c'est évident. Supposons que $\nabla F(x^*) = 0$. Alors, pour tout x de $\mathbb{R}^n$, on a

$$F(x) - F(x^*) \ \geq \ \langle \nabla F(x^*), x - x^* \rangle$$

car F est convexe. Or $\nabla F(x^*) = 0$ par hypothèse. Donc x^* est un minimum de F. □

Théorème 1.4.2. *Si la contrainte est qualifiée, tout point vérifiant les conditions nécessaires d'optimalité est un minimum du problème $(\mathcal{P})$.*

Preuve. Soit x^* vérifiant les conditions nécessaires d'optimalité. Alors il existe $\lambda^* \in \mathbb{R}^l_+$ tel que

$$\begin{cases} i) & (\lambda^*)^T g(x^*) = 0 \quad \text{condition d'exclusion} \\ ii) & \nabla_x L(x^*, \lambda^*) = 0 \quad \text{condition nécessaire} \end{cases}$$

L'égalité (ii) signifie que x^* est un minimum de la fonction $x \to L(x, \lambda^*)$. On a donc

$$\forall y \in \mathbb{R}^n, \ \ L(y, \lambda^*) \geq L(x^*, \lambda^*)$$

Or

$$L(x^*, \lambda^*) = f(x^*) + (\lambda^*)^T g(x^*) = f(x^*)$$

grâce à la condition d'exclusion. De plus, pour tout $y \in K$,

$$L(y, \lambda^*) = f(y) + (\lambda^*)^T g(y) \leq f(y)$$

car $\lambda_i^* \geq 0$ et $g_i(y) \leq 0$, c'est-à-dire $(\lambda^*)^T g(y) \leq 0$. On peut déduire des trois relations précédentes que

$$f(x^*) = L(x^*, \lambda^*) \leq L(y, \lambda^*) \leq f(y)$$

pour tout $y \in K$. Donc x^* est bien un minimum du problème $(\mathcal{P})$. □

1.4.4 Le théorème de dualité

Théorème 1.4.3. *On suppose que la contrainte est qualifiée et que le problème $(\mathcal{P})$ admet au moins une solution. Alors*

$$\min_{x \in K} f(x) = \sup_{\lambda \in \mathbb{R}^l_+} \inf_{x \in \mathbb{R}^n} L(x, \lambda) = \inf_{x \in \mathbb{R}^n} \sup_{\lambda \in \mathbb{R}^l_+} L(x, \lambda)$$

De plus, le problème $\sup_{\lambda \in \mathbb{R}^l_+} \inf_{x \in \mathbb{R}^n} L(x, \lambda)$ a au moins une solution λ^* et le problème $\inf_{x \in \mathbb{R}^n} L(x, \lambda^*)$ a une solution x^* qui est solution du problème $(\mathcal{P})$.

Attention : le résultat est faux si f ou un des g_i est non convexe. Exemple :

$$\min_{-\frac{1}{2} \leq x \leq \frac{1}{2}} \frac{x^4}{4} - \frac{x^2}{2}$$

Définition 1.4.1 (Problème dual)**.** *Posons*

$$d(\lambda) = \min_{x \in \mathbb{R}^n} L(x, \lambda)$$

Le problème

$$(\mathcal{D}) \qquad \max_{\lambda \in \mathbb{R}^l_+} d(\lambda)$$

est appelé problème dual de $(\mathcal{P})$.

Remarque : Le problème dual - quand il est connu - est souvent plus simple à résoudre numériquement que le problème primal (c.f. si dessous).

Preuve du théorème. 1. Rappelons que l'inégalité ci-dessous est toujours vraie :

Lemme 1.4.2. *Soit L une fonction de deux variables x et λ. Alors*

$$\inf_{x\in X}\sup_{\lambda\in\Lambda} L(x,\lambda) \geq \sup_{\lambda\in\Lambda}\inf_{x\in X} L(x,\lambda)$$

(c.f. exercice **??**)

2. Comme le problème $\mathcal{P}$ admet au moins une solution x^*, et que, par hypothèse la contrainte est qualifiée en tout point, il existe $\lambda^* \in \mathbb{R}^l_+$ tel que

$$\begin{cases} i) & (\lambda^*)^T g(x^*) = 0 \quad \text{condition d'exclusion} \\ ii) & \nabla_x L(x^*,\lambda^*) = 0 \quad \text{condition nécessaire} \end{cases}$$

D'autre part, les fonctions f, $g_1,\ldots,g_l$ étant convexes et les λ_i positifs, le lagrangien du problème est convexe. Donc la condition nécessaire $\nabla_x L(x^*,\lambda^*) = 0$ est suffisante au sens où x^* est un minimum de $L(\cdot,\lambda^*)$. On en déduit que

$$\sup_{\lambda\in\mathbb{R}^l_+}\inf_{x\in\mathbb{R}^n} L(x,\lambda) \geq \inf_{x\in\mathbb{R}^n} L(x,\lambda^*) = L(x^*,\lambda^*)$$

3. Or

$$L(x^*,\lambda^*) = f(x^*) + \sum_{i=1}^{l} \lambda_i^* g_i(x^*) = f(x^*)$$

grâce aux conditions d'exclusion. Donc

$$L(x^*,\lambda^*) = \min_{x\in K} f(x) \;\leq\; \sup_{\lambda\in\mathbb{R}^l_+}\inf_{x\in\mathbb{R}^n} L(x,\lambda)$$

4. Montrons maintenant que

$$\sup_{\lambda\in\mathbb{R}^l_+} L(x,\lambda) \;=\; \begin{cases} f(x) & \text{si } x\in K \\ +\infty & \text{sinon} \end{cases}$$

En effet, si $x \in K$

$$f(x) = L(x,0) \;\leq\; \sup_{\lambda\in\mathbb{R}^l_+} L(x,\lambda) = f(x) + \sum_i \lambda_i g_i(x) \;\leq\; f(x)$$

car, pour tout i, $\lambda_i g_i(x) \leq 0$. Donc on optimise l'expression en prenant $\lambda_i = 0$. Si $x \notin K$, il existe i_0 tel que $g_{i_0}(x) > 0$. Alors, on peut construire la suite $(\lambda^{(k)})_k$ de $\mathbb{R}^l$ définie par

$$\lambda_i^{(k)} = \begin{cases} 0 & \text{si } i \neq i_0 \\ k & \text{sinon} \end{cases}$$

On a

$$\sup_{\lambda\in\mathbb{R}^l_+} L(x,\lambda) \geq \sup_k L(x,\lambda^k) = \sup_k (f(x) + k g_{i_0}(x)) = +\infty$$

5. Donc finalement

$$\begin{array}{rcl} \inf_{x\in\mathbb{R}^n}\sup_{\lambda\in\mathbb{R}^l_+} L(x,\lambda) & = & \inf_{x\in K}\sup_{\lambda\in\mathbb{R}^l_+} L(x,\lambda) \\ & = & \inf_{x\in K} f(x) \\ & = & L(x^*,\lambda^*) \\ & \leq & \sup_{\lambda\in\mathbb{R}^l_+}\inf_{x\in\mathbb{R}^n} L(x,\lambda) \end{array}$$

Or l'inégalité inverse est donnée par le lemme 1.4.2. D'où le resultat.

□

1.5 Méthodes numériques

Dans cette partie, nous discutons rapidement de quelques méthodes de résolution numérique pour les problèmes d'optimisation sous contraintes. Soulignons que nous nous intéressons ici à des problèmes non linéaires et sans structure particulière. En particulier, cette partie ne survole que rapidement *l'algorithme du simplexe*, qui permet de traiter très efficacement le cas de critères affines avec contraintes affines.

1.5.1 Projection sur un ensemble convexe fermé

Soit K un fermé de $\mathbb{R}^n$. On appelle projection d'un point $y \notin K$ sur K tout point $x \in K$ réalisant le minimum du problème

$$\min_{x \in K} \|x - y\|^2$$

Proposition 1.5.1. *Si K est un convexe fermé, alors il existe une seule projection $\Pi_K(y)$ de y sur K. Le point $\Pi_K(y)$ est le seul élément de K satisfaisant l'inégalité*

$$\forall z \in K,\ \langle y - \Pi_K(y), z - \Pi_K(y)\rangle \ \leq\ 0\ .$$

De plus, l'application $y \to \Pi_K(y)$ est contractante au sens où

$$\forall (y_1, y_2) \in \mathbb{R}^n \times \mathbb{R}^n,\ \|\Pi_K(y_1) - \Pi_K(y_2)\| \leq \|y_1 - y_2\|$$

Preuve. Comme la fonction $x \to \|x - y\|^2$ est strictement convexe, le minimum est unique. Montrons que $\Pi_K(y)$ vérifie la relation énoncée. En effet, pour tout $z \in K$, le point $tz + (1-t)\Pi_K(y)$ appartient à K si $t \in [0,1]$. Donc

$$\|y - \Pi_K(y)\|^2 \leq \|y - (tz + (1-t)\Pi_K(y))\|^2 = \|y - \Pi_K(y)\|^2 - 2t\langle y - \Pi_K(y), z - \Pi_K(y)\rangle + t^2\|z - \Pi_K(y)\|^2$$

Si on retranche $\|y - \Pi_K(y)\|^2$ des deux côtés, que l'on divise par $t > 0$ et que l'on fait tendre t vers 0^+, on obtient l'inégalité désirée :

$$\langle y - \Pi_K(y), z - \Pi_K(y)\rangle \leq 0$$

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

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

alors $x = \Pi_K(y)$. En effet

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

Finalement, vérifions que l'opérateur Π_K est contractant. On a

$$\langle y_1 - \Pi_K(y_1), \Pi_K(y_2) - \Pi_K(y_1)\rangle \leq 0$$

et

$$\langle y_2 - \Pi_K(y_2), \Pi_K(y_1) - \Pi_K(y_2)\rangle \leq 0$$

On additionne pour obtenir

$$\langle y_1 - y_2 - (\Pi_K(y_1) - \Pi_K(y_2)), \Pi_K(y_2) - \Pi_K(y_1)\rangle \leq 0$$

c'est-à-dire

$$\|\Pi_K(y_2) - \Pi_K(y_1)\|^2 \ \leq\ \langle y_1 - y_2; \Pi_K(y_1) - \Pi_K(y_2)\rangle \ \leq\ \|y_1 - y_2\|\|\Pi_K(y_1) - \Pi_K(y_2)\|$$

En divisant par $\|\Pi_K(y_2) - \Pi_K(y_1)\|$ on obtient l'inégalité désirée. □

1.5.2 Le gradient projeté

Soient K un convexe fermé de $\mathbb{R}^n$ et $f : \mathbb{R}^n \to \mathbb{R}$ une application convexe de classe $\mathcal{C}^1$. L'algorithme du gradient projeté a la structure suivante : Algorithme du gradient projeté :

- *on se donne un paramètre strictement positif ρ et on initialise avec un $x^0 \in K$*
- *à l'étape k, on remet à jour x^k en posant $x^{k+1} = \Pi_K(x^k - \rho\nabla f(x^k))$*

Théorème 1.5.1. *On suppose que f est une fonction elliptique, avec pour constante d'ellipticité $\alpha > 0$ et que ∇f est $M-$Lipschitzienne. On suppose que $\rho \in]0, \frac{2\alpha}{M^2}[$. Alors l'algorithme du gradient projeté converge, au sens où x_k converge vers la solution x^* du problème.*

Remarques :

1. Sous les hypothèses du théorème, le minimum est unique.

2. L'algorithme du gradient projeté est difficile à mettre en oeuvre dans le cas général, car le calcul de la projection est souvent aussi difficile que le problème initial. Cependant, cet algorithme peut être utilisé pour des contraintes "simples", de la forme $x \leq 0$.

Preuve. Soit x^* l'unique minimum du problème. Le lemme suivant est le point clé de la démonstration. Nous en donnons la preuve plus loin.

Lemme 1.5.1. *Sous les hypothèses du théorème, on a, pour tout $y \in K$,*

$$\langle \nabla f(y) - \nabla f(x^*), y - x^* \rangle \ \geq \ \alpha \|y - x\|^2 \ .$$

De plus,

$$\Pi_K(x^* - \rho \nabla f(x^*)) = x^* \ .$$

On a par définition de x_{k+1},

$$x^{k+1} - x^* = \Pi_K(x^k - \rho \nabla f(x^k)) - x^*$$

Or l'application Π_K est contractante, et $x^* = \Pi_K(x^* - \rho \nabla f(x^*))$ d'après le lemme. Donc

$$\begin{array}{rcl} \|x^{k+1} - x^*\|^2 & \leq & \|(x^k - \rho \nabla f(x_k)) - (x^* - \rho \nabla f(x^*))\|^2 \\ & \leq & \|x^k - x^*\|^2 - 2\rho \langle x^k - x^*, \nabla f(x_k) - \nabla f(x^*) \rangle + \rho^2 \|\nabla f(x^k) - \nabla f(x^*)\|^2 \\ & \leq & (1 - 2\alpha\rho + M^2\rho^2) \|x^k - x^*\|^2 \end{array}$$

car $\|\nabla f(x^k) - \nabla f(x^*)\| \leq M \|x^k - x^*\|$ et f est $\alpha-$elliptique. Or, par hypothèse, $-2\alpha\rho + M^2\rho^2$ est strictement compris entre 0 et 1. D'où

$$\|x^k - x^*\|^2 \leq (1 - 2\alpha\rho + M^2\rho^2)^k \|x^0 - x^*\|^2$$

avec $(1 - 2\alpha\rho + M\rho^2)^k$ qui tend vers 0. Donc (x^k) converge vers x^*. **Preuve du lemme :** Soit

$$\phi(t) = \langle \nabla f(ty + (1-t)x^*), y - x^* \rangle$$

Alors

$$\phi'(t) = \langle Hess_f(ty + (1-t)x^*)(y - x^*), (y - x^*) \rangle \geq \alpha \|y - x^*\|^2$$

par hypothèse d'ellipticité. Donc, en intégrant entre 0 et 1,

$$\phi(1) - \phi(0) \geq \alpha \|y - x^*\|^2$$

c'est-à-dire,

$$\langle \nabla f(y) - \nabla f(x^*), y - x^* \rangle \geq \alpha \|y - x^*\|^2$$

On a démontré l'inégalité désirée. Comme x^* est un minimum de f sur K, on a, pour tout $y \in K$ et tout $t \in [0, 1]$:

$$f((1-t)x^* + ty) - f(x^*) \geq 0$$

Divisant cette inégalité par t et en faisant tendre t vers 0 donne

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

Donc

$$\forall y \in K \ , \ \langle (x^* - \rho \nabla f(x^*)) - x^*, y - x^* \rangle \ \leq \ 0$$

et la proposition 1.5.1 implique que $x^* - \rho \nabla f(x^*)$ a pour projection x^* sur K. □

1.5.3 Algorithme d'Uzawa : contraintes d'égalité affines

On cherche à résoudre le problème suivant :

$$\min_{Cx = d} f(x)$$

où f est une fonction convexe de classe $\mathcal{C}^1$, C est une matrice $m \times n$ et d est un vecteur de $\mathbb{R}^m$. On sait que le problème dual est

$$(\mathcal{D}) \quad \max_{\lambda \in \mathbb{R}^m} d(\lambda)$$

où

$$d(\lambda) = \min_{x\in\mathbb{R}^n} f(x) + \lambda^T(Cx - d)$$

L'idée de l'algorithme d'Uzawa est d'appliquer (au moins formellement) l'algorithme du gradient à la fonction d. Il faut donc savoir calculer ∇d. Il se trouve que d n'est pas toujours dérivable. Cependant, lorsque d est dérivable, on peut calculer explicitement sa dérivée :

Lemme 1.5.2. *Supposons que d soit différentiable en un point λ_0, et que x_0 soit un minimum du problème*

$$\min_{x\in\mathbb{R}^n} f(x) + \lambda_0^T(Cx - d)\ .$$

Alors

$$\nabla d(\lambda_0) = Cx_0 - d\ .$$

Remarques :

1. Il y a, en général, un lien très fort entre l'unicité du minimum du problème $\min_x L(x, \lambda_0)$, et la différentiabilité de la fonction d en λ_0.
2. Mentionnons par exemple que, si la fonction f est elliptique, alors la fonction d est différentiable en tout point.

Preuve. En effet, pour tout $v \in \mathbb{R}^l$, pour tout $h \in \mathbb{R}$, on a

$$d(\lambda_0 + hv) \leq L(x_0, \lambda_0 + hv) = f(x_0) + \lambda_0^T(Cx_0 - d) + hv^T(Cx_0 - d)$$

Donc

$$d(\lambda_0 + hv) - d(\lambda_0) \geq hv^T(Cx_0 - d)$$

Si on divise cette égalité par $h > 0$, et que l'on fait tendre h vers 0^+, on obtient (puisqu'on a supposé d différentiable)

$$\langle \nabla d(\lambda_0), v\rangle \ \geq\ v^T(Cx_0 - d)$$

tandis que, si l'on fait la même opération avec $h < 0$, et que l'on fait tendre h vers 0^-, on obtient

$$\langle \nabla d(\lambda_0), v\rangle \ \leq\ v^T(Cx_0 - d)$$

Donc $\langle \nabla d(\lambda_0), v\rangle = v^T(Cx_0 - d)$. Cette égalité étant vraie pour tout vecteur v, on a bien prouvé que

$$\nabla d(\lambda_0) = Cx_0 - d\ .$$

□

L'algorithme d'Uzawa n'est alors rien d'autre que l'algorithme du gradient appliqué au problème dual :
ALGORITHME D'UZAWA :

- *on se donne un paramètre strictement positif ρ et on initialise avec un $\lambda^0 \in \mathbb{R}^m$*
- *à l'étape k, (i) on résoud le problème (sans contrainte)*

$$\min_{x\in\mathbb{R}^n} f(x) + (\lambda^k)^T(Cx - d)$$

Soit x^k une solution de ce problème, (ii) on remet à jour λ^k en posant $\lambda^{k+1} = \lambda^k + \rho(Cx^k - d)$

Théorème 1.5.2. *On suppose que f est une fonction elliptique, avec pour constante d'ellipticité $\alpha > 0$. On suppose que $\rho \in]0, \frac{2\alpha}{\|C\|^2}[$. Alors l'algorithme d'Uzawa converge au sens où x^k converge vers le minimum x^* du problème. Si, de plus, la matrice C est de rang m, alors λ^k converge vers le multiplicateur associé λ^*.*

Remarque : Sous les hypothèses du théorème, le minimum est unique.

Preuve. Soit x^* la solution du problème, et λ^* un multiplicateur associé. On a donc

$$\left\{ \begin{array}{ll} (i) & Cx^* = d \\ (ii) & \nabla f(x^*) + C^T\lambda^* = 0 \end{array} \right.$$

Les conditions d'optimalité satisfaites par x^k sont

$$\nabla f(x^k) + C^T\lambda^k = 0$$

En soustrayant (ii) à l'égalité précédente, on obtient

$$\nabla f(x^k) - \nabla f(x^*) + C^T(\lambda^k - \lambda^*) = 0$$

D'autre part, par définition de λ^{k+1}, on a

$$\begin{array}{rcl} \lambda^{k+1} - \lambda^* & = & \lambda^k - \lambda^* + \rho(Cx^k - d) \\ & = & \lambda^k - \lambda^* + \rho C(x^k - x^*) \end{array}$$

grâce à (i). En prenant le carré de la norme des deux membres de cette dernière égalité, on obtient

$$\begin{array}{rcl} \|\lambda^{k+1} - \lambda^*\|^2 & = & \|\lambda^k - \lambda^*\|^2 + 2\rho\langle\lambda^k - \lambda^*, C(x^k - x^*)\rangle + \rho^2\|C(x^k - x^*)\|^2 \\ & = & \|\lambda^k - \lambda^*\|^2 + 2\rho\langle C^T(\lambda^k - \lambda^*), x^k - x^*\rangle + \rho^2\|C(x^k - x^*)\|^2 \\ & = & \|\lambda^k - \lambda^*\|^2 - 2\rho\langle\nabla f(x^k) - \nabla f(x^*), x^k - x^*\rangle + \rho^2\|C(x^k - x^*)\|^2 \\ & \leq & \|\lambda^k - \lambda^*\|^2 - \rho(2\alpha - \rho\|C\|^2)\|x^k - x^*\|^2 \end{array}$$

On a utilisé successivement le fait que

$$\nabla f(x^k) - \nabla f(x^*) = -C^T(\lambda^k - \lambda^*) \ ,$$

que f est elliptique, et que

$$\|C(x^k - x^*)\|^2 \leq \|C\|^2\|x^k - x^*\|^2$$

Comme, par hypothèse, $2\alpha - \rho\|C\|^2$ est strictement positif, on en déduit que la suite $\|\lambda^k - \lambda^*\|^2$ est décroissante. Puisqu'elle est minorée, elle est convergente. En particulier

$$\|\lambda^k - \lambda^*\|^2 - \|\lambda^{k+1} - \lambda^*\|^2$$

tend vers 0. Or on a

$$\rho(2\alpha - \rho\|C\|^2)\|x^k - x^*\|^2 \leq \|\lambda^k - \lambda^*\|^2 - \|\lambda^{k+1} - \lambda^*\|^2$$

Comme $\rho(2\alpha - \rho\|C\|^2)$ est strictement positif, on peut conclure à la convergence de (x^k) vers x^*. Montrons enfin que, si C est de rang m, alors λ^k converge vers λ^*. Comme $\|\lambda^k - \lambda^*\|^2$ est convergente, on peut en déduire que la suite (λ^k) est bornée. Considérons une sous-suite $(\lambda^{k'})$ convergente vers une limite λ. Comme

$$\nabla f(x^k) + C^T\lambda^k = 0$$

on obtient, en passant à la limite

$$\nabla f(x^*) + C^T\lambda = 0 \tag{1.3}$$

Or C étant de rang m (c'est-à-dire surjective), on en déduit que C^T est injective. Donc l'équation (1.3) a une unique solution, λ^*. Donc $\lambda = \lambda^*$. En conclusion, toute sous-suite convergente de la suite bornée (λ^k) converge vers λ^*. On peut donc conclure que (λ^k) converge vers λ^*. □

1.5.4 Algorithme d'Uzawa : contraintes d'inégalité affines

On cherche maintenant à résoudre le problème suivant :

$$\min_{Cx \leq d} f(x)$$

où f est une fonction convexe de classe $\mathcal{C}^1$ et C est une matrice $l \times n$, d est un vecteur de $\mathbb{R}^l$ et l'inégalité $Cx \leq d$ signifie que toute composante du vecteur Cx est inférieure ou égale à la composante correspondante du vecteur d :

$$\forall i \in \{1, \ldots, l\}, \ (Cx)_i = d_i \ .$$

Le problème dual associé est

$$(\mathcal{D}) \quad \max_{\lambda \in \mathbb{R}^l_+} d(\lambda)$$

où

$$d(\lambda) = \min_{x \in \mathbb{R}^n} f(x) + \lambda^T (Cx - d)$$

Bien noter que, maintenant, on a à résoudre un problème avec contrainte $\lambda \geq 0$. Nous avons vu précédemment que l'algorithme d'Uzawa, pour les contraintes d'égalité affines, correspondait à un algorithme du gradient sur le problème dual. Comme, ici, le problème dual est un problème avec contraintes, il faudra appliquer un algorithme de gradient projeté. Appelons Π la projection sur l'orthant positif $\mathbb{R}^l_+$:

$$\Pi(\lambda)_i = \begin{cases} 0 & \text{si } \lambda_i < 0 \\ \lambda_i & \text{sinon} \end{cases}$$

L'algorithme d'Uzawa devient alors : Algorithme d'Uzawa :

- *on se donne un paramètre positif ρ et on initialise avec un $\lambda^0 \in \mathbb{R}^l_+$*
- *à l'étape k, (i) on résoud le problème (sans contrainte)*

$$\min_{x \in \mathbb{R}^n} f(x) + (\lambda^k)^T (Cx - d)$$

Soit x^k une solution de ce problème, (ii) on remet à jour λ^k en posant $\lambda^{k+1} = \Pi(\lambda^k + \rho(Cx^k - d))$

On a le même résultat de convergence que précédemment :

Théorème 1.5.3. *On suppose que f est une fonction elliptique, avec pour constante d'ellipticité $\alpha > 0$. On suppose que $\rho \in]0, \frac{2\alpha}{\|C\|^2}[$. Alors l'algorithme d'Uzawa converge au sens où x^k converge vers le minimum x^* du problème. Si, de plus, la matrice C est de rang m, alors λ^k converge vers le multiplicateur associé λ^*.*

Remarques :

- Sous les hypothèses du théorème, le minimum est unique.
- La démonstration de ce résultat est identique à celle du théorème 1.5.2. En effet, l'opérateur de projection vérifie l'inégalité
$$\|\Pi(\lambda_1) - \Pi(\lambda_2)\| \leq \|\lambda_1 - \lambda_2\|$$
pour tout $(\lambda_1, \lambda_2) \in \mathbb{R}^l \times \mathbb{R}^l$.

1.5.5 Programmation linéaire

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.

Définition 1.5.1. *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

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

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

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

Théorème 1.5.4. *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 $\{1, \dots, n\}$ 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. 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

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

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 1.5.5. *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 1.5.4 $x^{\delta,+}$ et $x^{\delta,-}$ par

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

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 $la, x^{\delta,+} - x\rangle = -\langle a, x^{\delta,-} - x\rangle = 0$. 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 $x^{\delta,+}_j = 0$ ou $x^{\delta,-}_j = 0$, 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 1.5.5 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.

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

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

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 $x^{k,\delta}$:

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

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

Proposition 1.5.2. *Une et une seule des trois alternatives suivantes a lieu :*

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

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

(C) Soit il existe au moins un indice k tel que $a_k - \sum_{j\in J} \alpha_j^k a_j < 0$ et au moins un indice $\alpha_j^k > 0$ pour tout $j \in J$. Alors on prendra le plus grand réel $\delta > 0$ tel que $x^{k,\delta}$ est encore dans K. Alors il existe $j \in J^*(x)$ tel que $x_j - \delta\alpha_j^k = 0$. On remarque facilement que $x^{k,\delta}$ est alors un sommet, et que la nouvelle base associée à $x^{k,\delta}$ est $(C_j)_{j\in J^*(x)\cup\{k\})\setminus\{j\}}$.

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. □

1.5.6 Autres méthodes

On se contente de citer ici quelques méthodes numériques. Pour les détails, les preuves de convergence, etc.., voir par exemple les ouvrages de Minoux ou de Culioli.

Méthodes de pénalité

– **Les méthodes de pénalisations exterieures** concernent des problèmes très généraux de la forme

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

avec

$$K = \{x \in R^n,\ g_i(x) \leq 0,\ i \in I,\ h_j(x) = 0,\ j \in J\}$$

où $I = \{1, \ldots, l\}$ indexe les contraintes d'inégalité et $J = \{1, \ldots, m\}$ indexe les contraintes d'égalité. Les fonctions g_i et h_i sont toutes supposées de classe $\mathcal{C}^1$ de $\mathbb{R}^n$ dans $\mathbb{R}$. On approxime (par exemple) ce problème par le problème sans contrainte

$$(\mathcal{P}_\epsilon) \quad \min_{x\in\mathbb{R}^n}\ f(x) + \frac{1}{\epsilon}\left[\sum_{i=1}^{l}[[g_i(x)]_+]^2 + \sum_{j=1}^{m}[h_j(x)]^2\right]$$

où $[t]_+ = \max\{0, t\}$. La pénalisation est dite extérieure car on s'attend à ce que l'optimum $\bar{x}_\epsilon$ du problème $(\mathcal{P}_\epsilon)$ ne vérifie pas la contrainte $\bar{x}_\epsilon \in K$.

– **Les méthodes de pénalisation intérieure** ne concernent que des problèmes avec contraintes d'inégalité, c'est-à-dire que $J = \emptyset$. On définit

$$\Omega = \{x \in \mathbb{R}^n \mid g_i(x) < 0,\ i \in I\}$$

Le problème pénalisé prend alors (par exemple) la forme suivante

$$(\mathcal{P}_\epsilon) \quad \min_{x\in\Omega} f(x) + \epsilon\left[\sum_{i=1}^{l}\log g_i(x)\right]$$

Notez que s'il existe un optimum au problème pénalisé, alors celui-ci appartient à l'intérieur de la contrainte K.

Ces méthodes sont difficiles à mettre en oeuvre pour obtenir des résultats précis, car, quand $\epsilon > 0$ est "petit", la pénalisation rend les algorithmes de recherche très instables. Cependant, on peut se servir de ces méthodes pour obtenir un résultat approché à partir duquel on peut démarrer un algorithme plus fin.

Méthode de Frank et Wolfe

Elle concerne des problèmes de la forme

$$(\mathcal{P}) \quad \min_{Ax=b,\ x\geq 0} f(x)$$

où f est différentiable, mais pas nécessairement convexe. On remplace le problème non linéaire $(\mathcal{P})$ par une suite de problèmes de programmation linéaire. Supposons construits k points $x_1, \ldots, x_k$ (les k premières itérations de l'algorithme). Soit y_k un minimum du problème

$$(\mathcal{P}_k) \quad \min_{Ay=b,\ y\geq 0} \langle \nabla f(x_j), y\rangle$$

On choisit x_{k+1} de façon à minimiser f sur le segment $[x_k, y_k]$. Si on suppose que f est coercive, ou que la contrainte est bornée, alors on peut montrer que la suite x_k converge vers un point vérifiant les conditions nécessaires d'optimalité du problème $(\mathcal{P})$.

Approximation du système de Kuhn & Tucker

Une méthode d'approximation possible est de chercher les points vérifiant le système de conditions nécessaires de Kuhn & Tucker. Voir par exemple [Minoux, p. 219]

1.6 Conditions du second ordre

1.6.1 Une condition nécessaire du second ordre

On suppose que K est de la forme :

$$K = \{x \in \mathbb{R}^n \mid \forall i \in I,\ g_i(x) \leq 0 \text{ et } \forall j \in J,\ h_j(x) = 0\}\ ,$$

avec maintenant g_i et h_j de classe $\mathcal{C}^2$.

Posons

$$\forall x \in K,\ C(x) = \{v \in T_K(x) \mid \langle \nabla f(x), v\rangle = 0\}\ .$$

Remarquons que, si la contrainte est qualifiée au point x, l'ensemble $C(x)$ a pour expression :

$$\forall x \in K,\ C(x) = \left\{ v \in T_K(x) \mid \begin{array}{l} \forall i \in I(x),\ \langle \nabla g_i(x), v\rangle \leq 0, \\ \forall j \in J,\ \langle \nabla h_j(x), v\rangle = 0 \text{ et } \langle \nabla f(x), v\rangle = 0 \end{array} \right\}\ .$$

Pour tout x de K, on pose aussi

$$\Lambda(x) = \left\{ (\lambda, \mu) \in \mathbb{R}^l \times \mathbb{R}^m \mid \begin{array}{l} \lambda \geq 0 \text{ et } \frac{\partial}{\partial x} L(x, \lambda, \mu) = 0, \\ \forall i \notin I(x),\ \lambda_i = 0 \end{array} \right\}\ .$$

Notons que $\Lambda(x)$ est non vide si et seulement si x vérifie les conditions nécessaires d'optimalité du premier ordre. Dans ce cas, $\Lambda(x)$ est l'ensemble de tous les multiplicateurs de Lagrange associés au point x.

Théorème 1.6.1 (Conditions nécessaires d'ordre 2)**.** *Si x est un minimum local de f sur K et si K est qualifiée en x, alors*

$$\forall v \in C(x),\ \max_{(\lambda,\mu)\in\Lambda(x)} \langle \frac{\partial^2 L}{\partial x^2}(x, \lambda, \mu)v, v\rangle\ \geq\ 0\ .$$

Remarque : Il n'est pas vrai en général que la matrice symmétrique $\frac{\partial^2 L}{\partial x^2}(x, \lambda, \mu)$ soit positive même lorsque les multiplicateurs sont uniques. Par exemple, on montre facilement que la fonction (dans $\mathbb{R}^2$) $f(x, y) = -xy$ admet un minimum global sur $K = \{(x, y) \in \mathbb{R}^2 \mid x + y \leq 2,\ x \geq 0, y \geq 0\}$ au point $(1, 1)$. Les multiplicateurs sont ici uniques, et égaux à $(1, 0, 0)$. Or $L(x, y, \lambda_1, \lambda_2, \lambda_3) = -xy + \lambda_1(x + y - 1) - \lambda_2 - \lambda_3 y$, et donc

$$\frac{\partial^2 L}{\partial (x, y)^2}(1, 1, 1, 0, 0) = -\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\ .$$

Cependant, on a $C(1, 1) = \{(v_1, v_2) \in \mathbb{R}^2 \mid v_1 + v_2 = 0\}$, et la restriction de la forme quadratique associée à la matrice $\frac{\partial^2 L}{\partial (x,y)^2}(1, 1, 1, 0, 0)$ au sous-espace vectoriel $C(1, 1)$ est positive. C'est précisément ce que dit le théorème.

1.6.2 Preuve de la condition nécessaire d'ordre 2

La preuve du théorème 1.6.1 est assez longue et nécessite plusieurs étapes. Elle suit d'assez près la démonstration des conditions du premier ordre.

Introduisons d'abord la notion de cône tangent du second ordre :

$$\forall x \in \mathbb{R}^n, \forall v \in T_K(x),\ T_K^2(x, v) = \{w \in \mathbb{R}^n \mid \exists t_k \to 0^+,\ \exists w_k \to v \text{ avec } x + t_k v + t_k^2 w_k \in K\}\ .$$

On affirme que

Lemme 1.6.1 (Condition d'Euler du second ordre)**.** *Supposons que f possède un minimum sur K au point x. Alors*

$$\forall v \in C(x),\ \forall w \in T_K^2(x,v),\ \langle \nabla f(x), w\rangle + \frac{1}{2}\langle Hess_f(x)v, v\rangle \geq 0\ .$$

Preuve. En effet, par définition, pour tout $v \in C(x)$ et $w \in T_K^2(x,v)$, il existe une suite $t_k \to 0^+$ et une suite $w_k \to w$ telles que $x + t_k v + t_k^2 w_k \in K$. Donc

$$f(x) \leq f(x + t_k v + t_k^2 w_k) = f(x) + t_k\langle \nabla f(x), v\rangle + t_k^2(\langle \nabla f(x), w\rangle + \frac{1}{2}\langle Hess_f(x)v, v\rangle) + o(t_k^2)\ .$$

Or $\langle \nabla f(x), v\rangle = 0$ car $v \in C(x)$. Donc l'inégalité ci-dessus implique que

$$t_k^2(\langle \nabla f(x), w\rangle + \frac{1}{2}\langle Hess_f(x)v, v\rangle) + o(t_k^2) \geq 0\ .$$

On divise alors cette inégalité par t_k^2, et on fait tendre $k \to +\infty$ pour obtenir le résultat désiré. □

Comme dans le cas du théorème de Kuhn et Tucker, pour exploiter la condition d'Euler, il faut être capable de calculer l'ensemble $T_K^2(x)$. C'est l'objet du lemme suivant, que l'on admettra :

Lemme 1.6.2. *Posons, pour tout $v \in T_K(x)$,*

$$I^2(x,v) = \{i \in \{1,\dots,l\} \mid g_i(x) = 0 \text{ et } \langle \nabla g_i(x), v\rangle = 0\}\ .$$

Si la contrainte K est qualifiée en x, alors

$$T_K^2(x,v) = \left\{ w \in \mathbb{R}^n \mid \begin{array}{l} \langle \nabla g_i(x), w\rangle + \frac{1}{2}\langle Hess_{g_i}(x)v, v\rangle \leq 0\ \forall i \in I^2(x,v) \\ \langle \nabla h_j(x), w\rangle + \frac{1}{2}\langle Hess_{h_j}(x)v, v\rangle = 0\ \forall j \in J \end{array} \right\}$$

Preuve du théorème 1.6.1. Pour démontrer la condition nécessaire d'ordre 2, il reste à exploiter la condition d'Euler et la caractérisation de $T_K^2(x)$.

Pour cela, on considère le problème linéaire suivant :

$$\min_{w \in T_K^2(x,v)} F(w) \text{ où } F(w) = \langle \nabla f(x), w\rangle + \frac{1}{2}\langle Hess_f(x)v, v\rangle\ .$$

Comme, pour tout $w \in T_K^2(x)$, $F(w) \geq 0$ (cf. la première étape), une propriété bien connue des systèmes linéaires affirme que ce problème admet au moins une solution. Par conséquent, on peut utiliser les méthodes de dualité : notons $\mathcal{L}$ le lagrangien de ce problème :

$$\begin{array}{rl} \mathcal{L}(w,\lambda,\mu) = F(w) + \sum_{i\in I(x)} \lambda_i(\langle \nabla g_i(x), w\rangle + \frac{1}{2}\langle Hess_{g_i}(x)v, v\rangle) \\ \qquad + \sum_{j\in J} \mu_j(\langle \nabla h_j(x), w\rangle + \frac{1}{2}\langle Hess_{h_j}(x)v, v\rangle)\ . \end{array}$$

Notons que l'application $\mathcal{L}(\cdot,\lambda,\mu)$ est affine. Si elle admet un minimum sur $\mathbb{R}^n$, cela signifie donc qu'elle est constante et que

$$\frac{\partial \mathcal{L}}{\partial w}(\cdot,\lambda,\mu) = \nabla f(x) + \sum_{i\in I(x)} \lambda_i \nabla g_i(x) + \sum_{j\in J} \mu_j \nabla h_j(x) = 0\ .$$

Donc, si $\mathcal{L}(\cdot,\lambda,\mu)$ admet un minimum sur $\mathbb{R}^n$, alors $(\lambda,\mu) \in \Lambda(x)$ et $\mathcal{L}(w,\lambda,\mu)$, qui ne dépend pas de w, vaut

$$\begin{array}{rcl} \mathcal{L}(w,\lambda,\mu) & = & \frac{1}{2}\left[\langle Hess_f(x)v, v\rangle + \sum_{i\in I(x)} \lambda_i\langle Hess_{g_i}(x)v, v\rangle + \sum_{j\in J} \mu_j\langle Hess_{h_j}(x)v, v\rangle\right] \\ & = & \frac{1}{2}\langle \frac{\partial^2 L}{\partial x^2}(x,\lambda,\mu)v, v\rangle\ . \end{array}$$

Inversement, il est clair que si $(\lambda,\mu) \in \Lambda(x)$, alors $\mathcal{L}$ est constante.

Au contraire, si $(\lambda,\mu) \notin \Lambda(x)$, alors la fonction affine $\mathcal{L}(\cdot,\lambda,\mu)$ n'admet pas de minimum sur $\mathbb{R}^n$, et son infimum sur $\mathbb{R}^n$ est égal à $-\infty$.

Nous venons donc de montrer que

$$\delta(\lambda,\mu) = \inf_{w\in\mathbb{R}^n} \mathcal{L}(w,\lambda,\mu) = \begin{cases} \frac{1}{2}\langle \frac{\partial^2 L}{\partial x^2}(x,\lambda,\mu)v, v\rangle & \text{si } (\lambda,\mu)\in\Lambda(x) \\ -\infty & \text{sinon} \end{cases}$$

Le théorème de dualité affirme alors que

$$\max_{\lambda\geq 0,\ \mu} \delta(\lambda,\mu) = \min_{w\in T_K^2(x,v)} F(w) \geq 0 \ .$$

Or

$$\max_{\lambda\geq 0,\ \mu} \delta(\lambda,\mu) = \max_{(\lambda,\mu)\in\Lambda(x,v)} \frac{1}{2}\langle \frac{\partial^2 L}{\partial x^2}(x,\lambda,\mu)v, v\rangle \ .$$

Par conséquent, nous avons bien montré que

$$\max_{(\lambda,\mu)\in\Lambda(x,v)} \langle \frac{\partial^2 L}{\partial x^2}(x,\lambda,\mu)v, v\rangle \ \geq \ 0 \ .$$

□

1.6.3 Condition suffisante du second ordre

Comme pour les problèmes sans contraintes, il suffit de renforcer légèrement les conditions nécessaires d'ordre 2 pour obtenir une condition suffisante.

Théorème 1.6.2. *Soit x un point de K où la contrainte est qualifiée. On suppose que les conditions du premier ordre sont satisfaites en x : l'ensemble $\Lambda(x)$ n'est pas vide. Si de plus il existe une constante $\alpha > 0$ telle que*

$$\forall v \in C(x),\ \max_{(\lambda,\mu)\in\Lambda(x)} \langle \frac{\partial^2 L}{\partial x^2}(x,\lambda,\mu)v, v\rangle \ \geq \ \alpha\|v\|^2 \ ,$$

alors f possède un minimum local strict sur K en x.

Remarque : On rencontre fréquemment la condition suivante dans la littérature :
"s'il existe $(\lambda,\mu)\in\Lambda(x)$ tel que la matrice $\frac{\partial^2 L}{\partial x^2}(x,\lambda,\mu)$ est définie positive, alors f possède un minimum local sur K en x."
Cette condition, qui est beaucoup plus simple que celle donnée dans le théorème, est cependant très éloignée des conditions nécessaires du second ordre.

1.7 Optimisation de portefeuilles financiers

(Source : Quintard-Pinon, Théorie financière, Economica)

1.7.1 Description du modèle

On s'intéresse à une économie composée de n actifs financiers, caractérisés par leur taux de rendement aléatoire r_i. L'espérance du rendement de l'actif i est notée $\overline{r_i}$. La matrice de covariance $Q = (q_{ij})_n$ des variables aléatoires r_i est définie par

$$\forall (i,j)\in\{1,\dots,n\}^2,\ \ q_{ij} = \mathbb{E}((r_i - \overline{r_i})(r_j - \overline{r_j})) \ .$$

On rappelle que la matrice Q est symétrique et positive.

Un portefeuille w est une combinaison d'actifs, $w = \sum_i w_i r_i$ où les w_i sont des réels tels que $\sum_i w_i = 1$. Le réel w_i est la part de l'actif i dans le portefeuille w. En général, les w_i sont positifs. Cependant, on considèrera fréquemment le cas où w_i est négatif : cela correspondra à une vente à découvert sur l'actif i.

Le rendement moyen du portefeuille w est l'espérance $\overline{w} = \mathbb{E}(w)$ de cette variable aléatoire. Son expression est

$$\mathbb{E}(w) = \sum_{i=1}^{n} w_i \mathbb{E}(r_i) = \sum_{i=1}^{n} w_i \overline{r_i} \ .$$

Le risque est modélisé par la variance du portefeuille,

$$Var(w) = \mathbb{E}((w - \overline{w})^2) \ .$$

Un calcul rapide montre que

$$Var(w) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j q_{ij}$$

Une question générale en finance est d'optimiser le rendement moyen $\mathbb{E}(w)$ du portefeuille w pour un niveau de risque $Var(w)$ donné, ou de minimiser le risque $Var(w)$ pour un rendement moyen $\mathbb{E}(w)$ donné. Nous allons voir qu'en général, il existe plusieurs portefeuilles dits *efficients*, qui sont optimaux au sens où ils proposent un bon compromis entre le rendement moyen et le risque.

On appelle [1] *portefeuille efficient* un portefeuille w tel que tout portefeuille w' ayant un rendement moyen meilleur que w a un risque supérieur à celui de w : mathématiquement, cela s'écrit

$$w \text{ est efficient } \Leftrightarrow \text{ si } \mathbb{E}(w') \geq \mathbb{E}(w) \text{ alors } Var(w') \geq Var(w)$$

1.7.2 Formalisation du problème

Dans toute la suite, on note W et R les matrices colonnes de terme général $(w_i)_n$ et $(r_i)_n$. Le portefeuille w s'exprime alors matriciellement comme $w = W^T R = R^T W$. Le fait que $\sum_i w_i = 1$ s'écrit matriciellement comme

$$\sum_i w_i = 1 \ \Leftrightarrow \ W^T \mathbf{1} = 1$$

où $\mathbf{1}$ est la matrice colonne de $\mathbb{R}^n$ de coordonnées identiquement égales à 1.

En ces termes, l'espérance du portefeuille s'écrit

$$\mathbb{E}(w) = W^T \overline{R} = \overline{R}^T W \ ,$$

où $\overline{R} = (\overline{r_i})$, et sa variance

$$Var(w) = W^T Q W = \langle W, QW \rangle \ .$$

Exercice 1.8. Montrer que chercher un portefeuille efficient de rendement moyen donné r revient à chercher une solution au problème

$$(\mathcal{P}_1) \qquad \begin{cases} \min & W^T Q W \\ W^T \mathbf{1} = 1 \\ W^T \overline{R} \geq r \end{cases}$$

si on autorise les ventes à découvert. Si l'on n'autorise pas ces ventes à découvert, montrer que le problème à résoudre devient

$$(\mathcal{P}_2) \qquad \begin{cases} \min & W^T Q W \\ W^T \mathbf{1} = 1 \\ W^T \overline{R} \geq r \\ W \geq 0 \end{cases}$$

Reste à déterminer

1. s'il existe des solutions au problème ci-dessus (sous quelles hypothèses ?)
2. comment caractériser ces solutions ?

1. Il existe plusieurs définitions de portefeuilles efficients. Elles sont toutes plus ou moins équivalentes.

3. comment les calculer ?

Exercice 1.9. Montrer si Q est définie positive et *si les contraintes sont non vides* alors les problèmes précédents $(\mathcal{P}_1)$ et $(\mathcal{P}_2)$ ont chacun une unique solution.

Exercice 1.10. Montrer que l'hypothèse Q définie positive implique l'absence de portefeuille sans risque, c'est-à-dire de variance nulle.

Dans toute la suite, on supposera que Q est définie positive.

1.7.3 Le portefeuille de variance minimale

On cherche d'abord le portefeuille de risque minimal (sans se soucier du rendement moyen). Cela revient à minimiser

$$(\mathcal{P}_3) \quad \min_{W^T\mathbf{1}=1} W^T Q W$$

Exercice 1.11. Montrer qu'il existe un unique portefeuille de risque minimal. Montrer que ce portefeuille est efficient, et qu'il n'existe pas de portefeuille efficient de rendement moyen plus faible.

Exercice 1.12. Ecrire les conditions nécessaires d'optimalité du problème $(\mathcal{P}_3)$. Déterminer la solution optimale.

Exercice 1.13. Montrer que le portefeuille de variance minimale est corrélé positivement avec tout autre portefeuille.

1.7.4 Caractérisation des portefeuilles efficients

Exercice 1.14. Quels sont les rendements moyens possibles d'un portefeuille efficient

1. quand tous les actifs ont même rendement moyen,
2. quand au moins deux des actifs ont des rendements moyens différents, et que l'on autorise les ventes à découvert.

Dans cette partie, afin de simplifier les calculs, on autorise les ventes à découvert.

Exercice 1.15. Ecrire les conditions nécessaires d'optimalité du problème $(\mathcal{P}_1)$. Expliquer pourquoi elles sont suffisantes. Montrer que, si le rendement moyen r est supérieur au rendement moyen du portefeuille de risque minimum, alors la contrainte $W^T\overline{R} \geq r$ est saturée.

Exercice 1.16. Calculer le portefeuille efficient de rendement moyen r en fonction des matrices Q, d et D où

$$d = \begin{pmatrix} 1 \\ r \end{pmatrix} \text{ et } D = \begin{pmatrix} \mathbf{1}^T \\ \overline{R}^T \end{pmatrix}$$

(on supposera la matrice $DQ^{-1}D^T$ inversible).

Exercice 1.17. Déduire de l'exercice précédent que, si l'on connait deux portefeuilles efficients w_1 et w_2, de rendement moyen r_1 et r_2 ($r_1 \neq r_2$), on peut calculer tous les portefeuilles efficients comme combinaison affine de ces portefeuilles. En d'autre termes, si w est un portefeuille efficient, il existe $\lambda \in \mathbb{R}$ tel que

$$w = \lambda w_1 + (1-\lambda) w_2 \ .$$

Exercice 1.18. Démontrer le théorème de Roll : Soit w un portefeuille de rendement moyen strictement supérieur à celui du portefeuille de risque minimal. Alors une condition nécessaire et suffisante pour que w soit efficient est que sa covariance avec tout autre portefeuille w' soit une fonction affine du rendement moyen du portefeuille w'.

Exercice 1.19. Montrer que si l'on dessine la courbe $(\sigma(w), \mathbb{E}(w))$ dans le plan (σ, r) lorsque w parcourt tous les portefeuilles efficients, on obtient une portion d'hyperbole. Cette courbe est appelée *frontière efficiente*. Commenter.

Exercice 1.20. On considère deux actifs risqués dont les taux de rendements aléatoires r_1 et r_2 possèdent comme matrice de covariance

$$Q = \begin{pmatrix} 0,15 & 0,1 \\ 0,1 & 0,3 \end{pmatrix}$$

et comme espérance

$$\overline{R} = \begin{pmatrix} 0,1 \\ 0,18 \end{pmatrix}$$

1. Calculer le portefeuille de risque minimal.
2. Calculer tous les portefeuilles efficients.
3. Calculer la variance $v(r)$ du portefeuille efficient de rendement moyen r. Dessiner la frontière efficiente.

Exercice 1.21. On considère trois actifs risqués dont les taux de rendements aléatoires r_1, r_2 et r_3 possèdent comme matrice de covariance

$$Q = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$$

et comme espérance

$$\overline{R} = \begin{pmatrix} 0.4 \\ 0.8 \\ 0.8 \end{pmatrix}$$

1. Calculer le portefeuille de risque minimal.
2. Montrer qu'un portefeuille de rendement moyen $0,8$ ne contient pas d'actif 1. Montrer que le portefeuille efficient de rendement moyen $0,8$ *n'est pas* composé d'un unique actif de rendement moyen $0,8$.
3. Calculer le portefeuille efficient de rendement moyen $0,8$.
4. En déduire tous les portefeuilles efficients.

Exercice 1.22. On considère n actifs risqués r_i, de même espérance $\overline{r}$, et de variance σ_i^2. On suppose que les actifs r_i sont décorrélés, c'est-à-dire que $q_{ij} = 0$ si $i \neq j$.

1. Montrer qu'il n'y a qu'un portefeuille efficient.
2. Le calculer. Pourquoi n'est-il pas constitué que de l'actif de risque minimum ?

1.7.5 Le problème avec un actif sans risque

On suppose qu'un des actifs, noté r_0, est sans risque, c'est-à-dire que

$$Var(r_0) = 0,\ Covar(r_0, r_i) = 0\ \forall i \in \{1, \ldots, n\}\ .$$

Noter que r_0 est une variable aléatoire constante (c'est-à-dire qu'elle n'est pas aléatoire).

Un portefeuille efficient de rendement moyen $\overline{r}$ est un portefeuille minimisant le problème

$$\begin{cases} \min & W^T Q W \\ w_0 + W^T \mathbf{1} = 1 \\ w_0 r_0 + W^T \overline{R} = r \end{cases}$$

où W est la matrice colonne des $(w_i)_{i=1,\ldots,n}$, et w_0 est la part d'actif sans risque dans le portefeuille.

On peut remarquer que, même si la matrice Q est définie positive, le problème est dégénéré car w_0 n'intervient pas dans le critère.

On suppose dans toute la suite que $r_0 < \overline{r_i}$ pour tout i.

Exercice 1.23. Montrer qu'il suffit de trouver un portefeuille w vérifiant les conditions nécessaires d'optimalité pour conclure que le problème admet une solution, et que w en particulier est une solution.

Exercice 1.24. Calculer le portefeuille efficient de rendement moyen r en fonction des matrices Q, d, d' et D où

$$d = \begin{pmatrix} 1 \\ r \end{pmatrix},\ d' = \begin{pmatrix} -r_0 \\ 1 \end{pmatrix} \text{ et } D = \begin{pmatrix} \mathbf{1}^T \\ \overline{R}^T \end{pmatrix}$$

(on supposera toujours la matrice $DQ^{-1}D^T$ inversible).

Exercice 1.25. On dessine la courbe $(\sigma(w), E(w))$ lorsque w parcourt l'ensemble des portefeuilles efficients. Montrer que cette courbe est une droite passant par $(0, r_0)$. Montrer que cette droite est tangente à la courbe tracée dans la partie précédente.

Le point de contact entre cette droite et la courbe de la partie précédente est appelé **portefeuille de marché**. La droite elle-même s'appelle **droite du marché des capitaux**.

Exercice 1.26. On se donne deux actifs risqué r_1 et r_2. Les matrices d'espérance et de covariance associées sont

$$Q = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}\ \overline{R} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

On suppose que l'actif sans risque a pour rendement $r_0 = 1$.

1. On fixe w_0. Trouver le portefeuille $w = w(r, w_0)$ de risque minimum et de rendement moyen r.
2. Trouver le portefeuille efficient $w(r)$ de rendement moyen r en minimisant par rapport à w_0 la variance du portefeuille $w(r, w_0)$.
3. Dessiner la droite du marché des capitaux.
4. Mêmes questions si l'on n'autorise pas les ventes à découvert.

1.8 Tableau des conditions nécessaires d'optimalité

Contraintes	**Conditions de qualification**	**Conditions nécessaires**
$h(x) = 0$	$\nabla h(x^*) \neq 0$	$\exists \lambda \in \mathbb{R}$ avec $\nabla f(x^*) + \lambda \nabla h(x^*) = 0$
$h_1(x) = 0, \ldots, h_m(x) = 0$	$\{\nabla h_1(x^*), \ldots, \nabla h_m(x^*)\}$ famille libre	$\exists (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$ avec $\nabla f(x^*) + \sum_j \lambda_j \nabla h_j(x^*) = 0$
$g(x) \leq 0$	$\nabla g(x^*) \neq 0$ si $g(x^*) = 0$	$\exists \lambda \in \mathbb{R}_+$ avec $\nabla f(x^*) + \lambda \nabla g(x^*) = 0$ et $\lambda g(x^*) = 0$
$g(x) \leq 0$, g convexe	$\exists x_0$ avec $g(x_0) < 0$	$\exists \lambda \in \mathbb{R}_+$ avec $\nabla f(x^*) + \lambda \nabla g(x^*) = 0$ et $\lambda g(x^*) = 0$
$g_1(x) \leq 0, \ldots, g_l(x) \leq 0$	$\exists v \in \mathbb{R}^n$ avec $\forall i \in I_0(x^*), \langle \nabla g_i(x^*), v \rangle < 0$	$\exists (\lambda_1, \ldots, \lambda_l) \in \mathbb{R}_+^l$ avec $\nabla f(x^*) + \sum_i \lambda_i \nabla g_i(x^*) = 0$ et $\sum_i \lambda_i g_i(x^*) = 0$
$g_1(x) \leq 0, \ldots, g_l(x) \leq 0$, g_i convexes	$\exists x_0 \in \mathbb{R}^n$ avec $\forall i \in I$, $g_i(x_0) < 0$	$\exists (\lambda_1, \ldots, \lambda_l) \in \mathbb{R}_+^l$ avec $\nabla f(x^*) + \sum_i \lambda_i \nabla g_i(x^*) = 0$ et $\sum_i \lambda_i g_i(x^*) = 0$
$g_1(x) \leq 0, \ldots, g_l(x) \leq 0$, $h_1(x) = 0, \ldots, h_m(x) = 0$	$\{\nabla h_1(x^*), \ldots, \nabla h_m(x^*)\}$ libre, $\exists v \in \mathbb{R}^m$ avec $\forall j \in J$, $\langle \nabla h_j(x^*), v \rangle = 0$ et $\forall i \in I_0(x^*)$, $\langle \nabla g_i(x^*), v \rangle < 0$	$\exists \lambda \in \mathbb{R}_+^l$, $\exists \mu \in \mathbb{R}^m$ avec $\nabla f(x^*) + \sum_i \lambda_i \nabla g_i(x^*) + \sum_j \mu_j \nabla h_j(x^*) = 0$ et $\sum_i \lambda_i g_i(x^*) = 0$
$g_1(x) \leq 0, \ldots, g_l(x) \leq 0$, $h_1(x) = 0, \ldots, h_m(x) = 0$ g_i et h_j affines		idem

Dans ce tableau, les fonctions f, g_i et h_j sont toutes des applications de classe $\mathcal{C}^1$ de $\mathbb{R}^n$ dans $\mathbb{R}$. On a noté

$$I = \{1, \ldots, l\}\ ,\ J = \{1, \ldots, m\} \text{ et } I_0(x) = \{i \in I \mid g_i(x) = 0\}$$

Chapitre 2

Programmation dynamique

Ce chapitre 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 d'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 "Optimisation", nous avons expliqué comment écrire les conditions nécessaires d'optimalité avec contraintes sur l'état. Nous verrons qu'il est également possible dans le cadre du calcul des variations et du contrôle optimal d'écrire des conditions nécessaires d'optimalité : 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.

2.1 Problèmes en temps discret

On considère un système dynamique discret

$$x_{n+1} = f_n(x_n, u_n), \qquad n = 0,\ 1,\ 2, \dots,\ N-1, \qquad x_0 = \bar{x}$$

où les indices $n \in \mathbb{N}$ désignent les instants (discrets), l'instant final $N \in \mathbb{N}^*$ étant *l'horizon* du problème, x_n est *l'état du système* à l'instant n, 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 (i.e., $x_n \in X$ pour tout n), le contrôle à l'instant n dans un ensemble U_n ($u_n \in U_n$ pour tout n) et $f_n : X \times U_n \to X$ pour tout n. La condition initiale $\bar{x} \in X$ du système est fixée.

En général, dans un problème de contrôle optimal discret, on cherche à minimiser un coût

$$\min_{(u_n)} \sum_{n=0}^{N-1} L_n(x_n, u_n) + g(x_N)$$

sur tous les choix possibles des paramètres $u_0, \dots, u_{N-1}$. La quantité $L_n(x_n, u_n)$ est le *coût courant* à l'instant n, g étant le *coût terminal* : $L_n : X \times U_n \to \mathbb{R}$, $g : X \to \mathbb{R}$.

On peut aussi parfois avoir affaire à des problèmes en horizon infini (c'est-à-dire $N = +\infty$). Il n'y a alors pas de paiement terminal et le problème est généralement escompté par un taux d'escompte $r \in]0, 1[$ (dont la signification est qu'un euro aujourd'hui vaut r euros demain) :

$$\min_{(u_n)} \sum_{n=0}^{+\infty} r^n L(x_n, u_n)$$

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

2.1.1 Problème en horizon fini

A priori, on ne s'intéresse dans le problème de minimisation considéré plus haut qu'à la position initiale $\bar{x}$ qui est donnée. Pour mettre en oeuvre le principe de programmation dynamique, nous allons résoudre un grand nombre de problèmes (pour *toutes* les conditions initiales et en commençant à *n'importe quel instant*).

Pour cela, définissons la fonction valeur du problème comme étant la quantité, pour tout $\bar{x} \in X$ et $\bar{n} \in \{0, \dots, N-1\}$

$$V(\bar{n}, \bar{x}) := \inf_{(u_n)} \sum_{n=\bar{n}}^{N-1} L_n(x_n, u_n) + g(x_N)$$

où l'infimum est pris sur les éléments $(u_n) = (u_{\bar{n}}, \dots, u_{N-1})$ de $U_{\bar{n}} \times \dots \times U_{N-1}$ et où l'état $(x_n)_{n \in \{\bar{n}, \dots, N\}}$ est défini par récurrence par

$$\begin{cases} x_{\bar{n}} = \bar{x} \\ x_{n+1} = f_n(x_n, u_n), \qquad n = \bar{n},\ 1,\ 2, \dots,\ N-1 \end{cases}$$

La quantité qui nous intéresse est $V(0, \bar{x})$.

Théorème 2.1.1 (Programmation dynamique)**.** *Pour tout $x \in X$ et $\bar{n} \in \{0, \dots, N-1\}$, on a*

$$V(\bar{n}, x) = \inf_{u \in U_{\bar{n}}} \left\{ L_{\bar{n}}(x, u) + V(\bar{n}+1, f_{\bar{n}}(x, u)) \right\}, \qquad V(N, x) = g(x).$$

L'égalité ci-dessus porte le nom *d'équation de Bellman.*

Noter que le problème dans le membre de droite de l'égalité est en principe "plus simple" à résoudre que le problème initial, puisqu'il s'agit d'un problème de minimisation standard. Pour calculer $V(0, x_0)$, on résout par induction rétrograde les problèmes :

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

Preuve du théorème 2.1.1. Posons

$$W(\bar{n}, x) := \inf_{u \in U_{\bar{n}}} \left\{ L_{\bar{n}}(x, u) + V(\bar{n}+1, f_{\bar{n}}(x, u)) \right\}.$$

On veut montrer que $W = V$.

Soit $\epsilon > 0$ et $(u_n)_{n\geq\bar{n}}$ un contrôle $\epsilon-$optimal pour $V(\bar{n}, x)$. Alors

$$\begin{aligned} V(\bar{n},x)+\epsilon &\geq \sum_{n=\bar{n}}^{N-1} L_n(x_n,u_n)+g(x_N) \;=\; L_{\bar{n}}(x,u_{\bar{n}})+\sum_{n=\bar{n}+1}^{N-1} L_n(x_n,u_n)+g(x_N) \\ &\geq L_{\bar{n}}(x,u_{\bar{n}})+V(\bar{n}+1,f_{\bar{n}}(x,u_{\bar{n}})) \;\geq\; W(\bar{n},x) \end{aligned}$$

puisque $x_{\bar{n}+1} = f_{\bar{n}}(x, u_{\bar{n}})$. Comme ϵ est arbitraire, cela montre que $V \geq W$.

Inversement, soit $u_{\bar{n}} \in U_{\bar{n}}$ $\epsilon-$optimal pour $W(\bar{n}, x)$:

$$W(\bar{n},x)+\epsilon \geq L_{\bar{n}}(x,u_{\bar{n}})+V(\bar{n}+1,f_{\bar{n}}(x,u_{\bar{n}})).$$

Soit également $(u_n)_{n\geq\bar{n}+1}$ $\epsilon-$optimal pour $V(\bar{n}+1, f_{\bar{n}}(x, u_{\bar{n}}))$:

$$V(\bar{n}+1,f_{\bar{n}}(x,u_{\bar{n}}))+\epsilon \geq \sum_{n=\bar{n}+1}^{N-1} L_n(x_n,u_n)+g(x_N).$$

Définissons alors le contrôle

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

et notons $(\hat{x}_n)$ la solution associée issue de x en temps $\bar{n}$. Notons que $\hat{x}_{\bar{n}} = x$, $\hat{x}_{\bar{n}+1} = f_{\bar{n}}(x, u_{\bar{n}})$ et $\hat{x}_n = x_n$ pour $n \geq \bar{n}+1$. Alors

$$\begin{aligned} V(\bar{n},x) &\leq \sum_{n=\bar{n}}^{N-1} L_n(\hat{x}_n,\hat{u}_n)+g(\hat{x}_N) \;=\; L_{\bar{n}}(x,u_{\bar{n}})+\sum_{n=\bar{n}+1}^{N-1} L_n(x_n,u_n)+g(x_N) \\ &\leq L_{\bar{n}}(x,u_{\bar{n}})+V(\bar{n}+1,f_{\bar{n}}(x,u_{\bar{n}}))+\epsilon \;\leq\; W(\bar{n},x)+2\epsilon \end{aligned}$$

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

Un des principaux intérêts de la fonction valeur est de permettre le calcul des solutions optimales. Supposons que, pour tout $(n, x) \in \{0, \ldots, N-1\} \times X$, il existe $u_n^*(x)$ un "feedback optimal", i.e. 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 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. 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 2.1.1. *Soit $\bar{x}$ une position 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 :

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

Preuve. Montrons par récurrence que, pour tout $\bar{n} \in \{0, \ldots, N\}$,

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

Cette relation est clairement vraie pour $\bar{n} = 0$. Supposons-la pour un certain $\bar{n}$. En utilisant d'abord la programmation dynamique puis le choix de u^*, on a

$$V(\bar{n}, \bar{x}_{\bar{n}}) = \inf_{u \in U_{\bar{n}}} \{L_{\bar{n}}(\bar{x}_{\bar{n}}, u) + V(\bar{n}+1, f_{\bar{n}}(\bar{x}_{\bar{n}}, u))\} = L_{\bar{n}}(\bar{x}_{\bar{n}}, u^*_{\bar{n}}(\bar{x}_n)) + V(\bar{n}+1, f_{\bar{n}}(\bar{x}_{\bar{n}}, u^*_{\bar{n}}(x_{\bar{n}})))$$

où $u^*_{\bar{n}}(x_{\bar{n}}) = \bar{u}_{\bar{n}}$ et $f_{\bar{n}}(\bar{x}_{\bar{n}}, u^*_{\bar{n}}(x_{\bar{n}})) = \bar{x}_{\bar{n}+1}$. On utilise alors l'hypothèse de récurrence pour obtenir :

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

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

Reste à vérifier la continuité de V :

Proposition 2.1.2. *Sous les hypothèses ci-dessus, pour tout $n \in \{0, \dots, N-1\}$ l'application $x \to V(n, x)$ est continue.*

Preuve. Cela se montre (par exemple) 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 un résultat classique évoqué ci-dessous (cf. Exercice 2.1). □

Exercice 2.1. 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. Montrer que l'application marginale

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

est continue sur X. Montrer par un contre-exemple que le résultat n'est pas toujours vrai lorsque U n'est pas compact.

2.1.2 Problème 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

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

où l'infimum est pris sur les éléments $(u_n)_{n \in \mathbb{N}}$ de U 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 $\sum_{n=0}^{+\infty} r^n L(x_n, u_n)$ est bien convergente car L est bornée et $r \in]0,1[$.

Théorème 2.1.2 (Programmation dynamique)**.** *Pour tout $x \in X$, on a*

$$V(x) = \inf_{u \in U} \{L(x,u) + rV(f(x,u))\}. \tag{2.1}$$

L'égalité ci-dessus porte le nom *d'équation de Bellman.*

Contrairement au cas de l'horizon fini, la relation de programmation dynamique décrite ci-dessus ne permet pas un calcul explicite direct de la fonction valeur, puisque V apparaît à gauche et à droite de l'égalité. Cependant, nous expliquons plus bas que V est "l'unique solution" de cette équation implicite, ce qui peut fournir des schémas de calcul numérique pour V.

Preuve du théorème 2.1.2. 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

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

On veut prouver que $W = V$.

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

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

(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 $u \in U$ $\epsilon-$optimal pour $W(x)$, (u_n) $\epsilon-$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

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

et on note $(\hat{x}_n)_{n\geq 0}$ la trajectoire associée issue de x. Alors $\hat{x}_1 = f(\hat{x}_0, \hat{u}_0) = f(x,u) = x_0$ et donc, par récurrence, $\hat{x}_{n+1} = x_n$ pour tout $n \geq 0$ (on utilise encore le fait que f ne dépend pas de n). Par conséquent,

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

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

La relation (2.1) caractérise la fonction valeur, au moins dans certains cadres. Pour expliquer cela, posons

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

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

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

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

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

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 2.1.3. *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 2.1.2 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 :

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

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 2.1.3. *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 :

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

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

$$V(x) = \sum_{n=0}^{N-1} r^n L(\bar{x}_n, \bar{u}_n) + r^N V(\bar{x}_N). \tag{2.2}$$

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 (2.2). Comme V est borné et r appartient à $]0, 1[$, le terme $(r^N V(\bar{x}_N))$ tend vers 0 et (2.2) devient

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

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

2.2 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.

2.2.1 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 $f(x) := \int_0^1 x(t)dt$ sous la contrainte de longueur $g(x) := \int_0^1 \sqrt{1 + (x'(t))^2}dt = L$. 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...

2.2.2 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

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

où l'application $\epsilon : X \to \mathbb{R}$ vérifie $\lim_{x \to x_0} \epsilon(x) = 0$. 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 2.2.1. *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*

$$df(x_0)(v) = \int_0^1 \langle \frac{\partial L}{\partial x}(t, x_0(t), x_0'(t)), v(t)\rangle + \langle \frac{\partial L}{\partial p}(t, x_0(t), x_0'(t)), v'(t)\rangle \, dt \qquad \forall v \in X, \ \forall x_0 \in X \ . \tag{2.3}$$

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

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

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

Pour cela, fixons $\delta > 0$ petit et montrons qu'il existe $\eta > 0$ tel que, si $\|x - x_0\| \leq \eta$, on a $|\epsilon(x)| \leq \delta$. 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 :

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

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

$$\begin{aligned} &\Big| L(t, x(t), x'(t)) - L(t, x_0(t), x_0'(t)) - \langle \tfrac{\partial L}{\partial x}, x(t) - x_0(t)\rangle + \langle \tfrac{\partial L}{\partial p}, x'(t) - x_0'(t)\rangle \Big| \\ &\qquad \leq \delta(|x(t) - x_0(t)| + |x'(t) - x_0'(t)|) \leq \delta \|x - x_0\| \ , \end{aligned}$$

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

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

i.e., $|\epsilon(x)| \leq \delta$. Nous avons donc montré que $\lim_{x \to x_0} \epsilon(x) = 0$. Donc f est différentiable en tout point x_0 de X, et $df(x_0)$ est donnée par (2.3).

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 $\|v\| \leq 1$:

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

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 repoussons à plus tard la question de l'existence d'un minimum : celle-ci est assez délicate. Commençons par étudier les conditions nécessaires d'optimalité.

Théorème 2.2.1 (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*

$$\frac{d}{dt}\frac{\partial L}{\partial p}(t,x(t),x'(t)) = \frac{\partial L}{\partial x}(t,x(t),x'(t)) \qquad \forall t\in[0,1] \ . \tag{2.4}$$

La relation (2.4) 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 (2.4) sur un intervalle (non vide) $]a,b[$.

La preuve du théorème repose sur le lemme suivant :

Lemme 2.2.1 (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 2.2.1. 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

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

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 2.2.1. 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 2.2.1) 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 2.2.1 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 $\phi(t) = -\phi_1(t) + \langle \frac{\partial L}{\partial p}(t, x(t), x'(t)), v'(t)\rangle$. 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

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

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

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

□

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

Proposition 2.2.2. *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 (2.4), 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.

Exemple 2.2.1 (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 :

$$\min \left\{ \int_0^1 |(\Phi \circ x)'(t)|dt \mid x \in \mathcal{C}^1([0,1]; U),\ x(0) = A,\ x(1) = B \right\} \tag{2.5}$$

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

$$\min \left\{ \int_0^1 |(\Phi \circ x)'(t)|^2 dt \mid x \in \mathcal{C}^1([0,1]; U),\ x(0) = A,\ x(1) = B \right\} \tag{2.6}$$

Nous allons montrer que les problèmes (2.5) et (2.6) sont intimement liés.

Lemme 2.2.2. *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 (2.6), alors x est un minimum du problème (2.5). De plus, si x est une extrémale pour (2.6), alors x est une extrémale pour (2.5), 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 (2.5), alors x est un minimum pour (2.6). Cela est vrai après renormalisation (cf. la preuve du Lemme).

On appelle géodésique toute extrémale de (2.6), 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

$$\int_0^1 |(\Phi \circ x)'(t)|dt \le \left(\int_0^1 |(\Phi \circ x)'(t)|^2 dt \right)^{\frac{1}{2}} \tag{2.7}$$

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, (2.7) implique que, si x est un minimum de (2.6), alors x est un minimum de (2.5) puisque les infimum coïncident.

Supposons maintenant que x est extrémal pour (2.6). 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 (2.6), 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 (2.5). □

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

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

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 (2.6) dit que, si $x(t) = (\varphi(t), \theta(t))$ est une extrémale du problème, alors

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

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 2.2.2. *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*

$$z(0) = z(1) = 0 \qquad \text{et} \qquad dg(x)(z) \neq 0 \ , \tag{2.8}$$

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

$$\frac{d}{dt}\ \frac{\partial(L + \lambda M)}{\partial p}(t, x(t), x'(t)) = \frac{\partial(L + \lambda M)}{\partial x}(t, x(t), x'(t)) \qquad \forall t \in [0, 1] \ . \tag{2.9}$$

Remarque : la condition (2.8) 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

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

où z est défini par (2.8). Alors de la proposition 2.2.1 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 (2.8)), 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 2.2.1. □

Exemple 2.2.2 (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

$$x(t) = \lambda \sqrt{1 - (c-t)^2/\lambda^2} + cte \tag{2.10}$$

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 (2.10) pour $c = 1/2$ et λ l'unique solution de $2\lambda \sin(L/2\lambda) = 1$ telle que $L/2\lambda < \pi/2$.

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.

2.3.1 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. Rappelons que l'espace $W^{1,\infty}([0,T], \mathbb{R}^d)$ est l'ensemble des primitives de fonction L^∞. Cet espace coïncide avec l'ensemble des fonctions lipschitziennes de $[0,T]$

dans $\mathbb{R}^d$. Bien qu'on ne l'utilisera pas explicitement, il sera utile de garder en tête le fait qu'une fonction lipschitienne est dérivable presque partout (théorème de Rademacher). Cette dérivée p.p. coïncide avec la dérivée au sens des distributions.

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)

$$\begin{cases} \dot{x}(t) = f(t,x(t),u(t)), & t \in [t_0,T] \\ x(0) = x_0 \end{cases} \tag{2.11}$$

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 2.3.1 (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* (2.11)*.*

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 2.3.2 (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.

2.3.2 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

$$H(t,x,p) := \sup_{u \in U} \left\{ -\langle p, f(t,x,u)\rangle - L(t,x,u) \right\}.$$

On suppose que H est de classe C^1.

Théorème 2.3.3 (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.

2.3.3 Le principe de programmation dynamique

Définissons la fonction valeur $V : [0,T] \times \mathbb{R}^N \to \mathbb{R}$ par

$$V(t_0, x_0) := \inf_{u(\cdot)} \int_{t_0}^{T} L(t, x(t), u(t))\ dt + g(x(T) \tag{2.12}$$

sous la contrainte que le couple $x(\cdot)$ vérifie 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 2.3.4. *On a, pour tout $0 \le t_0 < t_1 \le T$,*

$$V(t_0, x_0) = \inf_{u(\cdot)} \int_{t_0}^{t_1} L(t, x(t), u(t))\ dt + V(t_1, x(t_1))$$

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

$$W(t_0, x_0) := \inf_{u(\cdot)} \int_{t_0}^{t_1} L(t, x(t), u(t))\ dt + V(t_1, x(t_1))$$

On veut montrer l'égalité $W = V$.

Soit $\epsilon > 0$ et $u(\cdot)$ $\epsilon-$optimal pour $V(t_0, x_0)$. Alors

$$\begin{aligned} V(t_0, x_0) - \epsilon \ &\ge \ \int_{t_0}^{t_1} L(t, x(t), u(t))\ dt + \int_{t_1}^{T} L(t, x(t), u(t))\ dt + g(x(T)) \\ &\ge \ \int_{t_0}^{t_1} L(t, x(t), u(t))\ dt + V(t_1, x(t_1)) \end{aligned}$$

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

$$V(t_0, x_0) - \epsilon \ \ge \int_{t_0}^{t_1} L(t, x(t), u(t))\ dt + V(t_1, x(t_1)) \ \ge \ W(t_0, x_0),$$

ce qui prouve l'inégalité $V \ge W$ car ϵ est arbitraire.

Inversement, soit $\epsilon > 0$ et (x_1, u_1) $\epsilon-$optimal pour $W(t_0, x_0)$. Choisissons également (x_2, u_2) $\epsilon-$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

$$\begin{aligned} V(t_0, x_0) &\leq \int_{t_0}^{T} L(t, x(t), u(t))\ dt + g(x(T)) \\ &= \int_{t_0}^{t_1} L(t, x_1(t), u_1(t))\ dt + \int_{t_1}^{T} L(t, x_2(t), u_2(t))\ dt + g(x_2(T)) \\ &\leq \int_{t_0}^{t_1} L(t, x_1(t), u_1(t))\ dt + V(t_1, x_1(t_1)) + \epsilon \qquad \text{(par définition de } (x_2, u_2)) \\ &\leq W(t_0, x_0) + 2\epsilon \qquad \text{(par définition de } (x_1, u_1)) \end{aligned}$$

Cela montre que $V(t_0, x_0) \leq W(t_0, x_0)$ car ϵ est arbitraire. D'où l'égalité $V = W$. □

2.3.4 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

$$H(t, x, p) := \sup_{u \in U} \left\{ -\langle p, f(t, x, u) \rangle - L(t, x, u) \right\}.$$

Vu nos hypothèses de continuité sur f et L, l'application H est continue.

Théorème 2.3.5. *On suppose que la fonction valeur V définie par* (2.12) *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 $\epsilon > 0$ et (t_0, x_0) tels que

$$-\frac{\partial V}{\partial t}(t_0, x_0) + H(t_0, x_0, \frac{\partial V}{\partial x}(t_0, x_0)) \ < \ -2\epsilon.$$

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) :

$$-\frac{\partial V}{\partial t}(t, x) + \sup_{u \in U} \left\{ -\langle \frac{\partial V}{\partial x}(t, x), f(t, x, u) \rangle - L(t, x, u) \right\} \ < \ -\epsilon \qquad \forall (t, x) \in B_\eta(t_0, x_0),$$

ce qui se réécrit

$$\frac{\partial V}{\partial t}(t, x) + \langle \frac{\partial V}{\partial x}(t, x), f(t, x, u) \rangle + L(t, x, u) \ > \ \epsilon \qquad \forall (t, x) \in B_\eta(t_0, x_0), \ \forall u \in U. \tag{2.13}$$

Soit $h > 0$ petit, (x_h, u_h) ϵh^2-optimal pour la programmation dynamique sur l'intervalle $[t_0, t_0 + h]$:

$$V(t_0, x_0) + \epsilon h^2 \geq \int_{t_0}^{t_0+h} L(t, x_h(t), u_h(t)) \ dt + V(t_0 + h, x_h(t_0 + h)).$$

Alors

$$\begin{aligned} &V(t_0, x_0) - \epsilon h^2 \\ &\quad \geq V(t_0, x_0) + \int_{t_0}^{t_0+h} L(t, x_h(t), u_h(t)) + \frac{\partial V}{\partial t}(t, x_h(t)) + \langle \frac{\partial V}{\partial x}(t, x_h(t)), f(t, x_h(t), u_h(t)) \rangle \ dt \end{aligned}$$

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 (2.13),

$$V(t_0, x_0) - \epsilon h^2 \ \geq \ V(t_0, x_0) + \int_{t_0}^{t_0+h} \epsilon \ dt \ = \ V(t_0, x_0) + \epsilon h,$$

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 2.3.6 (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. □

Annexe A

Convexité

A.1 Définitions générales et propriété élémentaires

Définition A.1.1.

- *Un sous-ensemble C de $\mathbb{R}^n$ est dit convexe si, pour tout x et y appartenant à C, le segment $[x, y]$ est contenu dans C :*
$$\forall (x, y) \in C^2,\ \forall \lambda \in [0, 1],\ \ \lambda x + (1 - \lambda)y \in C\ .$$
- *Soit C un sous-ensemble convexe de $\mathbb{R}^n$ et $f : C \to \mathbb{R}$ une fonction. On dit que f est convexe dans C si*
$$\forall (x, y) \in C^2,\ \forall \lambda \in [0, 1],\ f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)\ .$$
- *Soient C un sous-ensemble convexe de $\mathbb{R}^n$ et $f : C \to \mathbb{R}$ une fonction. On dit que f est strictement convexe dans C si*
$$\forall (x, y) \in C^2,\ \forall \lambda \in]0, 1[,\ f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y)\ .$$
- *Soient C un sous-ensemble convexe de $\mathbb{R}^n$ et $f : C \to \mathbb{R}$ une fonction. On dit que f est concave dans C si $-f$ est convexe dans C.*

Proposition A.1.1. *Soit C un convexe de $\mathbb{R}^n$. Alors*

1. *Si $f_1 : C \to \mathbb{R}$ et $f_2 : C \to \mathbb{R}$ sont convexes dans C, $f_1 + f_2$ l'est aussi.*
2. *Si I est un ensemble quelconque de paramètres et si $f_i : C \to \mathbb{R}$ sont des fonctions convexes, alors*
$$C' = \{x \in C \mid \sup_{i \in I} f_i(x) < +\infty\}$$
est un sous-ensemble convexe de C et l'application $x \to \sup_{i \in I} f_i(x)$ est convexe sur C'.
3. *Si $f : C \to \mathbb{R}$ est convexe dans C et $g : \mathbb{R} \to \mathbb{R}$ est convexe croissante, alors $g \circ f$ est convexe.*
4. *Si $f : C \to \mathbb{R}$ est à convexe et concave, alors f est affine, c'est-à-dire qu'il existe $a \in \mathbb{R}^n$ et $b \in \mathbb{R}$ tels que*
$$\forall x \in C,\ f(x) = \langle a, x \rangle + b\ .$$

Ces résultats sont très classiques. On laisse leur démonstration en exercice.

A.2 Convexité dans $\mathbb{R}$

Proposition A.2.1. *Les seuls sous-ensembles convexes de $\mathbb{R}$ sont les intervalles.*

Preuve. En effet, soit C un convexe non vide de $\mathbb{R}$. On note $a \in \mathbb{R} \cup \{-\infty\}$ sa borne inférieure et $b \in \mathbb{R} \cup \{+\infty\}$ sa borne supérieure. On affirme que $]a, b[\subset C$. En effet, si $x \in]a, b[$, par définition des bornes inférieure et supérieure il existe $a' \in C$ et $b' \in C$ avec $a' < x < b'$. Comme C est convexe, le point x appartient donc à C. D'où $]a, b[\subset C$.

On en déduit que C ne peut être que $]a, b[$, $[a, b[$, $]a, b]$ ou $[a, b]$, et donc que C est un intervalle. □

Théorème A.2.1. *Soit I un intervalle ouvert de $\mathbb{R}$ et $f : I \to \mathbb{R}$ de classe $\mathcal{C}^1$. La fonction f est convexe dans I si et seulement si sa dérivée f' est croissante dans I.*

Si f est de classe $\mathcal{C}^2$ dans I, alors f est convexe dans I si et seulement si sa dérivée seconde est positive sur I : $\forall x \in I$, $f"(x) \geq 0$.

En fait, nous allons montrer un résultat un peu plus précis : pour cela, rappelons les définitions des dérivées à gauche (notée $f_g(x)$) et à droite (notée $f_d(x)$) d'une fonction f en un point x :

$$f'_g(x) = \lim_{h \to 0^+} \frac{f(x) - f(x-h)}{h} \text{ et } f'_d(x) = \lim_{h \to 0^+} \frac{f(x+h) - f(x)}{h} .$$

Lemme A.2.1. *Soit I un intervalle ouvert de $\mathbb{R}$ et $f : I \to \mathbb{R}$ une fonction convexe. Alors f admet une dérivée à gauche et une dérivée à droite en tout point de I. De plus,*

$$\forall (x, y) \in I \times I,\ f'_g(x) \leq f'_d(x) \leq f'_g(y) \leq f'_d(y) .$$

Preuve du lemme. Notons d'abord que le quotient différentiel $\phi_d(x, h) = \frac{f(x+h)-f(x)}{h}$ est une fonction croissante de h sur $\mathbb{R}^+$. En effet, si $0 \leq h_1 < h_2$,

$$f(x + h_1) = f\left((1 - \frac{h_1}{h_2})x + \frac{h_1}{h_2}(x + h_2)\right) \leq (1 - \frac{h_1}{h_2})f(x) + \frac{h_1}{h_2} f(x + h_2) .$$

D'où

$$\frac{f(x + h_1) - f(x)}{h_1} \leq \frac{f(x + h_2) - f(x)}{h_2} .$$

On montre de même que le quotient différentiel $\phi_g(x, h) = \frac{f(x)-f(x-h)}{h}$ est une fonction décroissante de h sur $\mathbb{R}^+$. De plus, pour tout $h > 0$, on a :

$$\phi_d(x, h) - \phi_g(x, h) = \frac{f(x + h) + f(x - h) - 2f(x)}{h} \geq 0 .$$

Comme $\phi_d(x, \cdot)$ est croissante et $\phi_g(x, \cdot)$ décroissante, cela prouve que $\phi_d(x, h)$ et $\phi_g(x, h)$ ont une limite finie lorsque $h \to 0^+$ et que ces limites vérifient $f'_g(x) \leq f'_d(x)$.

Reste à montrer que, si $x < y$, alors $f'_d(x) \leq f'_g(y)$. Pour cela, il suffit d'utiliser ce que l'on vient de prouver :

$$f'_d(x) \leq \phi_d(x, y - x) = \phi_g(y, y - x) \leq f'_g(y) .$$

□

Preuve du théorème. Si f est convexe et $\mathcal{C}^1$ sur I, alors f' est croissante d'après le lemme.

Réciproquement, supposons que f' est croissante. Fixons $x < y$ dans I et $s \in]0, 1[$. La formule de Taylor-Lagrange à l'ordre 1 entre $z = sx + (1 - s)y$ et x affirme qu'il existe $z_1 \in]x, z[$ tel que

$$f(x) = f(z) + f'(z_1)(x - z) = f(z) + (1 - t)f'(z_1)(x - y) .$$

On montre de même, en utilisant la formule de Taylor-Lagrange entre z et y, qu'il existe $z_2 \in]z, y[$ tel que

$$f(y) = f(z) + tf'(z_2)(y - x) .$$

On multiplie la première de ces deux égalités par t et la deuxième par $(1 - t)$, puis on additionne pour obtenir :

$$tf(x) + (1 - t)f(y) = f(z) + t(1 - t)(f'(z_2) - f'(z_1))(y - x) .$$

Or $z_1 < z_2$, donc $f'(z_1) \leq f'(z_2)$ car f' est croissante. D'où le résultat désiré : $tf(x) + (1 - t)f(y) \geq f(z)$.

La dernière partie du théorème est alors immédiate. □

Voici une autre application du lemme :

Corollaire A.2.1. *Soit I un intervalle ouvert et $f : I \to \mathbb{R}$ une fonction convexe. Alors f est dérivable dans I, sauf éventuellement en un nombre au plus dénombrable de points.*

Preuve. Comme, d'après le lemme, f'_d est une fonction croissante, f'_d est continue sur I sauf éventuellement dans un ensemble $\mathcal{N}$ au plus dénombrable de points. Si $x \in I \backslash \mathcal{N}$, on a, toujours d'après le lemme, que

$$\lim_{h \to 0^+} f'_d(x - h) \leq f'_g(x) \leq f'_d(x) .$$

Comme f'_d est continue en x, cela prouve que $f'_g(x) = f'_d(x)$, c'est-à-dire que f est dérivable en x. □

A.3 Caractérisation des fonctions convexes : le cas régulier

Nous aurons d'abord besoin d'un lemme :

Lemme A.3.1. *Soient C un sous-ensemble convexe de $\mathbb{R}^n$ et $f : C \to \mathbb{R}$ une fonction. Alors f est convexe dans C, si et seulement si, pour tout $(x, y) \in C \times C$, l'application $\phi_{x,y} : [0, 1] \to \mathbb{R}$ définie par*

$$\forall s \in [0, 1],\ \phi_{x,y}(s) = f(sx + (1 - s)y)$$

est une fonction convexe.

Preuve. Si f est convexe, alors pour tout $(x, y) \in C \times C$, pour tout $(s_1, s_2) \in [0, 1]$ et pour tout $t \in [0, 1]$, on a

$$\begin{aligned} \phi_{x,y}(ts_1 + (1 - t)s_2) &= f((ts_1 + (1 - t)s_2)x + (1 - ts_1 - (1 - t)s_2)y) \\ &= f(t(s_1x + (1 - s_2)y) + (1 - t)(s_1x + (1 - s_2)y)) \\ &\leq tf(s_1x + (1 - s_2)y) + (1 - t)f(s_1x + (1 - s_2)y) \\ &= t\phi_{x,y}(s_1) + (1 - t)\phi_{x,y}(s_2) \end{aligned}$$

Donc $\phi_{x,y}$ est convexe dans $[0, 1]$.

Réciproquement, supposons que $\phi_{x,y}$ soit convexe pour tout $(x, y) \in C \times C$. Alors, pour tout $t \in [0, 1]$,

$$\phi_{x,y}(t) = \phi_{x,y}(t.1 + (1 - t).0) \leq t\phi_{x,y}(1) + (1 - t)\phi_{x,y}(0)$$

par convexité de $\phi_{x,y}$. Ce qui donne, d'après la définition de $\phi_{x,y}$:

$$f(sx + (1 - s)y) \leq tf(x) + (1 - t)f(y)\ .$$

Donc f est convexe. □

Nous commençons par caractériser la convexité dans le cas $\mathcal{C}^1$:

Théorème A.3.1. *Soient C un sous-ensemble ouvert et convexe de $\mathbb{R}^n$ et $f : C \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$ dans C. Les assertions suivantes sont équivalentes :*

i) f est convexe,

ii) $\forall (x, y) \in C \times C,\ \langle \nabla f(x) - \nabla f(y), x - y\rangle \geq 0$.

iii) $\forall (x, y) \in C \times C,\ f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$.

Preuve. Notons d'abord que f étant $\mathcal{C}^1$, $\phi_{x,y}$ l'est aussi pour $(x, y) \in C \times C$.

Montrons que $(i) \Rightarrow (ii)$. Comme f est convexe, pour $(x, y) \in C \times C$, la fonction $\phi_{x,y}$ est convexe. D'après la caractérisation de la convexité dans $\mathbb{R}$, cela donne que $\phi'_{x,y}$ est croissante sur $]0, 1[$. Par continuité, on a donc :

$$\forall s \in [0, 1],\ \phi'_{x,y}(0) = \langle \nabla f(y), x - y\rangle \leq \phi'_{x,y}(1) = \langle \nabla f(x), x - y\rangle\ .$$

D'où (ii).

Montrons maintenant que $(ii) \Rightarrow (iii)$. On écrit la formule de Taylor-Lagrange à l'ordre 1 en x et y : il existe $s \in]0, 1[$ tel que

$$f(y) = f(x) + \langle \nabla f(sx + (1 - s)y), y - x\rangle\ .$$

Or, si on pose $z = sx + (1 - s)y$, on a

$$\langle \nabla f(z), y - x\rangle - \langle \nabla f(x), y - x\rangle = \frac{1}{1 - s}\langle \nabla f(z) - \nabla f(x), z - x\rangle \geq 0$$

par hypothèse. Donc $\langle \nabla f(z), y - x\rangle \geq \langle \nabla f(x), y - x\rangle$, ce qui implique l'inégalité désirée : $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$.

Montrons enfin que $(iii) \Rightarrow (i)$. De (iii), on déduit que

$$\forall z \in C,\ f(z) = \sup_{x \in C} \{f(x) + \langle \nabla f(x), z - x\rangle\}\ .$$

En effet, l'inégalité $f(z) \geq \sup_{x \in C} f(x) + \langle \nabla f(x), z - x\rangle$ est donnée par (iii). L'inégalité inverse est évidente (prendre $x = z$ dans le sup.).

Par conséquent, f est égal au supremum des fonctions affines $z \to f(x) + \langle \nabla f(x), z - x\rangle$. Donc f est convexe. □

Théorème A.3.2. *Soient C un sous-ensemble ouvert et convexe de $\mathbb{R}^n$ et $f : C \to \mathbb{R}$ une fonction de classe $\mathcal{C}^2$. La fonction f est convexe, si et seulement si,*

$$\forall x \in C,\ Hess_f(x) \geq 0 \text{ (au sens des matrices symétriques)}$$

Si, de plus, on a

$$\forall x \in C,\ Hess_f(x) > 0 \text{ (au sens des matrices symétriques)}$$

alors la fonction f est strictement convexe sur C.

Preuve. Supposons d'abord que f soit convexe. Alors, pour tout $x \in C$, pour tout $v \in \mathbb{R}^n$, il existe $r > 0$ tel que $y = x + rv \in C$, car C est ouvert. Comme f est $\mathcal{C}^2$ et convexe, la fonction $\phi_{x,y}$ est aussi $\mathcal{C}^2$ et convexe, et on a

$$\phi_{x,y}"(s) = \langle Hess_f(sx + (1-s)y)v, v\rangle \geq 0 \ .$$

Lorsque $s \to 1^-$, cela donne que $\langle Hess_f(x)v, v\rangle \geq 0$ pour tout $v \in \mathbb{R}^n$, i.e., que $Hess_f(x) \geq 0$ au sens des matrices symmétriques.

Réciproquement, supposons que $Hess_f(x) \geq 0$ pour tout x de C. Fixons x et y dans C, et $s \in [0,1]$. Ecrivons la formule de Taylor-Lagrange à l'ordre 2 entre $sx + (1-s)y$ et x :

$$f(x) = f(sx + (1-s)y) + (1-s)\langle \nabla f(sx + (1-s)y), x - y\rangle\rangle + \frac{(1-s)^2}{2}\langle Hess_f(z)(x-y), (x-y)\rangle$$

pour z appartenant au segment $[sx + (1-s)y, x]$. Comme $Hess_f(z) \geq$, cela implique que

$$f(sx + (1-s)y) \geq f(x) + (1-s)\langle \nabla f(x), y - x\rangle \tag{A.1}$$

On écrit de même la formule de Taylor-Lagrange à l'ordre 2 entre $sx + (1-s)y$ et y et on obtient de façon similaire :

$$f(y) \geq f(sx + (1-s)y) + s\langle \nabla f(sx + (1-s)y), y - x\rangle \ .$$

En additionnant línégalité précédente multipliée par $(1-s)$ et l'inégalité (A.1) multipliée par s, on a alors l'inégalité désirée :

$$sf(x) + (1-s)f(y) \geq f(sx + (1-s)y) \ .$$

Donc f est convexe.

Si l'on suppose de plus que $Hess_f(x) > 0$ pour tout $x \in C$, la démonstration précédente prouve que f est strictement convexe. □

A.4 Caractérisation des fonctions convexes : le cas général

Nous aurons d'abord besoin du lemme suivant :

Lemme A.4.1. *Soient C un sous-ensemble convexe de $\mathbb{R}^n$ et $f : C \to \mathbb{R}$ une fonction. Alors f est convexe dans C si et seulement si, l'épigraphe de f, i.e., l'ensemble*

$$K = \{(x,t) \in C \times \mathbb{R} \mid t \geq f(x)\}$$

est un sous-ensemble convexe de $\mathbb{R}^n \times \mathbb{R}$.

Preuve. C'est un simple jeu d'écriture, que l'on laisse en exercice au lecteur. □

Théorème A.4.1. *Soit C un sous-ensemble ouvert et convexe de $\mathbb{R}^n$ et $f : C \to \mathbb{R}$ une fonction. Alors f est convexe dans C si et seulement si, pour tout $x \in C$, il existe $p \in \mathbb{R}^n$ tel que*

$$\forall y \in C,\ f(y) \geq f(x) + \langle p, y - x\rangle \ .$$

Remarques : 1) Si f est $\mathcal{C}^1$, le résultat est vrai avec $p = \nabla f(x)$. Le vecteur p est donc un "gradient généralisé" à la fonction f au point p.

2) Le résultat devient faux si C n'est pas ouvert (penser à $-\sqrt{x}$ sur $[0, +\infty[$).

Preuve. Supposons que f possède la propriété que, pour tout $x \in C$, il existe $p \in \mathbb{R}^n$ tel que

$$\forall y \in C,\ f(y) \geq f(x) + \langle p, y - x \rangle \ .$$

Notons

$$I = \{(a,b) \in \mathbb{R}^n \times \mathbb{R} \mid \forall y \in C, f(y) \geq c + \langle a, y \rangle\} \ .$$

Alors on voit facilement que l'hypothèse implique que

$$\forall x \in C,\ f(x) = \sup_{(a,b)\in A} c + \langle a, x \rangle \ .$$

Les applications $y \to c + \langle a, y \rangle$ sont convexes parce qu'affines, donc l'application $y \to \sup_{(a,b)\in A} c + \langle a, y \rangle$ est convexe. Cela prouve que f est convexe.

Réciproquement, si f est convexe, alors son épigraphe K est convexe. De plus, pour tout x appartenant à C, on voit facilement que $(x, u(x))$ appartient au bord de K. Donc le théorème de séparation affirme qu'il existe une forme linéaire α sur $\mathbb{R}^{n+1}$ telle que

$$\inf_{(y,t)\in K} \alpha(y,t) \ \leq \alpha(x, u(x)) \ . \tag{A.2}$$

Comme α est une forme linéaire sur $\mathbb{R}^n \times \mathbb{R}$, il existe $(u, v) \in \mathbb{R}^n \times \mathbb{R}$ avec $(u, v) \neq 0$ tel que

$$\forall (y,t) \in \mathbb{R}^n \times \mathbb{R},\ \alpha(y,t) = \langle u, x \rangle + vt \ .$$

Montrons que $v > 0$. Notons d'abord que, comme $(x, u(x) + 1)$ appartient à K, on a

$$\alpha(x, u(x) + s) = \langle u, x \rangle + v(u(x) + 1 > \geq \alpha(x, u(x)) = \langle u, x \rangle + vu(x) \ .$$

Donc $v \geq 0$. Pour montrer que $v > 0$, on raisonne par l'absurde en supposant $v = 0$. Alors, $u \neq 0$ car $(u, v) \neq 0$. Comme C est ouvert, il existe $r > 0$ tel que $x - ru$ appartient à C. Dans ce cas, l'inégalité (A.2) affirme que

$$\alpha(x - ru, f(x - ru)) = \langle u, x - ru \rangle = \langle u, x \rangle - r\|u\|^2 \geq \alpha(x) = \langle u, x \rangle \ .$$

C'est absurde. Donc on a montré que $v > 0$.

On applique à nouveau l'inégalité (A.2), que l'on divise par v :

$$\forall y \in C,\ \langle \frac{u}{v}, y \rangle + f(y) \geq \langle \frac{u}{v}, x \rangle + f(x) \ ,$$

ce qui donne le résultat désiré en posant $p = \frac{u}{v}$:

$$\forall y \in C,\ f(y) \geq f(x) + \langle p, y - x \rangle \ .$$

□

A.5 Régularité des fonctions convexes

En préliminaire, on a besoin de la notion de combinaison convexe : une combinaison convexe de points $a_1, \ldots, a_k$ de $\mathbb{R}^n$ est une somme de la forme $\sum_{i=1}^k \lambda_i a_i$ avec $\lambda_i \geq 0$ pour tout i et $\sum_{i=1}^k \lambda_i = 1$.

Lemme A.5.1. *Soit C un convexe de $\mathbb{R}^n$ et $f : C \to \mathbb{R}$ une fonction convexe. Si $a_1, \ldots a_k$ sont des points de C, pour toute combinaison convexe $\sum_{i=1}^k \lambda_i a_i$, on a :*

$$f\left(\sum_{i=1}^k \lambda_i a_i\right) \leq \sum_{i=1}^k \lambda_i f(a_i) \ .$$

En particulier,

$$f\left(\sum_{i=1}^k \lambda_i a_i\right) \leq \max_{i=1,\ldots,k} f(a_i) \ .$$

La preuve, qui est immédiate, se fait par récurrence.

Les fonctions convexes de $\mathbb{R}^n$ sont des fonctions assez régulières :

Théorème A.5.1. *Soit C un sous-ensemble ouvert et convexe de $\mathbb{R}^n$ et $f : C \to \mathbb{R}$ une fonction convexe. Alors f est localement lipschitzienne dans C, i.e. :*

pour tout $x_0 \in C$, il existe un rayon $r > 0$ tel que la boule $B_r(x_0)$ soit contenue dans C et il existe une constante L (dépendant de x_0), avec

$$\forall (x, y) \in B_r(x_0),\ |f(x) - f(y)| \leq L\|x - y\| .$$

En particulier, si f est convexe dans C, alors f est continue dans l'intérieur de C.

Preuve. On fixe $x_0 \in C$ et on suppose pour simplifier les notations que $f(x_0) = 0$ (on peut toujours se ramener à ce cas).

1. Montrons d'abord qu'il existe un rayon $r > 0$ tel que la boule $B_r(x_0)$ soit contenue dans C et f soit majorée sur $B_r(x_0)$ par une constante M.
   (Notons que $M \geq 0$ car $f(x_0) = 0$)
   Preuve : En effet, soit (e_i) une base orthonormée de $\mathbb{R}^n$ et $r > 0$ tel que la boule (pour la norme $\|\cdot\|_1$) $B_r(x_0)$ soit contenue dans C. Alors, pour tout $x \in B_r(x)$, il existe $(s_i)_{i=1,\dots,n}$ tel que
   $$x = \sum_i s_i e_i \text{ et } \forall i,\ \sum_i |s_i \leq 1 .$$
   Donc x est une combinaison convexe des $x_0 \pm re_i$, $i = 1, \dots, n$. Par conséquent,
   $$f(x) \leq \sup_{i=1,\dots,n} \max\{f(x_0 + re_i), f(x_0 - re_i)\} .$$
   En conclusion, f est majorée sur $B_r(x_0)$.
2. Montrons maintenant que f est minorée sur $B_r(x_0)$ par $-2M$.
   Preuve : Pour tout $x \in B_r(x_0)$, avec $x \neq x_0$, notons z le point de $\partial B_r(x_0)$ tel que x_0 appartient au segment $[x, z]$. Il existe alors $s \in [0, 1]$ tel que
   $$x_0 = sx + (1 - s)z,\ \text{c'est-à-dire } z = \frac{x_0 - (1 - s)z}{1 - s} .$$
   D'où
   $$r = \|z - x_0\| = \frac{s}{1 - s}\|x - x_0\| < r .$$
   On en déduit que $s > 1/2$. Comme f est convexe, l'égalité $x_0 = sx + (1 - s)z$ implique, puisque $f(x_0) = 0$, que
   $$0 = f(x_0) \leq sf(x) + (1 - s)f(z) \leq sf(x) + M$$
   car f est majorée par M. Donc
   $$f(x) \geq -\frac{M}{s} \geq -2M .$$
3. Montrons finalement que f est lipschitzienne de constante de lipschitz $L = \frac{6M}{r}$.
   Preuve : Soient $(x, y) \in B_{r/2}(x_0) \times B_{r/2}(x_0)$, avec $x \neq y$. Considérons le point z appartenant à $\partial B_r(x_0)$ tel que le point y appartient au segment $[x, z]$. Alors il existe $s \in]0, 1[$ tel que
   $$y = sx + (1 - s)z\ ,\ \text{c'est-à-dire } z = \frac{y - sx}{1 - s} .$$
   Alors
   $$r = \|z - x_0\| = \|\frac{y - x}{1 - s} + x - x_0\| \leq \frac{\|y - x\|}{1 - s} + \|x - x_0\| \leq \frac{\|y - x\|}{1 - s} + \frac{r}{2} .$$
   On en déduit que
   $$1 - s \leq \frac{2}{r}\|y - x\| .$$
   Comme f est convexe et $y = sx + (1 - s)z$, on a
   $$f(y) \leq sf(x) + (1 - s)f(z) \leq sf(x) + (1 - s)M ,$$
   car f est majorée par M. D'où
   $$f(y) - f(x) \leq -(1 - s)f(x) + (1 - s)M \leq 3M(1 - s) \leq \frac{6M}{r}\|x - y\| ,$$
   car $f(x) \geq -2M$ et $1 - s \leq \frac{2}{r}\|y - x\|$.
   En intervertissant les rôles de x et y, on obtient l'inégalité opposée : $f(x) - f(y) \leq L\|y - x\|$, ce qui implique le résultat.

□

En fait, le théorème suivant (dont la démonstration est beaucoup plus technique) affirme qu'une fonction convexe est presque de classe $\mathcal{C}^2$.

Théorème A.5.2 (Alexandroff)**.** *Soit C un ouvert convexe de $\mathbb{R}^n$ et $f : C \to \mathbb{R}$ une fonction convexe. Alors il existe un sous-ensemble $\Sigma \subset C$ de mesure nulle tel que, pour tout $x \notin \Sigma$, f admet un développement de Taylor à l'ordre 2 en x : il existe un vecteur $p \in \mathbb{R}^n$ et il existe une matrice symmétrique A de format $n \times n$ tels que :*

$$\forall y \in C,\ f(y) = f(x) + \langle p, y - x\rangle + \frac{1}{2}\langle A(y-x), (y-x)\rangle + \|y-x\|^2 \epsilon(y-x)\ ,$$

où $\epsilon(y-x)$ tend vers 0 lorsque y tend vers x.

A.6 Convexité et optimisation

Comme nous le verrons par la suite, la convexité joue un rôle très important en optimisation. Soulignons le résultat suivant, qui est élémentaire :

Proposition A.6.1. *Soient C un sous-ensemble fermé et convexe de $\mathbb{R}^n$ et $f : C \to \mathbb{R}$ une fonction convexe sur C. Alors l'ensemble des points où f atteint son minimum est un ensemble convexe (éventuellement vide) :*

$$\{x \in C \mid f(x) = \inf_{y \in C} f(y)\} \text{ est convexe.}$$

Si, de plus, f est strictement convexe, f admet au plus un minimum sur C.

C'est une application directe du lemme suivant, dont la démonstration est laissée au lecteur :

Lemme A.6.1. *Si $f : C \to \mathbb{R}$ est convexe, alors pour tout $c \in \mathbb{R}$, l'ensemble de niveau c de f, c'est-à-dire $\{x \in C \mid f(x) \leq c\}$ est un sous-ensemble convexe de C.*

Lorsque l'on étudiera les aspects numériques de la théorie, nous aurons souvent besoin d'une condition plus forte que la stricte convexité :

Définition A.6.1. *Soient un sous-ensemble ouvert et convexe de $\mathbb{R}^n$ et $f : C \to \mathbb{R}$ une fonction de classe $\mathcal{C}^2$ sur C. On dit que f est elliptique sur C s'il existe une constante $\alpha > 0$ telle que*

$$\forall x \in C,\ Hess_f(x) \geq \alpha I_n \text{ (au sens des matrices symétriques)}$$

La constante α est appelée constante d'ellipicité de f.

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