Optimisation et contrôle

Grégoire ALLAIRE, Alexandre ERN
Ecole Polytechnique

3 mars 2020

Table des matières

Préface

Ce cours traite de deux sujets distincts mais étroitement liés en mathématiques appliquées : l'optimisation et le contrôle. Avant même de présenter ces deux disciplines, disons tout de suite qu'à travers leur enseignement un des objectifs de ce cours est d'introduire le lecteur au monde de la **modélisation mathématique** et de son utilisation pour la **conception** et la **commande** de systèmes complexes, issus de tous les domaines de la science et des applications industrielles (ou sciences de l'ingénieur). La modélisation mathématique est l'art (ou la science, selon le point de vue) de représenter une réalité physique par des modèles abstraits accessibles à l'analyse et au calcul qui permettent d'apporter des réponses à la fois qualitatives et quantitatives à des questions concrètes. La conception des systèmes fait très souvent appel, explicitement ou implicitement, à la théorie de l'optimisation. D'ailleurs, on parle souvent de conception optimale. En effet, lorsqu'on conçoit un appareil, une structure, une organisation ou tout autre « système », on ne se contente pas en général de trouver une solution possible mais plutôt la « meilleure » solution possible et cela passe par l'utilisation de concepts et d'outils d'optimisation. De même le fonctionnement d'un système évoluant en temps nécessite de pouvoir le piloter ou le commander afin qu'il réagisse à des événements extérieurs : c'est ce qu'on appelle la contrôlabilité d'un système.

Les **objectifs de ce cours** sont de familiariser le lecteur avec les principales notions et résultats théoriques d'optimisation et de contrôle, ainsi que les algorithmes numériques qui en découlent. Le plan de ce cours est le suivant. Un premier chapitre d'introduction à **l'optimisation** et au **contrôle** donne de nombreux exemples et motivations pour l'étude de ces deux sujets. La première partie du cours (Chapitres 2 à 4) porte sur l'optimisation. Le Chapitre 2 présente quelques résultats d'existence de solutions à des problèmes d'optimisation, ainsi que des notion d'analyse convexe. Le Chapitre 3 dérive les conditions (nécessaires ou suffisantes) d'optimalité des solutions. Ces conditions sont importantes tant du point de vue théorique que numérique. Elles permettent de caractériser les optima, et elles sont à la base des algorithmes numériques que nous décrivons. Elles reposent sur des notions fondamentales comme celle des **multiplicateurs de Lagrange**, du **point-selle pour un Lagrangien** ou encore de la **dualité**. Finalement, le Chapitre 4 est dédié à la programmation linéaire qui est un outil essentiel pour la planification optimale des ressources et des tâches dans toutes les grandes entreprises (domaine de la recherche opérationnelle).

La deuxième partie du cours (Chapitres 5 à 7) est consacré à l'étude des systèmes de contrôle, c'est-à-dire des systèmes dynamiques sur lesquels on peut agir au moyen d'un contrôle ou d'une commande. Un premier objectif peut être d'amener le système d'un état initial donné à un état final (une cible), en respectant éventuellement certaines contraintes (par exemple, la valeur du contrôle ne peut être trop grande ou bien l'état du système doit appartenir à un domaine admissible). Il

s'agit du problème de la **contrôlabilité**. Un deuxième objectif peut être celui de déterminer un contrôle optimal, c'est-à-dire minimisant une fonctionnelle de coût dépendant du contrôle et de la trajectoire résultant de ce contrôle. Il s'agit du problème de **contrôle optimal**. Nous aborderons ces deux problèmes dans ce cours. Le champ d'applications est très vaste. On rencontre des problèmes de contrôlabilité et de contrôle optimal dans des domaines très variés, comme l'aéronautique, l'électronique, le génie des procédés, la médecine, l'économie et la finance, internet et les communications, etc. Le Chapitre 5 aborde le problème de la contrôlabilité. Le résultat phare est le **critère de Kalman** sur la contrôlabilité des systèmes linéaires autonomes et son extension à la contrôlabilité locale des systèmes non-linéaires. Le Chapitre 6 est consacré à l'étude du système linéaire-quadratique (dit système LQ) qui consiste à minimiser un critère quadratique pour un système de contrôle linéaire. Le système LQ étant particulièrement simple, il nous sera possible de mener une analyse complète du problème. Celle-ci repose sur diverses idées importantes, comme la notion d'état adjoint, de Hamiltonien et de feedback (ou rétro–action) grâce à l'équation de Riccati. Le Chapitre 7 étudie le problème du contrôle optimal par le biais du **principe du minimum de Pontryaguine** (PMP). Ce résultat est de portée générale car il permet de traiter des systèmes régis par des dynamiques non-linéaires et de considérer des fonctionnelles de coût non-quadratiques. Il s'étend également au cas où des contraintes d'atteinte de cible sont imposées et à celui où le temps final n'est pas fixé a priori. Finalement le Chapitre 8 est une annexe qui regroupe des rappels d'outils mathématiques utiles (espaces de Hilbert, sélections mesurables, équations différentielles ordinaires).

Le niveau de ce cours est introductif et il n'exige aucun autre prérequis que le niveau de connaissances acquis en classes préparatoires ou en premier cycle universitaire. Reconnaissons qu'il est difficile de faire preuve de beaucoup d'originalité sur ce sujet déjà bien classique dans la littérature. En particulier, notre cours doit beaucoup à ses prédécesseurs et notamment au cours de Pierre-Louis Lions [20].

Les auteurs remercient à l'avance tous ceux qui voudront bien leur signaler les inévitables erreurs ou imperfections de cette édition, par exemple par courrier électronique à l'adresse
`gregoire.allaire@polytechnique.fr`, `alexandre.ern@enpc.fr`.

G. Allaire, A. Ern
Paris, le 1er mars 2020

Chapitre 1

INTRODUCTION À L'OPTIMISATION ET AU CONTRÔLE

Ce chapitre est une introduction aux deux sujets de ce cours qui, quoique différents et indépendants, partagent de nombreux points communs. L'objectif est de présenter les motivations à ces deux sujets et d'illustrer leur portée et leur diversité à travers de nombreux exemples concrets.

1.1 Motivations

L'optimisation et le contrôle sont des sujets très anciens qui connaissent un nouvel essor depuis l'apparition des ordinateurs et dont les méthodes s'appliquent dans de très nombreux domaines : économie, gestion, planification, logistique, automatique, robotique, conception optimale, sciences de l'ingénieur, traitement du signal, etc. L'optimisation et le contrôle couvrent ainsi un champ scientifique relativement vaste, qui touche aussi bien au calcul des variations qu'à la recherche opérationnelle (en lien avec les processus de gestion ou de décision). Nous ne ferons souvent qu'effleurer ces sujets car il faudrait un polycopié complet pour chacun d'eux si nous voulions les traiter à fond.

Avant de considérer l'optimisation ou le contrôle d'un phénomène physique ou d'un système industriel, il faut déjà passer par une étape de **modélisation** qui permet de représenter cette réalité (et éventuellement de la simplifier si elle est trop complexe) par un modèle mathématique. Dans ce qui suit, nous considérerons des exemples de problèmes d'optimisation et de contrôle où les modèles peuvent être de nature très différente. Dans le cas le plus simple des problèmes d'optimisation, le modèle sera une simple équation algébrique et il s'agira simplement d'optimiser une fonction définie sur un espace de dimension finie (disons $\mathbb{R}^n$). Une deuxième catégorie de problèmes correspond au cas où la fonction à optimiser dépend de la solution d'une équation différentielle ordinaire (autrement dit, cette fonction est définie sur un espace de dimension infinie, par exemple l'espce $C[0,T]$ des fonctions

continues sur l'intervalle fermé $[0, T]$ en temps). On parle alors de contrôle (ou de commande) optimale, et les applications sont très nombreuses en automatique et robotique. La troisième et dernière catégorie correspond à l'optimisation de fonctions de la solution d'une équation aux dérivées partielles. Il s'agit alors de la théorie du contrôle optimal des systèmes distribués qui a de nombreuses applications, par exemple en conception optimale ou pour la stabilisation de structures mécaniques. Il ne nous sera pas possible dans ce cours de niveau introductif d'aborder le cas du contrôle des systèmes distribués. Remarquons néanmoins que ces catégories ne sont pas hermétiquement cloisonnées puisqu'après discrétisation spatiale une équation aux dérivées partielles se ramène à un système d'équations différentielles ordinaires et, qu'après discrétisation temporelle, une équation différentielle ordinaire se ramène à un système d'équations algébriques.

On peut aussi séparer l'optimisation en deux grandes branches aux méthodes fort différentes selon que les variables sont continues ou discrètes. Typiquement, si l'on minimise une fonction $f(x)$ avec $x \in \mathbb{R}^n$, il s'agit **d'optimisation en variables continues**, tandis que si $x \in \mathbb{Z}^n$ on a affaire à de **l'optimisation combinatoire** ou en variables discrètes. Malgré les apparences, l'optimisation en variables continues est souvent plus "facile" que l'optimisation en variables discrètes car on peut utiliser la notion de dérivée qui est fort utile tant du point de vue théorique qu'algorithmique. L'optimisation combinatoire est naturelle et essentielle dans de nombreux problèmes de la recherche opérationnelle. C'est un domaine où, à coté de résultats théoriques rigoureux, fleurissent de nombreuses "heuristiques" essentielles pour obtenir de bonnes performances algorithmiques. Dans ce cours de niveau introductif nous traiterons majoritairement d'optimisation continue et nous renvoyons au cours de troisième année [5] pour l'optimisation combinatoire.

Quant aux problèmes de contrôle, on peut également en distinguer deux branches principales selon que l'objectif soit d'amener le système en un état fixé (on parle alors de **contrôlabilité** ou de **commandabilité**), ou de minimiser une fonctionnelle évaluant le coût de l'action sur le système. Ce coût résulte bien souvent d'un compromis entre la réalisation de certaines performances (comme l'atteinte d'une cible ou le fait de s'en rapprocher) et le coût afférent à la réalisation de ce contrôle (et dû par exemple à la consommation énergétique, à la poussée des moteurs, etc.) On parle alors de problèmes de **contrôle optimal**.

Pour finir cette brève introduction nous indiquons le plan de la suite du cours. Le reste de ce chapitre présente à travers des exemples les applications de l'optimisation et du contrôle. Les Chapitres 2, 3 et 4 sont consacrés aux problèmes d'**optimisation**. Le Chapitre 2 va principalement porter sur les aspects théoriques. On y étudiera la question de l'existence et de l'unicité de solutions à des problèmes d'optimisation, que ce soit en dimension finie ou infinie. En particulier, nous verrons le rôle crucial de la **convexité** pour obtenir des résultats d'existence en dimension infinie. Nous y verrons aussi les conditions d'optimalité, reposant sur les dérivées des fonctions optimisées, qui permettent de caractériser les solutions possibles. Le Chapitre 3 développera les algorithmes numériques qui découlent de ces conditions d'optimalité. Le Chapitre 4 traite de la programmation linéaire dans le cas continu ou discret. Dans ce dernier cas, cela constitue une très brève, et très biaisée, introduction aux

méthodes de la recherche opérationnelle. Pour plus de détails sur l'optimisation nous renvoyons le lecteur aux ouvrages [6], [12], [14], [23].

Les Chapitres 5, 6 et 7 sont, quant à eux, consacrés au **contrôle**, en se restreignant au cas des systèmes régis par des équations différentielles ordinaires en temps (le cas du contrôle des systèmes distribués ne sera donc pas abordé dans ce cours introductif). En outre nous considérerons uniquement des systèmes **déterministes** et n'aborderons pas ici le cas (très important en pratique) des systèmes stochastiques comme les systèmes avec bruit. Le Chapitre 5 porte sur la contrôlabilité des systèmes linéaires et non-linéaires. Le résultat phare de ce chapitre est le **critère de Kalman**. Le Chapitre 6 aborde les problèmes de contrôle optimal sous le prisme d'un exemple relativement simple : le système **linaire-quadratique** où la dynamique du système est linéaire et la fonctionnelle de coût quadratique. La relative simplicité du problème nous permettra d'introduire plusieurs notions clés, comme l'**état adjoint**, le **Hamilotnien** et le **feedback** grâce à l'équation de Riccati. Enfin le Chapitre 7 considère le cas général d'une dyamique et d'une fonctionnelle non-linéaires, et le résultat phare est le **principe du minimum** de Pontryaguine.

Le lecteur désireux d'aller plus loin pourra par exemple consulter [27], ou des ouvrages plus spécialisés (en anglais) comme [2], [3], [15], [18], [19], [24], [26] ou [28].

1.2 Exemples en optimisation

Passons en revue quelques problèmes typiques d'optimisation, d'importance pratique ou théorique inégale, mais qui permettent de faire le tour des différentes "branches" de l'optimisation.

Commençons par quelques exemples en **recherche opérationnelle**, c'est-à-dire en optimisation de la gestion ou de la programmation des ressources.

Exemple 1.2.1 (problème de transport) Il s'agit d'un exemple de programme linéaire (ou programmation linéaire). Le but est d'optimiser la livraison d'une marchandise (un problème classique en logistique). On dispose de M entrepôts, indicés par $1 \leq i \leq M$, disposant chacun d'un niveau de stocks s_i. Il faut livrer N clients, indicés par $1 \leq j \leq N$, qui ont commandé chacun une quantité r_j. Le coût de transport unitaire entre l'entrepôt i et le client j est donné par c_{ij}. Les variables de décision sont les quantités v_{ij} de marchandise partant de l'entrepôt i vers le client j. On veut minimiser le coût du transport tout en satisfaisant les commandes des clients (on suppose que $\sum_{i=1}^{M} s_i \geq \sum_{j=1}^{N} r_j$). Autrement dit, on veut résoudre

$$\inf_{(v_{ij})} \left(\sum_{i=1}^{M} \sum_{j=1}^{N} c_{ij} v_{ij} \right)$$

sous les contraintes de limites des stocks et de satisfaction des clients

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

Lorsque les coûts c_{ij} sont les distances entre i et j et que $\sum_{i=1}^{M} s_i = \sum_{j=1}^{N} r_j$, il s'agit du célèbre problème "des déblais et des remblais" de Monge qui a ensuite été généralisé par Kantorovitch (cette théorie du transport, dit optimal, a de très nombreuses applications en gestion, finance, traitement des images, problèmes inverses... et mathématiques !). Nous étudierons au Chapitre 4 la résolution de ce problème de programmation linéaire. •

Exemple 1.2.2 (problème d'affectation) Il s'agit d'un exemple d'optimisation combinatoire ou en variables entières. Imaginez vous à la tête d'une agence matrimoniale... Soit N femmes, indicées par $1 \leq i \leq N$, et N hommes, indicés par $1 \leq j \leq N$. Si la femme i et l'homme j sont d'accord pour se marier leur variable d'accord a_{ij} vaut 1 ; dans le cas contraire elle vaut 0. Le but du jeu est de maximiser le nombre de mariages "satisfaisants" entres ces N femmes et N hommes. Autrement dit, on cherche une permutation σ dans l'ensemble des permutations $\mathcal{S}_N$ de $\{1, ..., N\}$ qui réalise le maximum de

$$\max_{\sigma \in \mathcal{S}_N} \sum_{i=1}^{N} a_{i\sigma(i)}.$$

Une variante consiste à autoriser des valeurs réelles positives de $a_{ij} \in \mathbb{R}^+$. Ce type de problèmes est appelé problème d'affectation (il intervient dans des contextes industriels plus sérieux comme l'affectation des équipages et des avions dans une compagnie aérienne). Bien que ce ne soit pas forcément la meilleure manière de poser le problème, on peut l'écrire sous une forme voisine de l'Exemple 1.2.1. Les variables de décision sont notées v_{ij} qui vaut 1 s'il y a mariage entre la femme i et l'homme j et 0 sinon. On veut maximiser

$$\sup_{(v_{ij})} \left(\sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} v_{ij} \right)$$

sous les contraintes

$$v_{ij} = 0 \text{ ou } 1, \ \sum_{j=1}^{N} v_{ij} \leq 1, \ \sum_{i=1}^{M} v_{ij} \leq 1 \ \text{ pour } 1 \leq i, j \leq N.$$

On pourrait croire que ce problème d'affectation est simple puisqu'il y a un nombre fini de possibilités qu'il "suffit" d'énumérer pour trouver l'optimum. Il s'agit bien sûr d'un leurre car la caractéristique des problèmes combinatoires est leur très grand nombre de combinaisons possibles qui empêche toute énumération exhaustive en pratique. Dans la formulation avec les variables de décision, si on relaxait la contrainte $v_{ij} = 0$ ou 1, en $0 \leq v_{ij} \leq 1$ (ce qui correspondrait à une sorte de polygamie ou mariage à temps partiel !), il s'agirait d'un simple problème de **programmation linéaire**, résolu au Chapitre 4. Le vrai problème correspond à des variables entières, $v_{ij} = 0$ ou 1, ce qui en fait un problème de programmation en nombres entiers qui est plus délicat mais peut encore se résoudre en le voyant comme un problème de

flot sur un graphe (voir la Section 4.3). Nous ne faisons qu'évoquer ces techniques dans ce cours et la résolution effective de ce problème n'est vraiment traitée que dans le cours de troisième année [5]. •

Exemple 1.2.3 (optimisation quadratique à contraintes linéaires) Soit A une matrice carrée d'ordre n, symétrique définie positive. Soit B une matrice rectangulaire de taille $m \times n$. Soit b un vecteur de $\mathbb{R}^n$. On veut résoudre le problème

$$\inf_{x \in \mathrm{Ker} B} \left\{ J(x) = \frac{1}{2} Ax \cdot x - b \cdot x \right\}.$$

La contrainte d'appartenance à $\mathrm{Ker} B$ rend cette minimisation non évidente (voir la Sous-section 2.5.2 pour sa résolution). •

Un autre exemple algébrique simple et d'une portée très générale est le problème de moindres carrés avec ajout éventuel d'une régularisation.

Exemple 1.2.4 (moindres carrés et régularisation) Soit A une matrice réelle d'ordre $p \times n$ et $b \in \mathbb{R}^p$. On considère le problème "aux moindres carrés"

$$\inf_{x \in \mathbb{R}^n} \|Ax - b\|^2. \tag{1.1}$$

Evidemment, si $p = n$ et si la matrice A est inversible, alors ce problème admet comme unique solution $x = A^{-1}b$. Mais lorsque A n'est pas inversible ou même pas carrée, ce problème donne une notion de solution approchée à ce système linéaire (cf. la Sous-section 2.5). Il existe de nombreuses motivations qui conduisent au problème (1.1). Donnons en juste un exemple en terme de problème inverse ou régression linéaire. Supposons que l'on fasse p expériences physiques qui dépendent de n paramètres. Le résultat de la $i-$ème expérience est un nombre b_i et les valeurs des paramètres correspondants sont la $i-$ème ligne de A. On veut expliquer ces résultats par une loi linéaire de coefficients x_j qui prédit "au mieux" les résultats, c'est-à-dire que, pour tout $1 \leq i \leq p$,

$$b_i \approx \sum_{j=1}^{n} a_{ij} x_j.$$

Le problème aux moindres carrés interprète le "au mieux" en minimisant la distance euclidienne entre résultats et prédictions.

Parfois, le nombre de paramètres n est très grand (beaucoup plus grand que p) et on cherche à expliquer les résultats avec un nombre aussi petit que possible de paramètres "vraiment pertinents". On dit que l'on cherche une solution "creuse". On pourrait donc remplacer (1.1) par la minimisation conjointe de l'écart entre mesures et prédictions et le nombre de composantes non nulles du vecteur x

$$\inf_{x \in \mathbb{R}^n} \|Ax - b\|^2 + \ell \|x\|_0, \tag{1.2}$$

où $\ell > 0$ est un coefficient de pondération et $\|x\|_0$ est le nombre de composantes non nulles du vecteur x (on l'appelle parfois norme l^0 de x bien qu'il ne s'agisse

pas d'une norme !). Malheureusement, le problème (1.2) est très difficile à résoudre (parce que de nature combinatoire). Pour cette raison il est remplacé par un autre problème qui fait intervenir la norme l^1 du vecteur x

$$\inf_{x\in\mathbb{R}^n} \|Ax-b\|^2 + \ell\|x\|_1 \quad \text{avec} \quad \|x\|_1 = \sum_{j=1}^{n} |x_j|. \tag{1.3}$$

Il se trouve que la solution de (1.3) est effectivement creuse, avec de nombreuses composantes nulles (ce nombre dépend du coefficient ℓ). Le terme $\|x\|_1$ est appelé terme de régularisation : il existe d'autres types de régularisation et cette idée est cruciale pour la résolution pratique et efficace des problèmes inverses. Le problème (1.3) est connu en statistiques sous le nom de LASSO (least absolute shrinkage and selection operator). •

Considérons un exemple classique en économie.

Exemple 1.2.5 (consommation des ménages) On considère un ménage qui peut consommer n types de marchandise dont les prix forment un vecteur $p \in \mathbb{R}^n_+$. Son revenu à dépenser est un réel $b > 0$, et ses choix de consommation sont supposés être modélisés par une fonction d'utilité $u(x)$ de $\mathbb{R}^n_+$ dans $\mathbb{R}$ (croissante et concave), qui mesure le bénéfice que le ménage tire de la consommation de la quantité x des n marchandises. La consommation du ménage sera le vecteur x^* qui réalisera le maximum de

$$\max_{x\in\mathbb{R}^n_+, x\cdot p\le b} u(x),$$

c'est-à-dire qui maximise l'utilité sous une contrainte de budget maximal (voir la Sous-section 2.6.2 pour la résolution). •

Voici maintenant un exemple issu du domaine de l'apprentissage machine (ou machine learning) qui connait un développement spectaculaire ces dernières années.

Exemple 1.2.6 (apprentissage machine) On dispose d'un très grand nombre de données $(x_i)_{1\le i\le n}$ (des images, du texte, des mesures expérimentales, etc.) caractérisées par des vecteurs $x_i \in \mathbb{R}^d$ et qu'on a déjà classées en les labellisant avec un label y_i qui est très souvent un booléen (ici, -1 ou $+1$) qui donne un type à la donnée x_i (une image de chat, ou pas ; un texte correct ou injurieux ; un courriel normal ou un "spam" ; une expérience physique couronnée de succès ou pas). Comme pour l'Exemple 1.2.4 on introduit une fonction affine, dite de prédiction,

$$h_{w,\tau}(x) = w\cdot x - \tau$$

où $w \in \mathbb{R}^d$ et $\tau \in \mathbb{R}$ sont des paramètres à optimiser. Si on note $\mathrm{sgn}(h)$ la fonction signe (qui retourne la valeur $+1$ si $h > 0$ et -1 si $h < 0$), on souhaite trouver des paramètres (w, τ) tels que la prédiction

$$\mathrm{sgn}\Big(h_{w,\tau}(x_i)\Big) \approx y_i$$

soit la meilleure possible. Une différence importante avec l'approche fondamentalement linéaire des moindres carrés est que, puisque seul le signe de cette fonction de prédiction compte, elle est combinée avec une fonction de "perte", très non-linéaire, comme par exemple la fonction définie sur $\mathbb{R} \times \{-1, +1\}$ par

$$P(h, y) = \log(1 + \exp(-hy))$$

qui, pour $y = -1$ (respectivement, $y = +1$), est convexe et croissante (respectivement convexe et décroissante). Autrement dit, on considère le problème d'optimisation

$$\inf_{w \in \mathbb{R}^d, \tau \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} P(h_{w,\tau}(x_i), y_i). \tag{1.4}$$

Comme pour l'Exemple 1.2.4 des moindres carrés, il est d'usage de régulariser le vecteur des paramètres w et, à l'aide d'un coefficient $\ell > 0$, de considérer une version augmentée de (1.4)

$$\inf_{w \in \mathbb{R}^d, \tau \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} P(h_{w,\tau}(x_i), y_i) + \frac{\ell}{2} \|w\|_2^2 \tag{1.5}$$

où $\|.\|_2$ est la norme euclidienne. On peut aussi remplacer cette norme par une norme l^1 si l'on souhaite trouver des vecteurs w creux. Il s'agit donc d'un problème d'apprentissage supervisé (puisqu'on dispose de labels y_i fournis par un expert pour apprendre le modèle à partir des données x_i) dont la solution optimale (w^*, τ^*) donne un algorithme de classification, c'est-à-dire de prédiction du label pour une nouvelle donnée x grâce à la formule $y = \operatorname{sgn}\Big(h_{w^*,\tau^*}(x)\Big)$. •

Exemple 1.2.7 (Entropie) La notion d'entropie est fondamentale, aussi bien en thermodynamique qu'en physique statistique ou qu'en théorie de l'information. Afin de ne considérer que des problèmes de minimisation, les mathématiciens changent le signe de l'entropie (afin de remplacer sa maximisation par sa minimisation). Ainsi, en théorie de l'information on minimise l'entropie de Shannon

$$\inf_{p \in \mathbb{R}^n_+, \sum_{i=1}^n p_i = 1} \sum_{i=1}^{n} p_i \log p_i,$$

voir l'Exercice 2.5.14 pour la solution. Un autre problème de minimisation d'entropie en théorie cinétique des gaz est proposé à l'Exercice 2.5.15. •

Donnons maintenant un exemple issu du calcul des variations. Il s'agit d'un problème de minimisation d'énergie qui se retrouve dans de très nombreux exemples en physique ou en mécanique. On se place ici dans une situation très simple en une dimension d'espace mais l'approche se généralise aux dimensions supérieures, au prix toutefois de complications techniques certaines.

Exemple 1.2.8 (calcul des variations) Pour fixer les idées, considérons une poutre uni-dimensionnelle qui au repos est représentée par le segment de droite $(0, L)$ où

$L > 0$ est la longueur de la poutre. On note $x \in (0, L)$ les points de ce segment. Sous l'action d'une force $f(x)$, le point x de la poutre se déplace perpendiculairement au segment d'une distance $u(x)$. On peut trouver la position d'équilibre de la poutre en résolvant un problème de minimisation d'énergie. L'énergie mécanique totale se décompose en deux termes : d'une part une énergie de déformation (le terme quadratique en la dérivée $u'(x)$ ci-dessous), d'autre part une énergie potentielle (le terme proportionnel à la force ci-dessous). Autrement dit, il s'agit du problème d'optimisation

$$\inf_{u \in C^1(0,L)} \frac{1}{2} \int_0^L \mu |u'(x)|^2 dx - \int_0^L f(x)\, u(x)\, dx,$$

où $\mu > 0$ est un coefficient de rigidité de la poutre. Si la poutre est encastrée à une de ses extrémités (ou aux deux) on rajoutera les contraintes $u(0) = 0$ et/ou $u(L) = 0$. Plus généralement, on peut vouloir résoudre des problèmes du type

$$\inf_{u \in C^1(0,L)} \int_0^L j\Big(u'(x), u(x)\Big)\, dx \tag{1.6}$$

où $j(\lambda, u)$ est une fonction régulière sur $\mathbb{R} \times \mathbb{R}$. Nous verrons que l'existence d'une solution au problème (1.6) n'est absolument pas une évidence et qu'il faut des conditions particulières sur l'intégrande j pour s'en assurer. Par ailleurs, la définition de l'espace sur lequel on minimise est aussi une question très délicate dont nous ne dirons rien ici et qui conduit à des développements importants en analyse (disons seulement que l'espace $C^1(0, L)$ doit être remplacé par des espaces plus généraux, de type Sobolev, utilisant la théorie des distributions, voir [1], [9], [14]). •

1.3 Exemples en contrôle

Dans ce cours nous considérerons des systèmes dynamiques à d degrés de liberté qui sont régis par un système d'équations différentielles ordinaires en temps où interviennent k fonctions du temps dont la valeur est choisie en vue de contrôler le système. On note $t \in [0,T]$ le temps où $T > 0$ est l'horizon temporel. Cet horizon temporel est en général fixé, mais on pourra également considérer des problèmes de contrôle optimal où T n'est pas fixé comme dans les problèmes de temps-optimalité où on cherche à atteindre une cible en temps optimal. La dynamique du système de contrôle s'écrit sous la forme générale

$$\dot{x}(t) = f(t, x(t), u(t)), \quad \forall t \in [0,T], \tag{1.7}$$

où la fonction $x : [0,T] \to \mathbb{R}^d$, $d \geq 1$, décrit l'état du système, $u : [0,T] \to \mathbb{R}^k$, $k \geq 1$, est le contrôle, et $f : [0,T] \times \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^d$ décrit la dynamique du système. En général, une condition initiale $x(0) = x_0 \in \mathbb{R}^d$ est également prescrite. Ainsi en choisissant un contrôle $u : [0,T] \to \mathbb{R}^k$, on montre sous des hypothèses relativement générales et des arguments de type Cauchy–Lipschitz qu'il existe une unique trajectoire $x : [0,T] \to \mathbb{R}^d$ associée à ce contrôle, au moins si l'horizon temporel n'est pas trop grand dans le cas non-linéaire. Donnons maintenant quelques exemples, tirés de [13].

Exemple 1.3.1 (contrôlabilité d'un tram) On considère un tram se déplaçant le long d'un axe unidirectionnel. L'état du tram est a priori décrit par sa position $X(t)$, et la variable de contrôle u est l'accélération du tram. En écrivant le principe fondamental de la dynamique (on considère une masse unité pour simplifier), il vient

$$\ddot{X}(t) = u(t), \quad \forall t \in [0,T].$$

Cette équation différentielle du second ordre en temps se récrit comme un système d'ordre un en temps en introduisant la vitesse $V(t) := \dot{X}(t)$. On obtient

$$\dot{x}(t) = \underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{=:A} x(t) + \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{=:B} u(t), \text{ en posant } x(t) := \begin{pmatrix} X(t) \\ V(t) \end{pmatrix}.$$

Ce système de contrôle se récrit sous la forme (1.7) avec $f(t,x,u) = Ax + Bu$, si bien qu'il y a $d = 2$ degrés de liberté et $k = 1$ variable de contrôle. En guise de condition initiale, on prescrit la position et la vitesse du tram à $t = 0$. Par exemple, le tram est initialement à l'arrêt ($V(0) = 0$) à la gare située en $X(0) = 0$. En se fixant un horizon temporel $T > 0$, le problème de la contrôlabilité du tram consiste à se demander s'il existe un contrôle $u : [0,T] \to \mathbb{R}$ capable d'amener le tram au temps T en une position X_1 avec une vitesse V_1 (par exemple, X_1 peut être la position de la gare suivante, et dans ce cas on prescrit $V_1 = 0$). Nous verrons dans ce cours que la réponse à cette question est toujours positive. On peut également poser cette question en rajoutant une contrainte sur le contrôle, par exemple que celui-ci soit en tout temps à valeurs dans un ensemble compact, comme $[-1,1]$, ce qui décrit le fait que les moteurs du tram ne peuvent fournir qu'une accélération bornée. On peut également, tout en conservant cette contrainte sur le contrôle, chercher à atteindre la cible le plus rapidement possible. Par exemple, s'il s'agit d'amener le tram d'une gare à la suivante, on s'attend à ce qu'un première phase d'accélération soit suivie par une phase de décélération jusqu'à l'arrêt complet du tram à la prochaine gare. •

Exemple 1.3.2 (système linéaire-quadratique) On considère un système différentiel linéaire avec critère quadratique. Le but est de guider un robot (ou un engin spatial, un véhicule, etc.) afin qu'il suive "au plus près" une trajectoire prédéfinie. L'état du robot à l'instant t est représenté par une fonction $x(t)$ à valeurs dans $\mathbb{R}^d$ (typiquement, la position et la vitesse). On agit sur le robot par l'intermédiaire d'un contrôle $u(t)$ à valeurs dans $\mathbb{R}^M$ (typiquement, la puissance du moteur, la direction des roues, etc.). En présence de forces $f(t) \in \mathbb{R}^d$, les lois de la mécanique conduisent à un système d'équations différentielles ordinaires (supposées linéaires pour simplifier)

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) + f(t) & \forall t \in [0,T], \\ x(0) = x_0, \end{cases} \tag{1.8}$$

où $x_0 \in \mathbb{R}^d$ est l'état initial du système, A et B sont deux matrices constantes de dimensions respectives $d \times d$ et $d \times k$. On note $z(t)$ une trajectoire "cible" et z_T une

position finale "cible". Pour approcher au mieux ces cibles et pour minimiser le coût du contrôle, on introduit trois matrices symétriques positives R, Q, D dont seule R est supposée en plus être définie positive. On définit alors un critère quadratique

$$J(u) = \int_0^T Ru(t)\cdot u(t)dt + \int_0^T Q(x-z)(t)\cdot(x-z)(t)dt + D\,(x(T) - z_T)\cdot(x(T) - z_T)\,.$$

Remarquons que la fonction $x(t)$ dépend de la variable u à travers (1.8). Comme les commandes admissibles sont éventuellement limitées (la puissance d'un moteur est souvent bornée...), on introduit un convexe fermé non vide K de $\mathbb{R}^k$ qui représente l'ensemble des contrôles admissibles. Le problème est donc de résoudre

$$\inf_{u(t)\in K,\ t\in[0,T]} J(u).$$

Il faudra, bien sûr, préciser dans quels espaces fonctionnels on minimise $J(u)$ et on définit la solution x de (1.8) (voir le Chapitre 6 pour la résolution). •

Exemple 1.3.3 (aspirateur robot (système de Dubbins)) Passons à un premier exemple de système de contrôle non-linéaire : un aspirateur robot. L'état du système est décrit par le triplet $(X, Y, \theta) : [0,T] \to \mathbb{R}^3$ ($d = 3$). Le couple (X, Y) repère la position de l'aspirateur dans le plan et θ l'angle des roues par rapport à l'axe des X. L'action sur le système s'exerce par le biais d'une fonction $u : [0,T] \to \mathbb{R}$ ($k = 1$) qui prescrit la vitesse angulaire de l'axe des roues. La dynamique du système est régie par le système différentiel suivant :

$$\begin{cases} \dot{X}(t) = v\cos(\theta(t)), \\ \dot{Y}(t) = v\sin(\theta(t)), \\ \dot{\theta}(t) = u(t), \qquad\qquad \leftarrow \text{action sur le système} \end{cases}$$

où v est la vitesse de l'aspirateur, supposée constante pour simplifier. Comme dans le cas du tram, on peut envisager plusieurs problèmes : celui d'atteindre une cible prescrite, de le faire en un minimum de temps ou encore de minimiser une fonctionnelle de coût résultant d'une pondération (analogue à celle du système liénaire-quadratique) entre l'adéquation à une trajectoire cible et des valeurs modérées pour le contrôle. •

Exemple 1.3.4 (traversée d'un canal (problème de Zermelo)) On considère une barque traversant un canal de largeur ℓ. On considère un repère cartésien où l'axe des X coïncide avec la berge de départ et l'axe des Y est transverse au canal. La barque a une vitesse d'amplitude constante notée v, et le courant a une vitesse $c(Y)$. On suppose que $c(Y) > v$, pour tout $Y \in [0, \ell]$; il s'agit d'une hypothèse dite de courant fort car l'amplitude du courant est toujours supérieure à la vitesse de la barque. La configuration est illustrée à la figure 1.1. Le contrôle est l'angle $u \in [0, 2\pi]$ de la vitesse de la barque par rapport à l'axe des X, la vitesse étant considérée dans le repère du courant. L'état de la barque est décrit par le couple $x := (X, Y) \in \mathbb{R}^2$

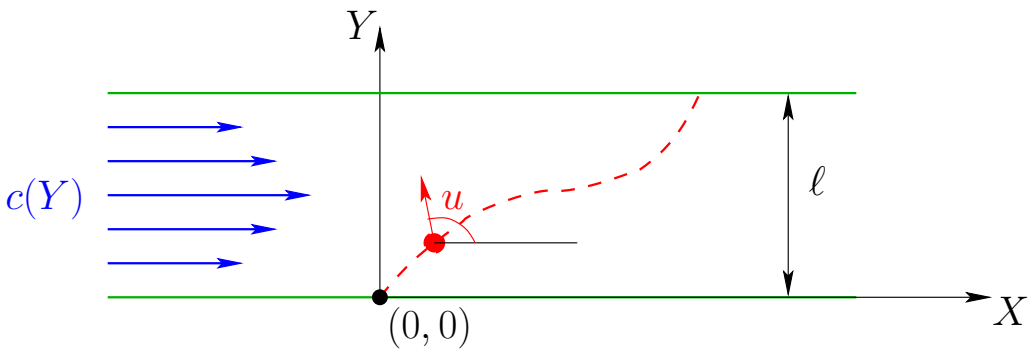

FIGURE 1.1 – Illustration du problème de Zermelo : barque traversant un canal.

donnant les coordonnées de la barque dans le repère fixé. La trajectoire de la barque est régie par la dynamique suivante :

$$\dot{x}(t) = f(x(t), u(t)) = \begin{pmatrix} v\cos(u(t)) + c(Y(t)) \\ v\sin(u(t)) \end{pmatrix}, \quad \forall t \in [0,T], \tag{1.9}$$

et la condition initiale est $x(0) = (0, 0)$. Nous pouvons considérer (entre autres) deux problèmes de contrôle optimal pour atteindre la berge opposée :

1. Minimiser le déport latéral : le critère à minimiser est

   $$J_1(u) = x(T),$$

   et on rajoute la contrainte de cible $Y(T) = \ell$. On obtient le problème de minimisation sous contraintes

   $$\inf_{u(t)\in[0,2\pi],\ t\in[0,T],\ Y(T)=\ell} J_1(u).$$

2. Minimiser le temps de traversée : le critère à minimiser est

   $$J_2(u) = \int_0^T dt = T,$$

   et on a toujours la contrainte de cible $Y(T) = \ell$. On obtient le problème de minimisation sous contraintes

   $$\inf_{u(t)\in[0,2\pi],\ t\in[0,T],\ Y(T)=\ell} J_2(u).$$

On notera que dans ces deux problèmes, le temps T n'est pas fixé et qu'il s'agit d'une inconnue supplémentaire. Ces deux problèmes peuvent être résolus en appliquant le principe du minimum de Pontryaguine que nous verrons dans ce cours. •

Chapitre 2

ASPECTS THÉORIQUES DE L'OPTIMISATION

Dans ce chapitre on introduit les principales notions et résultats d'optimisation. Les questions d'existence de solutions aux problèmes d'optimisation sont traitées dans la Section 2.2 pour le cas de la dimension finie, qui est assez simple, et dans la Section 2.3 pour le cas de la dimension infinie, qui est nettement plus délicat et, pour tout dire, insatisfaisant (autrement dit, très souvent il n'existe pas de solutions à des problèmes d'optimisation dont l'inconnue appartient à un espace de dimension infinie). Afin de pouvoir étudier les conditions d'optimalité qui caractérisent les éventuelles solutions, la Section 2.4 rappelle un certain nombre de résultats sur la différentiabilité des fonctions de plusieurs variables, ou plus généralement définies sur un espace de Hilbert. La Section 2.5 donne la forme des conditions nécessaires d'optimalité dans deux cas essentiels : lorsque l'ensemble des contraintes est convexe on obtient une **inéquation d'Euler** ; lorsqu'il s'agit de contraintes égalités ou inégalités, on obtient une équation faisant intervenir des **multiplicateurs de Lagrange**. La Section 2.6 est consacrée au **théorème de Kuhn et Tucker** qui affirme que, sous certaines hypothèses de convexité, les conditions nécessaires d'optimalité sont aussi suffisantes. On y donne aussi un bref aperçu de la théorie de la **dualité**.

2.1 Définitions et notations

L'optimisation a un vocabulaire particulier : introduisons quelques notations et définitions classiques. Nous considérons principalement des problèmes de minimisation (sachant qu'il suffit d'en changer le signe pour obtenir un problème de maximisation).

Tout d'abord, l'espace dans lequel est posé le problème, noté V, est supposé être un espace vectoriel normé, c'est-à-dire muni d'une norme notée $\|v\|$. Dans la Sous-section 2.2 V sera l'espace $\mathbb{R}^N$, tandis que dans la section suivante V sera un espace de Hilbert réel (on pourrait également considérer le cas, plus général, d'un espace de Banach, c'est-à-dire un espace vectoriel normé complet). On se donne également un sous-ensemble $K \subset V$ où l'on va chercher la solution : on dit que K

est l'ensemble des éléments **admissibles** du problème, ou bien que K définit les **contraintes** s'exerçant sur le problème considéré. Enfin, le **critère**, ou la **fonction coût**, ou la **fonction objectif**, à minimiser, noté J, est une fonction définie sur K à valeurs dans $\mathbb{R}$. Le problème étudié sera donc noté

$$\inf_{v \in K \subset V} J(v). \tag{2.1}$$

Lorsque l'on utilise la notation inf pour un problème de minimisation, cela indique que l'on ne sait pas, a priori, si la valeur du minimum est atteinte, c'est-à-dire s'il existe $\overline{v} \in K$ tel que

$$J(\overline{v}) = \inf_{v \in K \subset V} J(v).$$

Si l'on veut indiquer que la valeur du minimum est atteinte, on utilise de préférence la notation

$$\min_{v \in K \subset V} J(v),$$

mais il ne s'agit pas d'une convention universelle (quoique fort répandue). Pour les problèmes de maximisation, les notations sup et max remplacent inf et min, respectivement. Précisons quelques définitions de base.

Définition 2.1.1 *On dit que u est un minimum (ou un point de minimum) local de J sur K si et seulement si*

$$u \in K \quad et \quad \exists \delta > 0 \ , \ \forall v \in K \ , \ \|v - u\| < \delta \Longrightarrow J(v) \geq J(u) \ .$$

On dit que u est un minimum (ou un point de minimum) global de J sur K si et seulement si

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

Définition 2.1.2 *On appelle infimum de J sur K (ou, plus couramment, valeur minimum), que l'on désigne par la notation (2.1), la borne supérieure dans $\mathbb{R}$ des constantes qui minorent J sur K. Si J n'est pas minorée sur K, alors l'infimum vaut $-\infty$. Si K est vide, par convention l'infimum est $+\infty$.*

Une suite minimisante de J dans K est une suite $(u^n)_{n \in \mathbb{N}}$ telle que

$$u^n \in K \ \forall \, n \quad et \quad \lim_{n \to +\infty} J(u^n) = \inf_{v \in K} J(v).$$

Par la définition même de l'infimum de J sur K il existe toujours des suites minimisantes.

Lemme 2.1.3 *Pour tout problème d'optimisation, si K est non vide, il existe au moins une suite minimisante.*

Démonstration. Notons $m \in \mathbb{R} \cup \{-\infty\}$ la valeur de l'infimum de J sur K. Par définition, pour tout $v \in K$, $J(v) \geq m$. Supposons qu'il n'existe aucune suite $(u^n)_{n \in \mathbb{N}} \in K$ telle que $\lim_{n \to +\infty} J(u^n) = m$. Cela revient à dire qu'il existe $\epsilon > 0$ tel que, pour tout $v \in K$, $J(v) \geq m + \epsilon$. Mais cela est une contradiction avec le fait que m est définie comme la plus grande constante qui minore J sur K. □

2.2 Optimisation en dimension finie

Intéressons nous maintenant à la question de l'**existence de minima** pour des problèmes d'optimisation posés en dimension finie. Nous supposerons dans cette sous-section (sans perte de généralité) que $V = \mathbb{R}^N$ que l'on munit du produit scalaire usuel $u \cdot v = \sum_{i=1}^N u_i v_i$ et de la norme euclidienne $\|u\| = \sqrt{u \cdot u}$.

Un résultat assez général garantissant l'existence d'un minimum est le suivant.

Théorème 2.2.1 (Existence d'un minimum en dimension finie) *Soit K un ensemble fermé non vide de $\mathbb{R}^N$, et J une fonction continue sur K à valeurs dans $\mathbb{R}$ vérifiant la propriété, dite "infinie à l'infini",*

$$\forall (u^n)_{n\geq 0} \text{ suite dans } K \ , \ \lim_{n\to+\infty} \|u^n\| = +\infty \Longrightarrow \lim_{n\to+\infty} J(u^n) = +\infty \ . \tag{2.2}$$

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

Démonstration. Soit (u^n) une suite minimisante de J sur K. La condition (2.2) entraîne que u^n est bornée puisque $J(u^n)$ est une suite de réels majorée. Donc, il existe une sous-suite (u^{n_k}) qui converge vers un point u de $\mathbb{R}^N$. Mais $u \in K$ puisque K est fermé, et $J(u^{n_k})$ converge vers $J(u)$ par continuité, d'où $J(u) = \inf_{v\in K} J(v)$ d'après la Définition 2.1.2. □

Remarque 2.2.2 Notons que la propriété (2.2), qui assure que toute suite minimisante de J sur K est bornée, est automatiquement vérifiée si K est borné. Lorsque l'ensemble K n'est pas borné, cette condition exprime que, dans K, J est **infinie à l'infini**. •

Exercice 2.2.1 Montrer par des exemples que le fait que K est fermé ou que J est continue est en général nécessaire pour l'existence d'un minimum. Donner un exemple de fonction continue et minorée de $\mathbb{R}$ dans $\mathbb{R}$ n'admettant pas de minimum sur $\mathbb{R}$.

Exercice 2.2.2 Montrer que l'on peut remplacer la propriété "infinie à l'infini" (2.2) par la condition plus faible

$$\inf_{v\in K} J(v) < \lim_{R\to+\infty} \Big(\inf_{\|v\|\geq R} J(v) \Big) \ .$$

Exercice 2.2.3 Montrer que l'on peut remplacer la continuité de J par la semi-continuité inférieure de J définie par

$$\forall (u^n)_{n\geq 0} \text{ suite dans } K \ , \ \lim_{n\to+\infty} u^n = u \Longrightarrow \liminf_{n\to+\infty} J(u^n) \geq J(u) \ .$$

Exercice 2.2.4 Montrer qu'il existe un minimum pour les Exemples 1.2.1 et 1.2.3.

2.3 Existence d'un minimum en dimension infinie

2.3.1 Contre-exemples de non-existence

Cette sous-section est consacrée à deux exemples montrant que l'existence d'un minimum en dimension infinie n'est **absolument pas garantie** par des conditions du type de celles utilisées dans l'énoncé du Théorème 2.2.1. Cette difficulté est intimement liée au fait qu'en dimension infinie les fermés bornés ne sont pas compacts !

Commençons par donner un exemple abstrait qui explique bien le mécanisme de "fuite à l'infini" qui empêche l'existence d'un minimum.

Exemple 2.3.1 Soit l'espace de Hilbert (de dimension infinie) des suites de carré sommable dans $\mathbb{R}$

$$\ell_2(\mathbb{R}) = \left\{ x = (x_i)_{i\geq 1} \text{ tel que } \sum_{i=1}^{+\infty} x_i^2 < +\infty \right\},$$

muni du produit scalaire $\langle x, y\rangle = \sum_{i=1}^{+\infty} x_i y_i$. On considère la fonction J définie sur $\ell_2(\mathbb{R})$ par

$$J(x) = \left(\|x\|^2 - 1\right)^2 + \sum_{i=1}^{+\infty} \frac{x_i^2}{i}.$$

Prenant $K = \ell_2(\mathbb{R})$, on considère le problème

$$\inf_{x\in\ell_2(\mathbb{R})} J(x)\ , \tag{2.3}$$

pour lequel nous allons montrer qu'il n'existe pas de point de minimum. Vérifions tout d'abord que

$$\left(\inf_{x\in\ell_2(\mathbb{R})} J(x)\right) = 0.$$

Introduisons la suite x^n dans $\ell_2(\mathbb{R})$ définie par $x_i^n = \delta_{in}$ pour tout $i \geq 1$, où δ_{in} est le symbole de Kronecker qui vaut 1 si $i = n$ et 0 sinon. On vérifie aisément que

$$J(x^n) = \frac{1}{n} \to 0 \quad \text{quand} \quad n \to +\infty.$$

Comme J est positive, on en déduit que x^n est une suite minimisante et que la valeur du minimum est nulle. Cependant, il est évident qu'il n'existe aucun $\overline{x} \in \ell_2(\mathbb{R})$ tel que $J(\overline{x}) = 0$. Par conséquent, il n'existe pas de point de minimum pour (2.3). On voit dans cet exemple que la suite minimisante x^n "part à l'infini" et n'est pas compacte dans $\ell_2(\mathbb{R})$ (bien qu'elle soit bornée). •

Voici maintenant un exemple modèle qui, malgré son caractère simplifié, est très représentatif de problèmes réalistes et pratiques de minimisation d'énergies de transition de phases en science des matériaux.

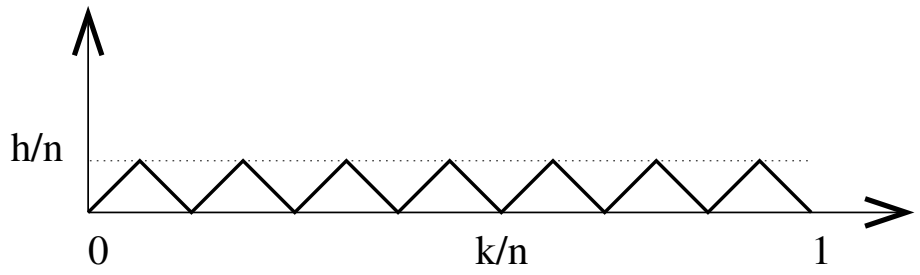

FIGURE 2.1 – Suite minimisante u^n pour l'Exemple 2.3.2.

Exemple 2.3.2 On considère l'espace V des fonctions continues et dérivables par morceaux sur le segment $(0,1)$, muni du produit scalaire

$$\langle u, v\rangle = \int_0^1 (u'(x)v'(x) + u(x)v(x))\, dx$$

et de la norme $\|v\| = \langle v, v\rangle^{1/2}$. On pose $K = V$ et, pour $1 \geq h > 0$, on considère

$$J_h(v) = \int_0^1 \Big((|v'(x)| - h)^2 + v(x)^2\Big) dx \ .$$

L'application J est continue sur V, et la condition (2.2) est vérifiée puisque

$$J_h(v) = \|v\|^2 - 2h\int_0^1 |v'(x)|dx + h^2 \geq \|v\|^2 - \frac{1}{2}\int_0^1 v'(x)^2 dx - h^2 \geq \frac{\|v\|^2}{2} - h^2 \ .$$

Montrons que

$$\inf_{v\in V} J_h(v) = 0 \ , \tag{2.4}$$

ce qui impliquera qu'il n'existe pas de minimum de J_h sur V : en effet, si (2.4) a lieu et si u était un minimum de J_h sur V, on devrait avoir $J_h(u) = 0$, d'où $u \equiv 0$ et $|u'| \equiv h > 0$ (presque partout) sur $(0,1)$, ce qui est impossible.
Pour obtenir (2.4), on construit une suite minimisante (u^n) définie pour $n \geq 1$ par

$$u^n(x) = \begin{cases} h(x - \dfrac{k}{n}) & \text{si } \dfrac{k}{n} \leq x \leq \dfrac{2k+1}{2n} \ , \\ h(\dfrac{k+1}{n} - x) & \text{si } \dfrac{2k+1}{2n} \leq x \leq \dfrac{k+1}{n} \ , \end{cases} \quad \text{pour } 0 \leq k \leq n-1 \ ,$$

comme le montre la Figure 2.1. On voit facilement que $u^n \in V$ et que la dérivée $(u^n)'(x)$ ne prend que deux valeurs : $+h$ et $-h$. Par conséquent, $J_h(u^n) = \int_0^1 u^n(x)^2 dx = \frac{h^2}{4n}$, ce qui prouve (2.4), c'est-à-dire que J_h n'admet pas de point de minimum sur V. Et pourtant, si $h = 0$, il est clair que J_0 admet un unique point de minimum $v \equiv 0$! •

Remarque 2.3.1 Le lecteur attentif pourrait faire remarquer que l'espace V dans l'Exemple 2.3.2 n'est pas un espace de Hilbert. Mais ce n'est pas là que se place la difficulté et le même résultat est vrai si on remplace V par l'espace de Sobolev $H^1(0,1)$ qui, muni du même produit scalaire, est bien un espace de Hilbert de dimension infinie, voir par exemple [1].

A la lumière de ces contre-exemples, examinons la difficulté qui se présente en dimension infinie et sous quelles hypothèses nous pouvons espérer obtenir un résultat d'existence pour un problème de minimisation posé dans un espace de Hilbert de dimension infinie.

Soit V un espace vectoriel de norme $\|v\|$. Soit J une fonction définie sur une partie K de V à valeurs dans $\mathbb{R}$, vérifiant la condition (2.2) (infinie à l'infini). Alors, toute suite minimisante (u^n) du problème

$$\inf_{v\in K} J(v) \tag{2.5}$$

est bornée. En dimension finie (si $V = \mathbb{R}^N$), on conclut aisément comme dans la Sous-section 2.2 en utilisant la compacité des fermés bornés (et en supposant que K est fermé et que J est continue ou semi-continue inférieurement). Malheureusement, un tel résultat est faux en dimension infinie, comme nous venons de le constater. De manière générale, on peut conclure si le triplet (V, K, J) vérifie la condition suivante : pour toute suite $(u^n)_{n\geq 1}$ dans K telle que $\sup_{n\in\mathbb{N}} \|u^n\| < +\infty$ on a

$$\lim_{n\to+\infty} J(u^n) = \ell < +\infty \Longrightarrow \exists\, u \in K \text{ tel que } J(u) \leq \ell\,. \tag{2.6}$$

Ainsi, sous les conditions (2.2) et (2.6), le problème (2.5) admet une solution.

Malheureusement, la condition (2.6) est inutilisable car invérifiable en général ! On peut cependant la vérifier pour une classe particulière de problèmes, très importants en théorie comme en pratique : les problèmes de minimisation **convexe**. Comme nous le verrons dans la Sous-section 2.3.3, si V est un espace de Hilbert, K un **convexe** fermé de V, et que J est une fonction **convexe** et continue sur K, alors (2.6) a lieu et le problème (2.5) admet une solution. Les motivations pour introduire ces conditions sont, d'une part, que les hypothèses de convexité sont souvent naturelles dans beaucoup d'applications, et d'autre part, qu'il s'agit d'une des rares classes de problèmes pour lesquels la théorie est suffisamment générale et complète. Mais ceci ne signifie pas que ces conditions sont les seules qui assurent l'existence d'un minimum ! Néanmoins, en dehors du cadre convexe développé dans les sous-sections suivantes, des difficultés du type de celles rencontrées dans les contre-exemples précédents peuvent survenir.

2.3.2 Analyse convexe

Dans tout ce qui suit, nous supposerons que V est un espace de Hilbert muni d'un produit scalaire $\langle u, v\rangle$ et d'une norme associée $\|v\|$. Rappelons qu'un ensemble K est convexe s'il contient tous les segments reliant deux quelconques de ses points (voir la Définition 8.1.2). Donnons quelques propriétés des fonctions convexes.

Définition 2.3.2 *On dit qu'une fonction J définie sur un ensemble convexe non vide $K \in V$ et à valeurs dans $\mathbb{R}$ est convexe sur K si et seulement si*

$$J(\theta u + (1-\theta)v) \leq \theta J(u) + (1-\theta)J(v) \quad \forall\, u, v \in K\ ,\ \forall\, \theta \in [0,1]\ . \tag{2.7}$$

De plus, J est dite strictement convexe si l'inégalité (2.7) est stricte lorsque $u \neq v$ et $\theta \in]0,1[$.

Remarque 2.3.3 Si J est une application définie sur K à valeurs dans $\mathbb{R}$, on appelle **épigraphe** de J l'ensemble $Epi(J) = \{(\lambda, v) \in \mathbb{R} \times K \ , \ \lambda \geq J(v)\}$. Alors J est convexe si et seulement si $Epi(J)$ est une partie convexe de $\mathbb{R} \times V$. •

Exercice 2.3.1 Soient J_1 et J_2 deux fonctions convexes sur $V, \lambda > 0$, et φ une fonction convexe croissante sur un intervalle de $\mathbb{R}$ contenant l'ensemble $J_1(V)$. Montrer que $J_1 + J_2$, $\max(J_1, J_2)$, λJ_1 et $\varphi \circ J_1$ sont convexes.

Exercice 2.3.2 Soit $(L_i)_{i\in I}$ une famille (éventuellement infinie) de fonctions affines sur V. Montrer que $\sup_{i\in I} L_i$ est convexe sur V. Réciproquement, soit J une fonction convexe continue sur V. Montrer que J est égale au $\sup_{L_i \leq J} L_i$ où les fonctions L_i sont affines.

Pour les fonctions convexes il n'y a pas de différence entre minima locaux et globaux comme le montre le résultat élémentaire suivant.

Proposition 2.3.4 *Si J est une fonction convexe sur un ensemble convexe K, tout point de minimum local de J sur K est un minimum global et l'ensemble des points de minimum est un ensemble convexe (éventuellement vide).*

Si de plus J est strictement convexe, alors il existe au plus un point de minimum.

Démonstration. Soit u un minimum local de J sur K. D'après la Définition 2.1.1, nous pouvons écrire

$$\exists\, \delta > 0 \ , \ \forall\, w \in K \ , \ \|w - u\| < \delta \implies J(w) \geq J(u) \ . \tag{2.8}$$

Soit $v \in K$. Pour $\theta \in]0, 1[$ suffisamment petit, $w_\theta = \theta v + (1-\theta)u$ vérifie $\|w_\theta - u\| < \delta$ et $w_\theta \in K$ puisque K est convexe. Donc, $J(w_\theta) \geq J(u)$ d'après (2.8), et la convexité de J implique que $J(u) \leq J(w_\theta) \leq \theta J(v) + (1-\theta)J(u)$, ce qui montre bien que $J(u) \leq J(v)$, c'est-à-dire que u est un minimum global sur K.

D'autre part, si u_1 et u_2 sont deux minima et si $\theta \in [0, 1]$, alors $w = \theta u_1 + (1-\theta)u_2$ est un minimum puisque $w \in K$ et que

$$\inf_{v\in K} J(v) \leq J(w) \leq \theta J(u_1) + (1-\theta)J(u_2) = \inf_{v\in K} J(v) \ .$$

Le même raisonnement avec $\theta \in]0, 1[$ montre que, si J est strictement convexe, alors nécessairement $u_1 = u_2$. □

Une propriété agréable des fonctions convexes "propres" (c'est-à-dire qui ne prennent pas la valeur $+\infty$) est qu'elles sont continues.

Lemme 2.3.5 *Soit $v_0 \in V$ et J une fonction convexe majorée sur la boule unité de centre v_0. Alors J est continue sur cette boule ouverte.*

Démonstration. Sans perte de généralité, par translation et addition on peut se ramener au cas où $v_0 = 0$ et $J(0) = 0$. Soit $v \neq 0$, $\|v\| < 1$ et M la majoration de J sur la boule unité. Par convexité de J pour $\theta = \|v\|$, on a

$$J(v) = J\left(\theta \frac{v}{\|v\|} + (1-\theta)0\right) \leq \theta J(\frac{v}{\|v\|}) + (1-\theta)J(0) \leq M\|v\|$$

Par ailleurs, par convexité

$$0 = J(0) \leq \frac{1}{1+\|v\|} J(v) + \frac{\|v\|}{1+\|v\|} J(\frac{-v}{\|v\|}) \leq \frac{1}{1+\|v\|}(J(v) + M\|v\|),$$

d'où l'on déduit la continuité en 0

$$|J(v)| \leq M\|v\|.$$

La continuité aux autres points s'obtient par un argument similaire. □

Nous nous servirons par la suite d'une notion de "forte convexité" **plus restrictive** que la stricte convexité.

Définition 2.3.6 *On dit qu'une fonction J définie sur un ensemble convexe K est fortement convexe si et seulement si il existe $\alpha > 0$ tel que, pour tout $u, v \in K$,*

$$J\left(\frac{u+v}{2}\right) \leq \frac{J(u)+J(v)}{2} - \frac{\alpha}{8}\|u-v\|^2 . \tag{2.9}$$

On dit aussi dans ce cas que J est α-convexe.

Dans la Définition 2.3.6, la forte convexité de J n'est testée que pour des combinaisons convexes de poids $\theta = 1/2$. Cela n'est pas une restriction comme le montre l'exercice suivant.

Exercice 2.3.3 Si J est α-convexe, localement majorée, montrer que, pour tout $\theta \in [0,1]$,

$$J(\theta u + (1-\theta)v) \leq \theta J(u) + (1-\theta)J(v) - \frac{\alpha\theta(1-\theta)}{2}\|u-v\|^2 . \tag{2.10}$$

Exercice 2.3.4 Soit A une matrice symétrique d'ordre N et $b \in \mathbb{R}^N$. Pour $x \in \mathbb{R}^N$, on pose $J(x) = \frac{1}{2}Ax \cdot x - b \cdot x$. Montrer que J est convexe si et seulement si A est semi-définie positive, et que J est strictement convexe si et seulement si A est définie positive. Dans ce dernier cas, montrer que J est aussi fortement convexe et trouver la meilleure constante α.

Exercice 2.3.5 Soit V un espace de Hilbert (avec les notations usuelles pour son produit scalaire et sa norme). Soit $L(v)$ une forme linéaire continue sur V, et soit $a(u,v)$ une forme bilinéaire continue et symétrique sur $V \times V$. Soit la fonction J définie sur V par

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

Montrer que J est convexe sur V si a est positive, c'est-à-dire $a(v,v) \geq 0$ pour tout $v \in V$. Montrer que J est fortement convexe sur V si a est coercive, c'est-à-dire s'il existe $C > 0$ tel que $a(v,v) \geq C\|v\|^2$ pour tout $v \in V$.

Le résultat suivant sera essentiel dans l'obtention d'un résultat d'existence d'un minimum en dimension infinie. En particulier, il permet de conclure qu'une fonction J fortement convexe et continue sur un ensemble K convexe fermé non vide est "infinie à l'infini" dans K, c'est-à-dire vérifie la propriété (2.2).

Proposition 2.3.7 *Si J est convexe continue sur un ensemble K convexe fermé non vide, alors il existe une forme linéaire continue $L \in V'$ et une constante $\delta \in \mathbb{R}$ telles que*

$$J(v) \geq L(v) + \delta \quad \forall\, v \in K\ . \tag{2.11}$$

Si de plus J est fortement convexe sur K, alors il existe deux constantes $\gamma > 0$ et $\eta \in \mathbb{R}$ telles que

$$J(v) \geq \gamma \|v\|^2 + \eta \quad \forall\, v \in K\ . \tag{2.12}$$

Démonstration. Prouvons d'abord (2.11). Si J est convexe continue (ou simplement semi-continue inférieurement) sur un ensemble K convexe fermé non vide, alors son épigraphe $Epi(J)$ (défini dans la Remarque 2.3.3) est convexe fermé non vide. Soit $v_0 \in K$ et $\lambda_0 < J(v_0)$. Puisque $(\lambda_0, v_0) \notin Epi(J)$, nous déduisons du Théorème 8.1.12 de séparation d'un point et d'un convexe l'existence de $\alpha, \beta \in \mathbb{R}$ et d'une forme linéaire continue $L \in V'$ tels que

$$\beta\lambda + L(v) > \alpha > \beta\lambda_0 + L(v_0) \quad \forall\, (\lambda, v) \in Epi(J)\ . \tag{2.13}$$

Comme, pour v fixé, on peut prendre λ arbitrairement grand dans le membre de gauche de (2.13), il est clair que $\beta \geq 0$; de plus, comme on peut prendre $v = v_0$ dans le membre de gauche de (2.13), β ne peut être nul. On a donc $\beta > 0$. On déduit alors de (2.13), en choisissant $\lambda = J(v)$, que $J(v) + L(v)/\beta > \alpha/\beta$ pour tout $v \in K$, ce qui prouve (2.11).

Prouvons maintenant (2.12). Soit encore $v_0 \in K$ fixé. Pour tout $v \in K$, (2.9) et (2.11) impliquent que

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

On en déduit

$$J(v) \geq \frac{\alpha}{4}\|v\|^2 - \frac{\alpha}{2}\langle v, v_0\rangle + L(v) + C_1\ ,$$

avec $C_1 = (\alpha/4)\|v_0\|^2 + L(v_0) - J(v_0) + 2\delta$. D'après l'inégalité de Cauchy-Schwarz appliqué à $\langle v, v_0\rangle$ et la continuité de L, i.e. $|L(v)| \leq \|L\|_{V'}\|v\|$ (voir la Définition 8.1.10), il vient

$$J(v) \geq \frac{\alpha}{4}\|v\|^2 - \left(\|L\|_{V'} + \frac{\alpha\|v_0\|}{2}\right)\|v\| + C_1 \geq \frac{\alpha}{8}\|v\|^2 + \eta\ ,$$

pour $\eta \in \mathbb{R}$ bien choisi. □

2.3.3 Résultats d'existence

Nous pouvons maintenant énoncer un premier résultat d'existence de minimum dans le cas particulier où J est fortement convexe (α-convexe).

Théorème 2.3.8 (Existence d'un minimum, cas fortement convexe) *Soit K un convexe fermé non vide d'un Hilbert V et J une fonction α-convexe continue sur K. Alors, il existe un unique minimum u de J sur K et on a*

$$\|v-u\|^2 \leq \frac{4}{\alpha}[J(v)-J(u)] \quad \forall\, v \in K\ . \tag{2.14}$$

En particulier, toute suite minimisante de J sur l'ensemble K converge vers u.

Démonstration. Soit (u^n) une suite minimisante de J sur K. D'après (2.12), $J(v) \geq \delta$ pour tout $v \in K$, c'est-à-dire que J est minorée sur K, donc $\inf_{v\in K} J(v)$ est une valeur finie. Pour $n, m \in \mathbb{N}$ la propriété (2.9) de forte convexité entraîne que

$$\begin{aligned}\frac{\alpha}{8}\|u^n-u^m\|^2 &\leq \frac{\alpha}{8}\|u^n-u^m\|^2 + J\left(\frac{u^n+u^m}{2}\right) - \inf_{v\in K} J(v)\\ &\leq \frac{1}{2}\left(J(u^n)-\inf_{v\in K} J(v)\right) + \frac{1}{2}\left(J(u^m)-\inf_{v\in K} J(v)\right)\ ,\end{aligned}$$

ce qui montre que la suite (u^n) est de Cauchy, et donc converge vers une limite u, qui est nécessairement un minimum de J sur K puisque J est continue et K fermé. L'unicité du point de minimum a été montrée dans la Proposition 2.3.4. Enfin, si $v \in K$, $(u+v)/2 \in K$ car K est convexe, d'où, toujours grâce à (2.9),

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

car $J\left(\dfrac{u+v}{2}\right) \geq J(u)$. □

Il est possible de généraliser en grande partie le Théorème 2.3.8 au cas de fonctions J qui sont seulement convexes (et non pas fortement convexes). Cependant, autant la démonstration du Théorème 2.3.8 est élémentaire, autant celle du théorème suivant est délicate. Elle repose en particulier sur la notion de convergence faible que l'on peut considérer comme "hors-programme" dans le cadre de ce cours.

Théorème 2.3.9 (Existence d'un minimum, cas convexe) *Soit K un convexe fermé non vide d'un espace de Hilbert V, et J une fonction convexe continue sur K, qui est "infinie à l'infini" dans K, c'est-à-dire qui vérifie la condition (2.2), à savoir,*

$$\forall (u^n)_{n\geq 0} \text{ suite dans } K\ ,\ \lim_{n\to+\infty} \|u^n\| = +\infty \Longrightarrow \lim_{n\to+\infty} J(u^n) = +\infty\ .$$

Alors il existe un minimum de J sur K.

Remarque 2.3.10 Le Théorème 2.3.9 donne l'existence d'un minimum comme le précédent Théorème 2.3.8, mais ne dit rien sur l'unicité ni sur l'estimation d'erreur (2.14). Remarquons au passage que (2.14) sera fort utile pour l'étude d'algorithmes numériques de minimisation puisqu'elle fournit une estimation de la vitesse de convergence d'une suite minimisante (u^n) vers le point de minimum u. •

Remarque 2.3.11 Le Théorème 2.3.9 reste vrai si l'on suppose simplement que V est un espace de Banach réflexif (i.e. que le dual de V' est V). •

Nous indiquons brièvement comment on peut démontrer le Théorème 2.3.9 dans le cas d'un espace de Hilbert séparable (c'est-à-dire qui admet une base hilbertienne dénombrable). On définit la notion de **convergence faible** dans V (pour plus de détails à ce sujet, voir les chapitres 3 et 5 de la référence [9]).

Définition 2.3.12 *On dit qu'une suite (u^n) de V converge faiblement vers $u \in V$ si*

$$\forall\, v \in V\ ,\ \lim_{n\to+\infty} \langle u^n, v\rangle = \langle u, v\rangle\ .$$

Soit $(e^i)_{i\geq 1}$ une base hilbertienne de V. Si on note $u_i^n = \langle u^n, e^i\rangle$ les composantes dans cette base d'une suite u^n, uniformément bornée dans V, il est facile de vérifier que la Définition 2.3.12 de la convergence faible est équivalente à la **convergence de toutes les suites de composantes** $(u_i^n)_{n\geq 1}$ pour $i \geq 1$.

Comme son nom l'indique la convergence faible est une notion "plus faible" que la convergence usuelle dans V. En effet, par simple application de l'inégalité de Cauchy-Schwarz, on voit que si u^n converge vers u, c'est-à-dire si $\lim_{n\to+\infty} \|u^n - u\| = 0$, alors u^n converge faiblement vers u. Réciproquement, en dimension infinie il existe des suites qui convergent faiblement mais pas au sens usuel (que l'on appelle parfois "convergence forte" par opposition). Par exemple, la suite $u^n = e^n$ converge faiblement vers zéro, mais pas fortement puisqu'elle est de norme constante égale à 1. L'intérêt de la convergence faible vient du résultat suivant.

Lemme 2.3.13 *De toute suite u^n bornée dans V on peut extraire une sous-suite qui converge faiblement.*

Démonstration. Comme la suite u^n est bornée, chaque suite d'une composante u_i^n est bornée dans $\mathbb{R}$. Pour chaque i, il existe donc une sous-suite, notée $u_i^{n_i}$, qui converge vers une limite u_i. Par un procédé d'extraction diagonale de suites, on obtient alors une sous-suite commune n' telle que, pour tout i, $u_i^{n'}$ converge vers u_i. Ce qui prouve que $u^{n'}$ converge faiblement vers u (on vérifie que $u \in V$). □

Si on appelle "demi-espace fermé" de V tout ensemble de la forme $\{v \in V\ ,\ L(v) \leq \alpha\}$, où L est une forme linéaire continue non identiquement nulle sur V et $\alpha \in \mathbb{R}$, on peut caractériser de façon commode les ensembles convexes fermés.

Lemme 2.3.14 *Une partie convexe fermée K de V est l'intersection des demi-espaces fermés qui contiennent K.*

Démonstration. Il est clair que K est inclus dans l'intersection des demi-espaces fermés qui le contiennent. Réciproquement, supposons qu'il existe un point u_0 de cette intersection qui n'appartiennent pas à K. On peut alors appliquer le Théorème 8.1.12 de séparation d'un point et d'un convexe et construire ainsi un demi-espace fermé qui contient K mais pas u_0. Ceci est une contradiction avec la définition de u_0, donc $u_0 \in K$. □

Lemme 2.3.15 *Soit K un ensemble convexe fermé non vide de V. Alors K est fermé pour la convergence faible.*

De plus, si J est convexe et semi-continue inférieurement sur K (voir l'Exercice 2.2.3 pour cette notion), alors J est aussi semi-continue inférieurement sur K pour la convergence faible.

Démonstration. Par définition, si u^n converge faiblement vers u, alors $L(u^n)$ converge vers $L(u)$. Par conséquent, un demi-espace fermé de V est fermé pour la convergence faible. Le Lemme 2.3.14 permet d'obtenir la même conclusion pour K.

D'après les hypothèses sur J, l'ensemble $Epi(J)$ (défini à la Remarque 2.3.3) est un convexe fermé de $\mathbb{R} \times V$, donc il est aussi fermé pour la convergence faible. On en déduit alors facilement le résultat : si la suite (v^n) tend faiblement vers v dans K, alors $\liminf_{n\to+\infty} J(v^n) \geq J(v)$. □

Nous avons maintenant tous les ingrédients pour finir.

Démonstration du Théorème 2.3.9. D'après (2.2), toute suite minimisante (u^n) est bornée. On déduit alors du Lemme 2.3.13 qu'il existe une sous-suite $(u^{n'})$ convergeant faiblement vers une limite $u \in V$. Mais, d'après le Lemme 2.3.15, $u \in K$ et

$$J(u) \leq \liminf_k J(u^{n_k}) = \inf_{v\in K} J(v) \ .$$

Le point u est donc bien un minimum de u sur K. □

2.4 Différentiabilité

Jusqu'ici nous ne nous sommes intéressés qu'aux questions d'existence de minimum aux problèmes d'optimisation. Mais il importe aussi de caractériser les points de minimum, autant d'un point de vue théorique que pratique, afin de les calculer. Pour ce faire, on utilise des **conditions d'optimalité**, c'est-à-dire des conditions nécessaires et parfois suffisantes de minimalité. Ces conditions d'optimalité s'écrivent avec les dérivées de la fonction objectif et des éventuelles contraintes. Nous allons donc rappeler comment on calcule ces dérivées ou différentielles de fonctions définies sur des espaces de Hilbert.

Avant d'en venir à ces considérations techniques, nous motivons l'étude de ces conditions d'optimalité en rappelant le cas, très simple, du calcul des minima d'une fonction dérivable $J(x)$ définie sur l'intervalle $[a, b] \subset \mathbb{R}$ à valeurs réelles. Il est bien connu que si x_0 est un point de minimum local de J sur l'intervalle $[a, b]$, alors on a

$$J'(x_0) \geq 0 \text{ si } x_0 = a \ , \ J'(x_0) = 0 \text{ si } x_0 \in]a, b[\ , \ J'(x_0) \leq 0 \text{ si } x_0 = b \ .$$

Rappelons la démonstration élémentaire de cette remarque : si $x_0 \in [a,b[$, on peut choisir $x = x_0 + h$ avec $h > 0$ petit et écrire $J(x) \geq J(x_0)$, d'où $J(x_0) + hJ'(x_0) + o(h) \geq J(x_0)$, ce qui donne $J'(x_0) \geq 0$ en divisant par h et en faisant tendre h vers 0. De même obtient-on $J'(x_0) \leq 0$ si $x_0 \in]a,b]$ en considérant $x = x_0 - h$, ce qui permet de conclure.

La stratégie d'obtention et de démonstration des conditions de minimalité est donc claire : on tient compte des contraintes ($x \in [a,b]$ dans l'exemple ci-dessus) pour tester la minimalité de x_0 dans des directions particulières qui respectent les contraintes ($x_0 + h$ avec $h > 0$ si $x_0 \in [a,b[$, $x_0 - h$ avec $h > 0$ si $x_0 \in]a,b]$) : on parlera de **directions admissibles**. On utilise ensuite la définition de la dérivée pour conclure. C'est exactement ce que nous ferons dans la section suivante !

Nous introduisons maintenant la notion de dérivée première d'une fonction $J(u)$ définie sur un espace de Hilbert réel V, et à valeurs dans $\mathbb{R}$. Le produit scalaire dans V est toujours noté $\langle u, v\rangle$ et la norme associée $\|u\|$. Dès que l'espace V n'est plus la droite réelle $\mathbb{R}$ (et même si V est l'espace vectoriel $\mathbb{R}^N$, cas particulièrement simple d'espace de Hilbert), la "bonne" notion théorique de dérivabilité, appelée différentiabilité au sens de Fréchet, est donnée par la définition suivante.

Définition 2.4.1 *On dit que la fonction J, définie sur un voisinage de $u \in V$ à valeurs dans $\mathbb{R}$, est dérivable (ou différentiable) au sens de Fréchet en u s'il existe une forme linéaire continue sur V, $L \in V'$, telle que*

$$J(u+w) = J(u) + L(w) + o(w) \quad , \quad \textit{avec} \quad \lim_{w\to 0} \frac{|o(w)|}{\|w\|} = 0 \, . \tag{2.15}$$

On appelle L la dérivée (ou la différentielle, ou le gradient) de J en u et on note $L = J'(u)$.

Remarque 2.4.2 La Définition 2.4.1 est en fait valable si V est seulement un espace de Banach (on n'utilise pas de produit scalaire dans (2.15)). Cependant, si V est un espace de Hilbert, on peut préciser la relation (2.15) en identifiant V et son dual V' grâce au Théorème de représentation de Riesz 8.1.11. En effet, il existe un unique $p \in V$ tel que $\langle p, w\rangle = L(w)$, donc (2.15) devient

$$J(u+w) = J(u) + \langle p, w\rangle + o(w) \quad , \quad \text{avec} \quad \lim_{w\to 0} \frac{|o(w)|}{\|w\|} = 0 \, . \tag{2.16}$$

On note aussi parfois $p = J'(u)$, ce qui peut prêter à confusion... La formule (2.16) est souvent plus "naturelle" que (2.15), notamment si $V = \mathbb{R}^n$ ou $V = L^2(\Omega)$. •

Dans la plupart des applications, il suffit souvent de déterminer la forme linéaire continue $L = J'(u) \in V'$ car on n'a pas besoin de l'expression explicite de $p = J'(u) \in V$ lorsque V' est identifié à V. En pratique, il est plus facile de trouver l'expression explicite de L que celle de p, comme le montrent les exercices suivants.

Exercice 2.4.1 Montrer que (2.15) implique la continuité de J en u. Montrer aussi que, si deux formes linéaires continues L_1, L_2 vérifient

$$\begin{cases} J(u+w) \geq J(u) + L_1(w) + o(w) \, , \\ J(u+w) \leq J(u) + L_2(w) + o(w) \, , \end{cases} \tag{2.17}$$

alors J est dérivable et $L_1 = L_2 = J'(u)$.

Exercice 2.4.2 (essentiel !) Soit a une forme bilinéaire symétrique continue sur $V \times V$. Soit L une forme linéaire continue sur V. On pose $J(u) = \frac{1}{2}a(u,u) - L(u)$. Montrer que J est dérivable sur V et que $\langle J'(u), w\rangle = a(u,w) - L(w)$ pour tout $u, w \in V$.

Exercice 2.4.3 Soit A une matrice symétrique $N \times N$ et $b \in \mathbb{R}^N$. Pour $x \in \mathbb{R}^N$, on pose $J(x) = \frac{1}{2}Ax \cdot x - b \cdot x$. Montrer que J est dérivable et que $J'(x) = Ax - b$ pour tout $x \in \mathbb{R}^N$.

Exercice 2.4.4 On reprend l'Exercice 2.4.2 avec $V = L^2(\Omega)$ (Ω étant un ouvert de $\mathbb{R}^N$), $a(u,v) = \int_\Omega uv\,dx$, et $L(u) = \int_\Omega fu\,dx$ avec $f \in L^2(\Omega)$. En identifiant V et V', montrer que $J'(u) = u - f$.

Remarque 2.4.3 Il existe d'autres notions de différentiabilité, plus faible que celle au sens de Fréchet. Par exemple, on rencontre souvent la définition suivante. On dit que la fonction J, définie sur un voisinage de $u \in V$ à valeurs dans $\mathbb{R}$, est différentiable au sens de Gâteaux en u s'il existe $L \in V'$ tel que

$$\forall\, w \in V \quad , \quad \lim_{\delta \to 0,\, \delta > 0} \frac{J(u+\delta w) - J(u)}{\delta} = L(w)\ . \tag{2.18}$$

On parle aussi de différentiabilité directionnelle et w est la direction de dérivation dans (2.18). L'intérêt de cette notion est que la vérification de (2.18) est plus aisée que celle de (2.15). Cependant, si une fonction dérivable au sens de Fréchet l'est aussi au sens de Gâteaux, la réciproque est fausse, même en dimension finie, comme le montre l'exemple suivant dans $\mathbb{R}^2$

$$J(x,y) = \frac{x^6}{(y-x^2)^2 + x^8} \quad \text{pour} \quad (x,y) \neq (0,0) \quad , \quad J(0,0) = 0\ .$$

Convenons que, dans ce qui suit, nous dirons qu'une fonction est dérivable lorsqu'elle l'est au sens de Fréchet, sauf mention explicite du contraire. •

Examinons maintenant les propriétés de base des fonctions convexes dérivables.

Proposition 2.4.4 *Soit J une application différentiable de V dans $\mathbb{R}$. Les assertions suivantes sont équivalentes*

$$J \quad \textit{est convexe sur} \quad V\ , \tag{2.19}$$

$$J(v) \geq J(u) + \langle J'(u), v-u\rangle \quad \forall\, u, v \in V\ , \tag{2.20}$$

$$\langle J'(u) - J'(v), u - v\rangle \geq 0 \quad \forall\, u, v \in V\ . \tag{2.21}$$

Proposition 2.4.5 *Soit J une application différentiable de V dans $\mathbb{R}$ et $\alpha > 0$. Les assertions suivantes sont équivalentes*

$$J \quad \textit{est } \alpha\textit{-convexe sur} \quad V\ , \tag{2.22}$$

$$J(v) \geq J(u) + \langle J'(u), v-u\rangle + \frac{\alpha}{2}\|v-u\|^2 \quad \forall\, u, v \in V\ , \tag{2.23}$$

$$\langle J'(u) - J'(v), u - v\rangle \geq \alpha\|u-v\|^2 \quad \forall\, u, v \in V\ . \tag{2.24}$$

Remarque 2.4.6 Les conditions (2.23) et (2.20) ont une interprétation géométrique simple. Elles signifient que la fonction convexe $J(v)$ est toujours au dessus de son plan tangent en u (considéré comme une fonction affine de v). Les conditions (2.24) et (2.21) sont des hypothèses de monotonie ou de croissance de J'. Par ailleurs, si $J(u) = \frac{1}{2}a(u,u) - L(u)$ avec a une forme bilinéaire symétrique continue sur V et L une forme linéaire continue sur V, alors l'Exercice 2.4.2 montre que (2.24) est exactement la définition de la coercivité de a. •

Démonstration. Il suffit de démontrer la Proposition 2.4.5 en observant que le cas $\alpha = 0$ donne la Proposition 2.4.4. Montrons que (2.22) implique (2.23). Comme J est α-convexe, on vérifie aisément (par récurrence) que, pour tout $k \geq 1$,

$$J\left(\left(1-\frac{1}{2^k}\right)u + \frac{1}{2^k}v\right) \leq \left(1-\frac{1}{2^k}\right)J(u) + \frac{1}{2^k}J(v) - \frac{\alpha}{2^{k+1}}\left(1-\frac{1}{2^k}\right)\|u-v\|^2 ,$$

d'où

$$2^k\left[J\left(u+\frac{1}{2^k}(v-u)\right) - J(u)\right] \leq J(v) - J(u) - \frac{\alpha}{2}\left(1-\frac{1}{2^k}\right)\|u-v\|^2 .$$

En faisant tendre k vers $+\infty$, on trouve (2.23). Pour obtenir (2.24) il suffit d'additionner (2.23) avec lui-même en échangeant u et v.

Montrons que (2.24) implique (2.22). Pour $u, v \in V$ et $t \in \mathbb{R}$, on pose $\varphi(t) = J(u + t(v-u))$. Alors φ est dérivable et donc continue sur $\mathbb{R}$, et $\varphi'(t) = \langle J'(u + t(v-u)), v-u\rangle$, de sorte que, d'après (2.24)

$$\varphi'(t) - \varphi'(s) \geq \alpha(t-s)\|v-u\|^2 \quad \text{si} \quad t \geq s . \tag{2.25}$$

Soit $\theta \in]0,1[$. En intégrant l'inégalité (2.25) de $t = \theta$ à $t = 1$ et de $s = 0$ à $s = \theta$, on obtient

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

c'est-à-dire (2.22). □

Exercice 2.4.5 Montrer qu'une fonction J dérivable sur V est strictement convexe si et seulement si

$$J(v) > J(u) + \langle J'(u), v-u\rangle \quad \forall\, u, v \in V \quad \text{avec} \quad u \neq v ,$$

ou encore

$$\langle J'(u) - J'(v), u-v\rangle > 0 \quad \forall\, u, v \in V \quad \text{avec} \quad u \neq v .$$

Terminons cette section en définissant la **dérivée seconde** de J. Remarquons tout d'abord qu'il est très facile de généraliser la Définition 2.4.1 de différentiabilité au cas d'une fonction f définie sur V à valeurs dans un autre espace de Hilbert W (et non pas seulement dans $\mathbb{R}$). On dira que f est différentiable (au sens de Fréchet) en u s'il existe une application linéaire continue L de V dans W telle que

$$f(u+w) = f(u) + L(w) + o(w) \quad , \quad \text{avec} \quad \lim_{w\to 0}\frac{\|o(w)\|_W}{\|w\|_V} = 0 . \tag{2.26}$$

On appelle $L = f'(u)$ la différentielle de f en u. La définition (2.26) est utile pour définir la dérivée de $f(u) = J'(u)$ qui est une application de V dans son dual V'.

Définition 2.4.7 *Soit J une fonction de V dans $\mathbb{R}$. On dit que J est deux fois dérivable en $u \in V$ si J est dérivable dans un voisinage de u et si sa dérivée $J'(u)$ est dérivable en u. On note $J''(u)$ la dérivée seconde de J en u qui vérifie*

$$J'(u+w) = J'(u) + J''(u)w + o(w) \quad , \quad avec \quad \lim_{w\to 0} \frac{\|o(w)\|_{V'}}{\|w\|_V} = 0 \, .$$

Telle qu'elle est définie la dérivée seconde est difficile à évaluer en pratique car $J''(u)w$ est un élément de V'. Heureusement, en la faisant agir sur $v \in V$ on obtient une brave forme bilinéaire continue sur $V \times V$ que l'on notera $J''(u)(w,v)$ en lieu et place de $(J''(u)w)\, v$. Nous laissons au lecteur le soin de prouver le résultat élémentaire suivant.

Lemme 2.4.8 *Si J est une fonction deux fois dérivable de V dans $\mathbb{R}$, elle vérifie*

$$J(u+w) = J(u) + J'(u)w + \frac{1}{2}J''(u)(w,w) + o(\|w\|^2), \ avec \ \lim_{w\to 0} \frac{o(\|w\|^2)}{\|w\|^2} = 0 \, , \quad (2.27)$$

où $J''(u)$ est identifiée à une forme bilinéaire continue sur $V \times V$.

En pratique c'est $J''(u)(w,w)$ que l'on calcule.

Exercice 2.4.6 Soit a une forme bilinéaire symétrique continue sur $V \times V$. Soit L une forme linéaire continue sur V. On pose $J(u) = \frac{1}{2}a(u,u) - L(u)$. Montrer que J est deux fois dérivable sur V et que $J''(u)(v,w) = a(v,w)$ pour tout $u, v, w \in V$. Appliquer ce résultat aux exemples des Exercices 2.4.3, 2.4.4.

Lorsque J est deux fois dérivable on retrouve la condition usuelle de convexité : si la dérivée seconde est positive, alors la fonction est convexe.

Exercice 2.4.7 Montrer que si J est deux fois dérivable sur V les conditions des Propositions 2.4.4 et 2.4.5 sont respectivement équivalentes à

$$J''(u)(w,w) \geq 0 \quad \text{et} \quad J''(u)(w,w) \geq \alpha\|w\|^2 \quad \forall\, u, w \in V \, . \quad (2.28)$$

2.5 Conditions d'optimalité

2.5.1 Inéquations d'Euler et contraintes convexes

Nous commençons par formuler les conditions de minimalité lorsque l'ensemble des contraintes K est convexe, cas où les choses sont plus simples (nous supposerons toujours que K est fermé non vide et que J est continue sur un ouvert contenant K). L'idée essentielle du résultat qui suit est que, pour tout $v \in K$, on peut tester l'optimalité de u dans la "direction admissible" $(v-u)$ car $u + h(v-u) \in K$ si $h \in [0,1]$.

Théorème 2.5.1 (Inéquation d'Euler, cas convexe) *Soit $u \in K$ convexe. On suppose que J est différentiable en u. Si u est un point de minimum local de J sur K, alors*

$$\langle J'(u), v - u \rangle \geq 0 \quad \forall\, v \in K\ . \tag{2.29}$$

Si $u \in K$ vérifie (2.29) et si J est convexe, alors u est un minimum global de J sur K.

Remarque 2.5.2 On appelle (2.29), "inéquation d'Euler". Il s'agit d'une condition **nécessaire** d'optimalité qui devient **nécessaire et suffisante** si J est convexe. La condition (2.29) exprime que la dérivée directionnelle de J au point u dans toutes les directions $(v - u)$, qui sont **rentrantes** dans K, est positive, c'est-à-dire que la fonction J ne peut que croître localement à l'intérieur de K. Il faut aussi remarquer que, dans deux cas importants, (2.29) **se réduit simplement à l'équation d'Euler** $J'(u) = 0$. En premier lieu, si $K = V$, $v - u$ décrit tout V lorsque v décrit V, et donc (2.29) entraîne $J'(u) = 0$. D'autre part, si u est intérieur à K, la même conclusion s'impose. •

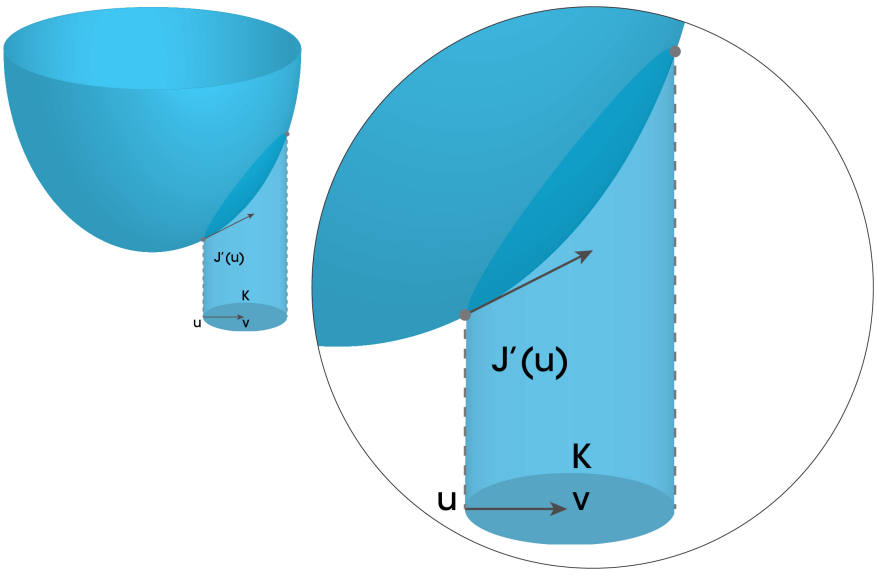

FIGURE 2.2 – Inéquation d'Euler : l'angle entre la dérivée $J'(u)$ et la direction rentrante $(v - u)$ est aigu.

Démonstration. Pour $v \in K$ et $h \in]0, 1]$, $u + h(v - u) \in K$, et donc

$$\frac{J(u + h(v - u)) - J(u)}{h} \geq 0\ . \tag{2.30}$$

On en déduit (2.29) en faisant tendre h vers 0. Le caractère suffisant de (2.29) pour une fonction convexe découle immédiatement de la propriété de convexité (2.20). □

Exercice 2.5.1 Soit K un convexe fermé non vide de V. Pour $x \in V$, on cherche la projection $x_K \in K$ de x sur K (voir le Théorème 8.1.3)

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

Montrer que la condition nécessaire et suffisante (2.29) se ramène exactement à (8.1).

Exercice 2.5.2 On reprend l'Exemple 1.2.4 du problème "aux moindres carrés". Montrer que ce problème admet toujours une solution et écrire l'équation d'Euler correspondante.

Exercice 2.5.3 On reprend l'Exemple 1.2.3

$$\inf_{x \in \mathrm{Ker} B} \left\{ J(x) = \frac{1}{2} Ax \cdot x - b \cdot x \right\}$$

avec A matrice symétrique carrée d'ordre n, et B de taille $m \times n$ ($m \leq n$). Montrer qu'il existe une solution si A est positive et $b \in (\mathrm{Ker} A \cap \mathrm{Ker} B)^{\perp}$, et qu'elle est unique si A est définie positive. Montrer que tout point de minimum $\overline{x} \in \mathbb{R}^n$ vérifie

$$A\overline{x} - b = B^* p \ \text{ avec } p \in \mathbb{R}^m.$$

Exercice 2.5.4 On reprend l'Exemple 1.2.8 et on considère le problème d'optimisation

$$\inf_{u \in K} J(u) = \frac{1}{2} \int_0^L \mu |u'(x)|^2 dx - \int_0^L f(x)\, u(x)\, dx,$$

où $f(x)$ est une fonction continue sur $[0, L]$, $\mu > 0$ et K est défini par

$$K = \{u \in C^1[0, L] \text{ tel que } u(0) = 0 \text{ et } u(L) = 0\}.$$

Vérifier que K est convexe et calculer la dérivée directionnelle $J'(u)(w)$. Montrer que le point de minimum $u_* \in K$ de J (s'il existe et s'il est de classe C^2) vérifie la condition nécessaire suivante

$$\begin{cases} -\mu u_*''(x) = f(x) & \text{si } 0 < x < L, \\ u_*(0) = u_*(L) = 0. \end{cases}$$

En déduire une formule pour ce point de minimum et donc qu'il est unique.

Exercice 2.5.5 Soit K un convexe fermé non vide de V, soit a une forme bilinéaire symétrique continue coercive sur V, et soit L une forme linéaire continue sur V. Montrer que $J(v) = \frac{1}{2} a(v, v) - L(v)$ admet un unique point de minimum dans K, noté u. Montrer que u est aussi l'unique solution du problème (appelé inéquation variationnelle)

$$u \in K \quad \text{et} \quad a(u, v - u) \geq L(v - u) \quad \forall\, v \in K\ .$$

Exercice 2.5.6 Soit J_1 et J_2 deux fonctions convexes continues sur une partie convexe fermée non vide $K \subset V$. On suppose que J_1 seulement est dérivable. Montrer que $u \in K$ est un minimum de $J_1 + J_2$ si et seulement si

$$\langle J_1'(u), v - u\rangle + J_2(v) - J_2(u) \geq 0 \quad \forall\, v \in K\ .$$

Les remarques suivantes, qui sont des applications simples du Théorème 2.5.1, vont nous donner l'intuition de la notion de "multiplicateur de Lagrange" qui sera développée à la Sous-section suivante.

Exercice 2.5.7 On suppose que K est un sous-espace affine fermé de V, $K = u_0 + \mathcal{P}$, avec $u_0 \in V$, où on suppose aussi que $\mathcal{P}$ est un sous-espace vectoriel fermé de V, défini comme une intersection finie d'hyperplans, c'est-à-dire que

$$\mathcal{P} = \{v \in V \quad , \quad \langle a_i, v\rangle = 0 \quad \text{pour} \quad 1 \leq i \leq M\}\ ,$$

où $a_1, \ldots, a_m$ sont donnés dans V. Montrer que la condition d'optimalité (2.29) s'écrit sous la forme

$$u \in K \quad \text{et} \quad \exists\, \lambda_1, \ldots, \lambda_M \in \mathbb{R} \quad , \quad J'(u) + \sum_{i=1}^{M} \lambda_i a_i = 0\ , \tag{2.31}$$

avec des réels λ_i (qui seront appelés multiplicateurs de Lagrange dans le Théorème 2.5.6).

Exercice 2.5.8 Soit $(a_1, \ldots, a_M)$ une famille libre dans V. Supposons que K est défini par

$$K = \{v \in V \quad , \quad \langle a_i, v\rangle \leq 0 \quad \text{pour} \quad 1 \leq i \leq M\}\,.$$

Vérifier que K est un cône convexe fermé, ce qui signifie que K est un ensemble convexe fermé tel que $\lambda v \in K$ pour tout $v \in K$ et tout $\lambda \geq 0$. Montrer que la condition d'optimalité (2.29) implique que

$$\langle J'(u), u\rangle = 0 \quad \text{et} \quad \langle J'(u), w\rangle \geq 0 \quad \forall\, w \in K\ . \tag{2.32}$$

En déduire que si u vérifie (2.29), alors il existe des réels positifs ou nuls $\lambda_i \geq 0$ tels que

$$u \in K \quad \text{et} \quad \exists\, \lambda_1, \ldots, \lambda_M \geq 0 \quad , \quad J'(u) + \sum_{i=1}^{M} \lambda_i a_i = 0\ , \tag{2.33}$$

et, de plus, $\lambda_i = 0$ si $\langle a_i, u\rangle < 0$. Nous verrons que ces réels λ_i sont encore appelés multiplicateurs de Lagrange au Théorème 2.5.16.

Terminons cette sous-section en donnant une **condition d'optimalité du deuxième ordre**.

Proposition 2.5.3 *On suppose que $K = V$ et que J est deux fois dérivable en u. Si u est un point de minimum local de J, alors*

$$J'(u) = 0 \quad et \quad J''(u)(w,w) \geq 0 \quad \forall\, w \in V\ . \tag{2.34}$$

Réciproquement, si, pour tout v dans un voisinage de u,

$$J'(u) = 0 \quad et \quad J''(v)(w,w) \geq 0 \quad \forall\, w \in V\ , \tag{2.35}$$

alors u est un minimum local de J.

Démonstration. Si u est un point de minimum local, on sait déjà que $J'(u) = 0$ et la formule (2.27) nous donne (2.34). Réciproquement, si u vérifie (2.35), on écrit un développement de Taylor à l'ordre deux (au voisinage de zéro) avec reste exact pour la fonction $\phi(t) = J(u + tw)$ avec $t \in \mathbb{R}$ et on en déduit aisément que u est un minimum local de J (voir la Définition 2.1.1). □

2.5.2 Contraintes d'égalité et d'inégalité : multiplicateurs de Lagrange

Cherchons maintenant à écrire des conditions de minimalité lorsque l'ensemble K n'est pas convexe. Plus précisément, nous étudierons des ensembles K définis par des **contraintes d'égalité** ou des **contraintes d'inégalité** (ou les deux à la fois). Nous commençons par une remarque générale sur les **directions admissibles**.

Définition 2.5.4 *En tout point $v \in K$, l'ensemble*

$$K(v) = \left\{ \begin{array}{l} w \in V\ ,\ \exists\, (v^n) \in K^{\mathbb{N}}\ ,\ \exists\, (\varepsilon^n) \in (\mathbb{R}_+^*)^{\mathbb{N}}\ , \\ \lim_{n\to+\infty} v^n = v\ ,\ \lim_{n\to+\infty} \varepsilon^n = 0\ ,\ \lim_{n\to+\infty} \frac{v^n - v}{\varepsilon^n} = w \end{array} \right\}$$

est appelé le cône des directions admissibles au point v.

En termes plus imagés, on peut dire aussi que $K(v)$ est l'ensemble de tous les vecteurs qui sont tangents en v à une courbe contenue dans K et passant par v (si K est une variété régulière, $K(v)$ est simplement l'espace tangent à K en v). Autrement dit, $K(v)$ est l'ensemble de toutes les directions possibles de variations à partir de v qui restent infinitésimalement dans K.

En posant $w^n = (v^n - v)/\varepsilon^n$, on peut aussi dire de façon équivalente que $w \in K(v)$ si et seulement si il existe une suite w^n dans V et une suite ε^n dans $\mathbb{R}_+^*$ telles que

$$\lim_{n\to+\infty} w^n = w\ ,\ \lim_{n\to+\infty} \varepsilon^n = 0 \text{ et } v + \varepsilon^n w^n \in K \quad \forall\, n\ .$$

Il est facile de vérifier que $0 \in K(v)$ (prendre la suite constante $v^n = v$) et que l'ensemble $K(v)$ est un cône, c'est-à-dire que $\lambda w \in K(v)$ pour tout $w \in K(v)$ et tout $\lambda \geq 0$.

Exercice 2.5.9 Montrer que $K(v)$ est un cône fermé et que $K(v) = V$ si v est intérieur à K. Donner un exemple où $K(v)$ est réduit à $\{0\}$.

L'intérêt du cône des directions admissibles réside dans le résultat suivant, qui donne une condition **nécessaire** d'optimalité. La démonstration, très simple, est laissée au lecteur.

Proposition 2.5.5 (Inéquation d'Euler, cas général) *Soit u un minimum local de J sur K. Si J est différentiable en u, on a*

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

Nous allons maintenant préciser la condition nécessaire de la Proposition 2.5.5 dans le cas où K est donné par des **contraintes d'égalité** ou **d'inégalité**. Les résultats que nous obtiendrons vont généraliser ceux des Remarques 2.5.7 et 2.5.8.

Contraintes d'égalité

Dans ce premier cas on suppose que K est donné par

$$K = \{v \in V\ , \quad F(v) = 0\}\ , \tag{2.36}$$

où $F(v) = (F_1(v), ..., F_M(v))$ est une application de V dans $\mathbb{R}^M$, avec $M \geq 1$. La condition **nécessaire** d'optimalité prend alors la forme suivante.

Théorème 2.5.6 *Soit $u \in K$ où K est donné par (2.36). On suppose que J est dérivable en $u \in K$ et que les fonctions $(F_i)_{1\leq i\leq M}$ sont continûment dérivables dans un voisinage de u. On suppose de plus que les vecteurs $\big(F_i'(u)\big)_{1\leq i\leq M}$ sont linéairement indépendants. Alors, si u est un minimum local de J sur K, il existe $\lambda_1, \ldots, \lambda_M \in \mathbb{R}$, appelés* **multiplicateurs de Lagrange**, *tels que*

$$J'(u) + \sum_{i=1}^{M} \lambda_i F_i'(u) = 0\ . \tag{2.37}$$

Démonstration. Montrons d'abord que le cône des directions admissibles $K(u)$ est précisément l'espace tangent à la variété K, définie par (2.36), au point u (voir la Figure 2.3), c'est-à-dire

$$K(u) = \mathrm{Ker}F'(u), \tag{2.38}$$

avec, par définition

$$\mathrm{Ker}F'(u) = \bigcap_{i=1}^{M} [F_i'(u)]^{\perp} = \Big\{w \in V\ , \quad \langle F_i'(u), w\rangle = 0 \quad \text{pour} \quad i = 1, \ldots, M\Big\}.$$

Soit $w \in K(u)$: il existe donc deux suites $v^n \in V$ et $\varepsilon^n > 0$ telles que $F(v^n) = 0$, $\lim_{n\to+\infty} v^n = u$, $\lim_{n\to+\infty} \varepsilon^n = 0$ et $\lim_{n\to+\infty} \frac{v^n - u}{\varepsilon^n} = w$. Un développement de Taylor conduit à

$$0 = F(v^n) = F(u) + \varepsilon^n \langle F'(u), w\rangle + o(\varepsilon^n)$$

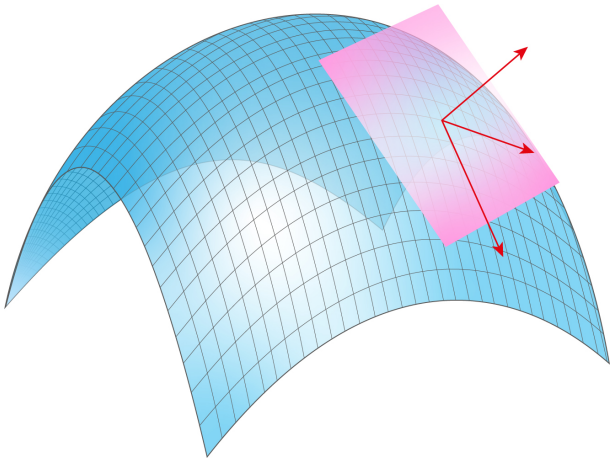

FIGURE 2.3 – Variété K (en bleu) et son hyperplan tangent $K(u)$ en un point u (en rose).

et comme $F(u) = 0$, divisant par ε^n, on obtient

$$0 = \langle F'(u), w\rangle + o(1),$$

c'est-à-dire, à la limite $n \to +\infty$, que w appartient à $\mathrm{Ker}F'(u)$. Pour démontrer l'inclusion inverse, on utilise le Lemme 2.5.7 ci-dessous avec, pour tout $w \in \mathrm{Ker}F'(u)$, $\mathcal{F}(\varepsilon, v) = F(v + \varepsilon w)$. Pour $v_0 = u$, on vérifie sans peine les hypothèses du Lemme 2.5.7, en particulier car la différentielle partielle $D_v\mathcal{F}(0, u) = F'(u)$ est égale à la matrice de lignes $\big(F_i'(u)\big)_{1\le i\le M}$ qui est surjective de V dans $\mathbb{R}^M$ puisque ses lignes sont libres. Pour tout $\varepsilon > 0$, on choisit $v = u + \varepsilon w$ qui vérifie $F(v) = o(\varepsilon)$ puisque $\langle F'(u), w\rangle = 0$. On en déduit donc que la suite v^ε fournie par le Lemme 2.5.7 est telle que $\|v - v_\varepsilon\|_V = o(\varepsilon)$, c'est-à-dire qu'elle est admissible pour la direction w dans la définition de $K(u)$. Par conséquent, $w \in K(u)$, ce qui démontre que $K(u) = \mathrm{Ker}F'(u)$.

Comme nous venons de montrer que $K(u)$ est un espace vectoriel, on peut prendre successivement w et $-w$ dans la Proposition 2.5.5, ce qui conduit à

$$\langle J'(u), w\rangle = 0 \quad \forall\, w \in \mathrm{Ker}F'(u) = \bigcap_{i=1}^{M} [F_i'(u)]^\perp\,,$$

c'est-à-dire que $J'(u)$ est engendré par les $\big(F_i'(u)\big)_{1\le i\le M}$ (notons que les multiplicateurs de Lagrange sont définis de manière unique). Une autre démonstration (plus géométrique) est proposé dans la preuve de la Proposition 2.5.12. □

Lemme 2.5.7 *Soit $\mathcal{F}(\varepsilon, v)$ une fonction de $\mathbb{R}\times V$ dans $\mathbb{R}^M$. On suppose qu'il existe $v_0 \in V$ tel que $\mathcal{F}(0, v_0) = 0$ et que $\mathcal{F}(\varepsilon, v)$ est continûment différentiable dans un voisinage de $(0, v_0)$ (c'est-à-dire différentiable de différentielle continue). Si la différentielle (partielle) $V \ni h \to \langle D_v\mathcal{F}(0, v_0), h\rangle \in \mathbb{R}^M$ est surjective, alors il existe un autre voisinage $\mathcal{V}$ de $(0, v_0)$ et une constante $C > 0$ tels que pour tout $(\varepsilon, v) \in \mathcal{V}$ il existe $v_\varepsilon \in V$ qui vérifie*

$$\mathcal{F}(\varepsilon, v_\varepsilon) = 0 \quad \text{et} \quad \|v - v_\varepsilon\|_V \le C\|\mathcal{F}(\varepsilon, v)\|_{\mathbb{R}^M}.$$

Une démonstration du Lemme 2.5.7 peut être trouvée, dans un cadre plus général (théorème de l'application surjective), dans [4].

Remarque 2.5.8 Lorsque les vecteurs $\left(F_i'(u)\right)_{1\leq i\leq M}$ sont linéairement indépendants (ou libres), on dit que l'on est dans un **cas régulier**. Dans le cas contraire, on parle de **cas non régulier** et la conclusion du Théorème 2.5.6 est fausse comme le montre l'exemple suivant.

Prenons $V=\mathbb{R}$, $M=1$, $F(v)=v^2$, $J(v)=v$, d'où $K=\{0\}$, $u=0$, $F'(u)=0$: il s'agit donc d'un cas non régulier. Comme $J'(u)=1$, (2.37) n'a pas lieu. •

Pour bien comprendre la portée du Théorème 2.5.6, nous l'appliquons sur l'Exemple 1.2.3

$$\min_{x\in \mathrm{Ker}B}\left\{J(x)=\frac{1}{2}Ax\cdot x-b\cdot x\right\},$$

où A est symétrique définie positive d'ordre n, et B de taille $m\times n$ avec $m\leq n$. On note $(b_i)_{1\leq i\leq m}$ les m lignes de B et on a donc m contraintes $b_i\cdot x=0$. Pour simplifier on suppose que le rang de B est m, c'est-à-dire que les vecteurs (b_i) sont libres. Si le rang de B est $m'<m$, alors $(m-m')$ lignes de B sont engendrées par m' autres lignes libres de B. Il y a donc $(m-m')$ contraintes redondantes que l'on peut éliminer et on se ramène au cas d'une matrice B' de taille $m'\times n$ et de rang maximal m'. Comme le rang de B est m, les (b_i) sont libres et on peut appliquer la conclusion (2.37). Il existe donc un multiplicateur de Lagrange $p\in\mathbb{R}^m$ tel que un point de minimum $\overline{x}$ vérifie

$$A\overline{x}-b=\sum_{i=1}^{m}p_ib_i=B^*p.$$

Comme A est inversible, on en déduit la valeur $\overline{x}=A^{-1}(b+B^*p)$. Par ailleurs $B\overline{x}=0$ et, comme B est de rang maximal, la matrice $BA^{-1}B^*$ est inversible, ce qui conduit à

$$p=-\left(BA^{-1}B^*\right)^{-1}BA^{-1}b\quad\text{et}\quad\overline{x}=A^{-1}\left(\mathrm{Id}-B^*\left(BA^{-1}B^*\right)^{-1}BA^{-1}\right)b.$$

Notons que le multiplicateur de Lagrange p est unique. Si B n'est pas de rang m, l'Exercice 2.5.3 montre qu'il existe quand même p solution de $BA^{-1}B^*p=-BA^{-1}b$ mais qui n'est unique qu'à l'addition d'un vecteur du noyau de B^* près.

Exercice 2.5.10 Généraliser les résultats ci-dessus pour cette variante de l'Exemple 1.2.3

$$\min_{Bx=c}\left\{J(x)=\frac{1}{2}Ax\cdot x-b\cdot x\right\},$$

où $c\in\mathbb{R}^m$ est un vecteur donné.

Exercice 2.5.11 Soit A une matrice carrée d'ordre n, symétrique. On veut caractériser et calculer les solutions de

$$\inf_{x\in\mathbb{R}^n,\|x\|=1}J(x)=Ax\cdot x,$$

où $\|x\|$ est la norme euclidienne de x. Appliquer le Théorème 2.5.6 et en déduire que les points de minimum de J sur la sphère unité sont des vecteurs propres de A associés à la plus petite valeur propre.

Exercice 2.5.12 Selon la légende rapportée par Virgile dans l'Énéide, la fondation de la ville de Carthage par la reine Didon conduisit à un problème typique du calcul des variations. Il s'agissait de trouver la plus grande surface possible (pour la ville) s'appuyant sur le rivage (une droite) et de frontière terrestre de longueur donnée. La réponse est intuitivement un demi disque. Mathématiquement, le problème est de trouver la courbe plane, définie par le graphe d'une fonction $y(x)$ pour $x \in (0, \xi)$, de longueur fixée $l \geq 0$ qui enclôt avec le segment $(0, \xi)$ reliant ses deux extrémités l'aire maximum. Autrement dit, on résout

$$\sup_{\xi \geq 0, y \in C^1[0,\xi]} \int_0^\xi y(x)\, dx,$$

avec les contraintes

$$y(0) = y(\xi) = 0, \ \int_0^\xi \sqrt{1 + y'(x)^2} dx = l.$$

La longueur ξ du segment est une variable d'optimisation, de même que la fonction $y(x)$ qui donne la position de la courbe au dessus du point x du segment. En s'inspirant de l'Exercice 2.5.4, calculer les dérivées directionnelles de la fonction objectif et de la contrainte intégrale. En déduire que la solution du problème de Didon est nécessairement un arc de cercle.

Exercice 2.5.13 Soit A une matrice $n \times n$ symétrique définie positive et $b \in \mathbb{R}^n$ non nul.

1. Montrer que les problèmes

$$\sup_{Ax \cdot x \leq 1} b \cdot x \quad \text{et} \quad \sup_{Ax \cdot x = 1} b \cdot x$$

   sont équivalents et qu'ils ont une solution. Utiliser le Théorème 2.5.6 pour calculer cette solution et montrer qu'elle est unique.

2. On introduit un ordre partiel dans l'ensemble des matrices symétriques définies positives d'ordre n en disant que $A \geq B$ si et seulement si $Ax \cdot x \geq Bx \cdot x$ pour tout $x \in \mathbb{R}^n$. Déduire de la question précédente que, si $A \geq B$, alors $B^{-1} \geq A^{-1}$.

Exercice 2.5.14 Montrer que l'entropie de Shannon de l'Exemple 1.2.7 admet un unique point de minimum que l'on calculera. Montrer aussi que, pour tout $p \in \mathbb{R}^n_+$ tel que $\sum_{i=1}^n p_i = 1$,

$$-\sum_{i=1}^n p_i \log p_i = \inf_{q \in \mathbb{R}^n_+, \sum_{i=1}^n q_i = 1} -\sum_{i=1}^n p_i \log q_i.$$

Exercice 2.5.15 En théorie cinétique des gaz les molécules de gaz sont représentées en tout point de l'espace par une fonction de répartition $f(v)$ dépendant de la vitesse microscopique $v \in \mathbb{R}^N$. Les quantités macroscopiques, comme la densité du gaz ρ, sa vitesse u, et sa température T, se retrouvent grâce aux moments de la fonction $f(v)$

$$\rho = \int_{\mathbb{R}^N} f(v)\,dv\ , \quad \rho u = \int_{\mathbb{R}^N} v\,f(v)\,dv\ , \quad \frac{1}{2}\rho u^2 + \frac{N}{2}\rho T = \frac{1}{2}\int_{\mathbb{R}^N} |v|^2 f(v)\,dv\ . \tag{2.39}$$

Boltzmann a introduit l'entropie cinétique $H(f)$ définie par

$$H(f) = \int_{\mathbb{R}^N} f(v)\log\big(f(v)\big)\,dv\ .$$

Montrer que H est strictement convexe sur l'espace des fonctions $f(v) > 0$ mesurables telle que $H(f) < +\infty$. On minimise H sur cet espace sous les contraintes de moment (2.39), et on admettra qu'il existe un unique point de minimum $M(v)$. Montrer que ce point de minimum est une Maxwellienne définie par

$$M(v) = \frac{\rho}{(2\pi T)^{N/2}}\exp\left(-\frac{|v-u|^2}{2T}\right).$$

On peut obtenir un nouvel éclairage sur le Théorème 2.5.6 en introduisant la notion de Lagrangien.

Définition 2.5.9 *On appelle* **Lagrangien** *du problème de minimisation de J sur K, défini par (2.36), la fonction de deux variables définie sur $V \times \mathbb{R}^M$ par*

$$\mathcal{L}(v,\mu) = J(v) + \sum_{i=1}^{M} \mu_i F_i(v) = J(v) + \mu \cdot F(v).$$

La nouvelle variable $\mu \in \mathbb{R}^M$ est appelée **multiplicateur de Lagrange** *pour la contrainte $F(v) = 0$.*

Si $u \in K$ est un minimum local de J sur K, le Théorème 2.5.6 nous dit alors que, dans le cas régulier, il existe $\lambda \in \mathbb{R}^M$ tel que

$$\frac{\partial \mathcal{L}}{\partial v}(u,\lambda) = 0 \quad , \quad \frac{\partial \mathcal{L}}{\partial \mu}(u,\lambda) = 0\ ,$$

puisque $\dfrac{\partial \mathcal{L}}{\partial \mu}(u,\lambda) = F(u) = 0$ si $u \in K$ et $\dfrac{\partial \mathcal{L}}{\partial v}(u,\lambda) = J'(u) + \lambda F'(u) = 0$ d'après (2.37). On peut ainsi écrire la contrainte et la condition d'optimalité comme l'annulation du gradient (la stationnarité) du Lagrangien.

Une autre propriété importante du Lagrangien est la suivante (qui permet de faire "disparaître" la contrainte au prix de l'ajout d'une variable).

Lemme 2.5.10 *Le problème de minimisation sous contrainte est équivalent à un problème de min-max*

$$\inf_{v\in V,\ F(v)=0} J(v) = \inf_{v\in V}\sup_{\mu\in\mathbb{R}^M} \mathcal{L}(v,\mu). \tag{2.40}$$

Démonstration. Si $F(v) = 0$ on a évidemment $F(v) = \mathcal{L}(v, \mu)$ pour tout $\mu \in \mathbb{R}^M$, tandis que, si $F(v) \neq 0$, alors $\sup_{\mu \in \mathbb{R}^M} \mathcal{L}(v, \mu) = +\infty$, d'où l'on déduit l'égalité (2.40). □

Remarque 2.5.11 On pourrait croire que le Lemme 2.5.10 est seulement une astuce ou un jeu d'écriture pour faire disparaitre la contrainte de manière artificielle. Il n'en est rien et la notion de Lagrangien est extrêmement utile pour les algorithmes numériques de minimisation. Donnons en un rapide aperçu. Nous verrons que tous les algorithmes d'optimisation sont itératifs, c'est-à-dire que l'on construit une suite de solutions approchées u^n qui convergerait vers une solution u. A chaque itération, on définit la nouvelle itérée u^{n+1} à partir de la précédente u^n. En général, il est très difficile, voire impossible, de garantir que les solutions approchées u^n vérifient exactement les contraintes $F(u^n) = 0$. L'utilisation du Lagrangien permet de contourner cette difficulté. Imaginons, par exemple dans le cas d'une seule contrainte ($M = 1$), que $F(u^n) > 0$. Alors, si on choisit un multiplicateur de Lagrange $\mu^n > 0$, la minimisation (totale ou partielle) sans contrainte de $\mathcal{L}(v, \mu^n)$ permet d'obtenir une nouvelle itérée u^{n+1} qui minimise au mieux la fonction objectif et la violation de la contrainte (la positivité de μ^n conduit à minimiser $F(v)$, comme $J(v)$). Inversement, si on avait $F(u^n) < 0$, on aurait choisi un multiplicateur de Lagrange $\mu^n < 0$ et la minimisation sans contrainte de $\mathcal{L}(v, \mu^n)$ conduirait à maximiser $F(v)$ (tout en minimisant toujours $J(v)$), c'est-à-dire à réduire la violation de la contrainte. Tout ceci sera précisé dans la Sous-section 3.3.2 lorsque sera présenté l'algorithme d'Uzawa. •

Nous donnons maintenant une condition **nécessaire** d'optimalité du deuxième ordre.

Proposition 2.5.12 *On se place sous les hypothèses du Théorème 2.5.6 et on suppose que les fonctions J et $F_1, ..., F_M$ sont deux fois continûment dérivables et que les vecteurs $\big(F_i'(u)\big)_{1 \leq i \leq M}$ sont linéairement indépendants. Soit $\lambda \in \mathbb{R}^M$ le multiplicateur de Lagrange défini par le Théorème 2.5.6. Alors tout minimum local u de J sur K vérifie*

$$\left(J''(u) + \sum_{i=1}^{M} \lambda_i F_i''(u) \right)(w, w) \geq 0 \quad \forall\, w \in K(u) = \bigcap_{i=1}^{M} [F_i'(u)]^{\perp}. \tag{2.41}$$

Démonstration. Supposons qu'il existe un chemin admissible de classe C^2, c'est-à-dire une fonction $t \to u(t)$ de $[0, 1]$ dans V telle que $u(0) = u$ et $F(u(t)) = 0$ pour tout $t \in [0, 1]$. Par définition, la dérivée $u'(0)$ appartient au cône des directions admissibles $K(u)$. On pose

$$j(t) = J\big(u(t)\big) \quad \text{et} \quad f_i(t) = F_i\big(u(t)\big) \ \text{ pour } 1 \leq i \leq M.$$

En dérivant on obtient

$$j'(t) = \langle J'\big(u(t)\big), u'(t) \rangle \quad \text{et} \quad f_i'(t) = \langle F_i'\big(u(t)\big), u'(t) \rangle \ \text{ pour } 1 \leq i \leq M,$$

et

$$\begin{aligned} j''(t) &= J''\big(u(t)\big)\big(u'(t),u'(t)\big) + \langle J'\big(u(t)\big), u''(t)\rangle \\ f_i''(t) &= F_i''\big(u(t)\big)\big(u'(t),u'(t)\big) + \langle F_i'\big(u(t)\big), u''(t)\rangle \quad \text{pour } 1 \le i \le M. \end{aligned}$$

Comme $f_i(t) = 0$ pour tout t et puisque 0 est un minimum de $j(t)$, on en déduit $j'(0) = 0$, $j''(0) \ge 0$, et $f_i'(0) = f_i''(0) = 0$. Les conditions $f_i'(0) = 0$ nous disent que $u'(0)$ est orthogonal au sous-espace engendré par $\big(F_i'(u)\big)_{1\le i\le M}$ (qui est égal à $K(u)$ quand cette famille est libre), tandis que $j'(0) = 0$ signifie que $J'(u)$ est orthogonal à $u'(0)$. Si $u'(0)$ décrit tout $K(u)$ lorsque on fait varier les chemins admissibles, on en déduit que $J'(u)$ et les $F_i'(u)$ appartiennent au même sous-espace (l'orthogonal de $K(u)$). On retrouve ainsi la condition du premier ordre : il existe $\lambda \in \mathbb{R}^M$ tel que

$$J'(u) + \sum_{i=1}^{M} \lambda_i F_i'(u) = 0. \tag{2.42}$$

Les conditions $f_i''(0) = 0$ impliquent que

$$0 = \sum_{i=1}^{M} \lambda_i \Big(F_i''(u)\big(u'(0),u'(0)\big) + \langle F_i'(u), u''(0)\rangle \Big),$$

tandis que $j''(0) \ge 0$ donne

$$J''(u)\big(u'(0),u'(0)\big) + \langle J'(u), u''(0)\rangle \ge 0.$$

Grâce à (2.42) on peut éliminer les dérivées premières et $u''(0)$ pour obtenir (en sommant les deux dernières équations)

$$\left(\sum_{i=1}^{M} \lambda_i F_i''(u) + J''(u) \right) \big(u'(0),u'(0)\big) \ge 0,$$

qui n'est rien d'autre que (2.41) lorsque $u'(0)$ parcourt $K(u)$.

L'existence de tels chemins admissibles $u(t)$ et le fait que l'ensemble des $u'(0)$ décrit la totalité du cône des directions admissibles $K(u)$ est une conséquence du théorème des fonctions implicites que l'on peut appliquer grâce à l'hypothèse que la famille $\big(F_i'(u)\big)_{1\le i\le M}$ est libre (voir par exemple [4]). □

Exercice 2.5.16 Calculer la condition nécessaire d'optimalité du second ordre pour l'Exemple 1.2.3 et l'Exercice 2.5.11.

Contraintes d'inégalité

Dans ce deuxième cas on suppose que K est donné par

$$K = \{v \in V \ , \quad F_i(v) \le 0 \quad \text{pour} \quad 1 \le i \le M\} \ , \tag{2.43}$$

où $F_1, \ldots, F_M$ sont des fonctions continues de V dans $\mathbb{R}$. Lorsque l'on veut déterminer le cône des directions admissibles $K(v)$, la situation est un peu plus compliquée que précédemment car toutes les contraintes dans (2.43) ne jouent pas le même rôle selon le point v où l'on calcule $K(v)$. En effet, si $F_i(v) < 0$, il est clair que, pour toute direction $w \in V$ et pour ε suffisamment petit, on aura aussi $F_i(v+\varepsilon w) \leq 0$ (on dit que la contrainte i est inactive en v). Par contre, si $F_i(v) = 0$, il faudra imposer des conditions sur le vecteur $w \in V$ pour que, pour tout $\varepsilon > 0$ suffisamment petit, $F_i(v + \varepsilon w) \leq 0$. Afin que toutes les contraintes dans (2.43) soient satisfaites pour $(v + \varepsilon w)$ il va donc falloir imposer des conditions sur w, appelées **conditions de qualification**. Grosso modo, ces conditions vont garantir que l'on peut faire des "variations" autour d'un point v afin de tester son optimalité. Il existe différents types de conditions de qualification (plus ou moins sophistiquées et générales). Nous allons donner une définition dont le principe est de regarder sur le problème **linéarisé** s'il est possible de faire des variations respectant les contraintes linéarisées. Ces considérations de "calcul des variations" motivent les définitions suivantes.

Définition 2.5.13 *Soit $u \in K$. L'ensemble $I(u) = \{i \in \{1, \ldots, M\}\ ,\ F_i(u) = 0\}$ est appelé l'ensemble des contraintes* **actives** *en u.*

Définition 2.5.14 *On dit que les contraintes (2.43) sont* **qualifiées** *en $u \in K$ si et seulement si il existe une direction $\overline{w} \in V$ telle que l'on ait pour tout $i \in I(u)$*

$$\begin{array}{ll} \textit{ou bien} & \langle F_i'(u), \overline{w} \rangle < 0\ , \\ \textit{ou bien} & \langle F_i'(u), \overline{w} \rangle = 0 \quad \textit{et} \quad F_i \quad \textit{est affine}\ . \end{array} \tag{2.44}$$

Remarque 2.5.15 La direction $\overline{w}$ est en quelque sorte une "direction rentrante" puisque on déduit de (2.44) que $u+\varepsilon\overline{w} \in K$ pour tout $\varepsilon \geq 0$ suffisamment petit. Bien sûr, si toutes les fonctions F_i sont affines, on peut prendre $\overline{w} = 0$ et les contraintes sont automatiquement qualifiées. Le fait de distinguer les contraintes affines dans la Définition 2.5.14 est justifié non seulement parce que celles-ci sont qualifiées sous des conditions moins strictes, mais surtout en regard de l'importance des contraintes affines dans les applications (comme le montre les exemples du Chapitre 1). •

Nous pouvons alors énoncer les conditions **nécessaires** d'optimalité sur l'ensemble (2.43).

Théorème 2.5.16 *On suppose que K est donné par (2.43), que les fonctions J et $F_1, \ldots, F_M$ sont dérivables en u et que les contraintes sont qualifiées en u. Alors, si u est un minimum local de J sur K, il existe $\lambda_1, \ldots, \lambda_M \geq 0$, appelés multiplicateurs de Lagrange, tels que*

$$J'(u) + \sum_{i=1}^{M} \lambda_i F_i'(u) = 0, \quad \lambda_i \geq 0, \quad \lambda_i = 0 \ \textit{si}\ F_i(u) < 0 \quad \forall\, i \in \{1, \ldots, M\}. \tag{2.45}$$

Remarque 2.5.17 On peut réécrire la condition (2.45) sous la forme suivante

$$J'(u) + \sum_{i=1}^{M} \lambda_i F_i'(u) = 0\ , \quad \lambda \geq 0 \quad , \quad \lambda \cdot F(u) = 0\ ,$$

où $\lambda \geq 0$ signifie que chacune des composantes du vecteur $\lambda = (\lambda_1, ..., \lambda_M)$ est positive, puisque, pour tout indice $i \in \{1, \ldots, M\}$, on a soit $F_i(u) = 0$, soit $\lambda_i = 0$. Le fait que $\lambda \cdot F(u) = 0$ est appelée condition des écarts complémentaires. •

Démonstration. Considérons tout d'abord l'ensemble

$$\tilde{K}(u) = \{w \in V \quad , \quad \langle F_i'(u), w\rangle \leq 0 \quad \forall\, i \in I(u)\} \ . \tag{2.46}$$

(On peut montrer que $\tilde{K}(u)$ n'est autre que le cône $K(u)$ des directions admissibles, voir [4]). Soit $\overline{w}$ une direction admissible satisfaisant (2.44), $w \in \tilde{K}(u)$, et un réel $\delta > 0$. Nous allons montrer que $u + \varepsilon(w + \delta\overline{w}) \in K$ pour tout réel $\varepsilon > 0$ assez petit. Il faut examiner trois cas de figure.

1. Si $i \notin I(u)$, on a $F_i(u) < 0$ et $F_i(u + \varepsilon(w + \delta\overline{w})) < 0$ par continuité de F_i si ε est assez petit.
2. Si $i \in I(u)$ et $\langle F_i'(u), \overline{w}\rangle < 0$, alors

$$\begin{aligned} F_i(u + \varepsilon(w + \delta\overline{w})) &= F_i(u) + \varepsilon\langle F_i'(u), w + \delta\overline{w}\rangle + o(\varepsilon) \\ &\leq \varepsilon\delta\langle F_i'(u), \overline{w}\rangle + o(\varepsilon) \ < \ 0 \ , \end{aligned} \tag{2.47}$$

   pour $\varepsilon > 0$ assez petit.
3. Enfin, si $i \in I(u)$ et $\langle F_i'(u), \overline{w}\rangle = 0$, alors F_i est affine et

$$F_i(u + \varepsilon(w + \delta\overline{w})) = F_i(u) + \varepsilon\langle F_i'(u), w + \delta\overline{w}\rangle = \varepsilon\langle F_i'(u), w\rangle \leq 0 \ . \tag{2.48}$$

Finalement, si u est un minimum local de J sur K, on déduit de ce qui précède que

$$\langle J'(u), w + \delta\overline{w}\rangle \geq 0 \quad \forall\, w \in \tilde{K}(u) \quad , \quad \forall\, \delta \in \mathbb{R}_+^* \ .$$

En faisant tendre δ vers 0, on obtient $\langle J'(u), w\rangle \geq 0$ pour toute direction $w \in \tilde{K}(u)$ et on termine la démonstration grâce au Lemme de Farkas 2.5.18 ci-dessous. □

Lemme 2.5.18 (de Farkas) *Soient $a_1, \ldots, a_M$ des éléments fixés de V. On considère les ensembles*

$$\mathcal{K} = \Big\{ w \in V \ , \ \langle a_i, w\rangle \leq 0 \ \textit{pour} \ 1 \leq i \leq M \Big\} \ ,$$

et

$$\hat{\mathcal{K}} = \Big\{ q \in V \ , \ \exists\, \lambda_1, \ldots, \lambda_M \geq 0 \ , \ q = -\sum_{i=1}^{M} \lambda_i a_i \Big\} \ .$$

Alors pour tout $p \in V$, on a l'implication

$$\langle p, w\rangle \geq 0 \ \forall\, w \in \mathcal{K} \implies p \in \hat{\mathcal{K}} \ .$$

(La réciproque étant évidente, il s'agit en fait d'une équivalence.)

Démonstration. Commençons par montrer que $\hat{\mathcal{K}}$ est fermé. Supposons d'abord que les vecteurs $(a_i)_{1\leq i\leq M}$ sont linéairement indépendants. Soit $(q^n) = \left(-\sum_{i=1}^M \lambda_i^n a_i\right)$ une suite d'éléments de $\hat{\mathcal{K}}$ (donc avec $\lambda_i^n \geq 0 \ \forall\, i \,\forall\, n$), convergeant vers une limite $q \in V$. Alors il est clair que chaque suite (λ_i^n) converge dans $\mathbb{R}_+$ vers une limite $\lambda_i \geq 0$ (pour $1 \leq i \leq M$) puisque les vecteurs $(a_i)_{1\leq i\leq M}$ forment une base de l'espace qu'ils engendrent. On a donc $q = -\sum_{i=1}^M \lambda_i a_i \in \hat{\mathcal{K}}$, qui est donc fermé.

Si les vecteurs $(a_i)_{1\leq i\leq M}$ sont linéairement dépendants, nous procédons par récurrence sur leur nombre M. La propriété est évidente lorsque $M = 1$, et nous supposons qu'elle est vraie lorsque le nombre de vecteurs a_i est inférieur à M. Comme les vecteurs $(a_i)_{1\leq i\leq M}$ sont liés, il existe une relation de la forme $\sum_{i=1}^M \mu_i a_i = 0$, avec au moins un des coefficients μ_i qui est strictement positif. Soit alors $q = -\sum_{i=1}^M \lambda_i a_i$ un élément de $\hat{\mathcal{K}}$. Pour tout $t \leq 0$, on peut aussi écrire $q = -\displaystyle\sum_{i=1}^M (\lambda_i + t\mu_i)a_i$, et on peut choisir $t \leq 0$ pour que

$$\lambda_i + t\mu_i \geq 0 \ \forall\, i \in \{1, \ldots, M\} \quad \text{et} \quad \exists\, i_0 \in \{1, \ldots, M\} \ , \ \lambda_{i_0} + t\mu_{i_0} = 0 \ .$$

Ce raisonnement montre que

$$\hat{\mathcal{K}} = \bigcup_{i_0=1}^M \left\{ q \in V \ , \ \exists\, \lambda_1, \ldots, \lambda_M \geq 0 \ , \ q = -\sum_{i\neq i_0} \lambda_i a_i \right\} . \tag{2.49}$$

Par notre hypothèse de récurrence, chacun des ensembles apparaissant dans le membre de droite de (2.49) est fermé, et il en est donc de même de $\hat{\mathcal{K}}$.

Raisonnons maintenant par l'absurde : supposons que $\langle p, w\rangle \geq 0 \ \forall\, w \in \mathcal{K}$ et que $p \notin \hat{\mathcal{K}}$. On peut alors utiliser le Théorème 8.1.12 de séparation d'un point et d'un convexe pour séparer p et $\hat{\mathcal{K}}$ qui est fermé et, à l'évidence, convexe et non vide. Il existe donc $w \neq 0$ dans V et $\alpha \in \mathbb{R}$ tels que

$$\langle p, w\rangle < \alpha < \langle q, w\rangle \ \forall\, q \in \hat{\mathcal{K}} \ . \tag{2.50}$$

Mais alors, on doit avoir $\alpha < 0$ puisque $0 \in \hat{\mathcal{K}}$; d'autre part, pour tout $i \in \{1, \ldots, M\}$ nous pouvons choisir dans (2.50) $q = -\lambda a_i$ avec λ arbitrairement grand, ce qui montre que $\langle a_i, w\rangle \leq 0$. On obtient donc que $w \in \mathcal{K}$ et que $\langle p, w\rangle < \alpha < 0$, ce qui est impossible. □

On peut donner un autre éclairage au Théorème 2.5.16 grâce à la notion de Lagrangien.

Définition 2.5.19 *On appelle* **Lagrangien** *du problème de minimisation de $J(v)$, sous les contraintes d'inégalité $F(v) \leq 0$, la fonction $\mathcal{L}(v, \mu)$ définie par*

$$\mathcal{L}(v, \mu) = J(v) + \sum_{i=1}^M \mu_i F_i(v) = J(v) + \mu \cdot F(v) \qquad \forall (v, \mu) \in V \times (\mathbb{R}^+)^M.$$

La nouvelle variable **positive** $\mu \in (\mathbb{R}^+)^M$ *est appelée* **multiplicateur de Lagrange** *pour la contrainte $F(v) \leq 0$.*

Comme dans le Lemme 2.5.10 la maximisation du Lagrangien permet de faire "disparaître" la contrainte.

Lemme 2.5.20 *Le problème de minimisation sous contrainte d'inégalité est équivalent à un problème de min-max*

$$\inf_{v\in V,\ F(v)\leq 0} J(v) = \inf_{v\in V} \sup_{\mu\in(\mathbb{R}^+)^M} \mathcal{L}(v,\mu). \tag{2.51}$$

La démonstration du Lemme 2.5.20 est très similaire à celle du Lemme 2.5.10 et laissée au lecteur en guise d'exercice.

Remarque 2.5.21 Comme pour le cas des contraintes d'égalité, la notion de Lagrangien est extrêmement utile pour les algorithmes numériques de minimisation. Si la contrainte $F(u^n) \leq 0$ est violée pour une solution approchée u^n, alors, pour un multiplicateur de Lagrange $\mu^n \geq 0$, dont les composantes correspondant à une contrainte violée sont strictement positives tandis que les composantes correspondant à une contrainte inactive sont nulles, la minimisation sans contrainte de $\mathcal{L}(v,\mu^n)$ permet d'obtenir une nouvelle itérée u^{n+1} qui minimise au mieux la fonction objectif et la violation de la contrainte. Cela sera précisé dans la Sous-section 3.3.2 lorsque sera présenté l'algorithme d'Uzawa. •

La condition **nécessaire** d'optimalité (2.45) du Théorème 2.5.16 est bien la stationnarité du Lagrangien de la Définition 2.5.19 puisque

$$\frac{\partial \mathcal{L}}{\partial v}(u,\lambda) = J'(u) + \lambda \cdot F'(u) = 0,$$

et que la condition $\lambda \geq 0$, $F(u) \leq 0$, $\lambda \cdot F(u) = 0$ est équivalente à l'inéquation d'Euler (2.29) pour la maximisation par rapport à μ dans le convexe fermé $(\mathbb{R}^+)^M$

$$\frac{\partial \mathcal{L}}{\partial \mu}(u,\lambda)\cdot(\mu-\lambda) = F(u)\cdot(\mu-\lambda) \leq 0 \quad \forall \mu \in (\mathbb{R}^+)^M.$$

Exercice 2.5.17 Soit A une matrice symétrique définie positive d'ordre n, et B une matrice de taille $m \times n$ avec $m \leq n$ et de rang m. On considère le problème de minimisation

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

Appliquer le Théorème 2.5.16 pour obtenir l'existence d'un multiplicateur de Lagrange $p \in \mathbb{R}^m$ tel qu'un point de minimum $\overline{x}$ vérifie

$$A\overline{x} - b + B^* p = 0\ , \quad p \geq 0\ , \quad p\cdot(B\overline{x} - c) = 0.$$

Exercice 2.5.18 Soit f une fonction définie sur un ouvert borné Ω. Pour $\epsilon > 0$ on considère le problème de régularisation suivant

$$u \in V_0\ ,\ \inf_{\int_\Omega |u-f|^2 dx \leq \epsilon^2} \int_\Omega |\nabla u|^2 dx,$$

où $V_0 = \{v \in C^1(\overline{\Omega}),\ v = 0 \text{ sur } \partial\Omega\}$. Montrer qu'un point de minimum u_ϵ vérifie, soit $u_\epsilon = f$, soit il existe $\lambda \geq 0$ tel que u_ϵ est solution de

$$\begin{cases} -\Delta u_\epsilon + \lambda(u_\epsilon - f) = 0 & \text{dans } \Omega, \\ u_\epsilon = 0 & \text{sur } \partial\Omega. \end{cases}$$

Contraintes d'égalité et d'inégalité

On peut bien sûr mélanger les deux types de contraintes. On suppose donc que K est donné par

$$K = \{v \in V \quad , \quad G(v) = 0 \ , \quad F(v) \leq 0\} \ , \tag{2.52}$$

où $G(v) = (G_1(v), ..., G_N(v))$ et $F(v) = (F_1(v), ..., F_M(v))$ sont deux applications de V dans $\mathbb{R}^N$ et $\mathbb{R}^M$. Dans ce nouveau contexte, il faut donner une définition adéquate de la qualification des contraintes. On note toujours $I(u) = \{i \in \{1, \dots, M\} \ , \ F_i(u) = 0\}$ l'ensemble des contraintes d'inégalité actives en $u \in K$.

Définition 2.5.22 *On dit que les contraintes (2.52) sont* **qualifiées** *en $u \in K$ si et seulement si les vecteurs $\left(G_i'(u)\right)_{1 \leq i \leq N}$ sont linéairement indépendants et il existe une direction $\overline{w} \in \bigcap_{i=1}^N \left[G_i'(u)\right]^\perp$ telle que l'on ait pour tout $i \in I(u)$*

$$\langle F_i'(u), \overline{w} \rangle < 0 \ . \tag{2.53}$$

Nous pouvons alors énoncer les conditions **nécessaires** d'optimalité sur l'ensemble (2.52).

Théorème 2.5.23 *Soit $u \in K$ où K est donné par (2.52). On suppose que J et F sont dérivables en u, que G est dérivable dans un voisinage de u, et que les contraintes sont qualifiées en u (au sens de la Définition 2.5.22). Alors, si u est un minimum local de J sur K, il existe des multiplicateurs de Lagrange $\mu_1, \dots, \mu_N$, et $\lambda_1, \dots, \lambda_M \geq 0$, tels que*

$$J'(u) + \sum_{i=1}^N \mu_i G_i'(u) + \sum_{i=1}^M \lambda_i F_i'(u) = 0 \ , \quad \lambda \geq 0 \ , \ F(u) \leq 0 \ , \ \lambda \cdot F(u) = 0 \ . \tag{2.54}$$

La démonstration du Théorème 2.5.23 est une simple adaptation de celles des Théorèmes 2.5.6 et 2.5.16, que nous laissons au lecteur en guise d'exercice.

Autres formes des conditions de qualification

Les conditions de qualification sont des conditions **suffisantes** de type "géométrique" qui permettent de faire des variations internes à l'ensemble K à partir d'un point $u \in K$. La condition de qualification de la Définition 2.5.14 est assez générale (quoique loin d'être nécessaire), mais parfois difficile à vérifier dans les applications. C'est pourquoi les remarques qui suivent donnent des conditions de qualifications plus simples (donc plus faciles à vérifier en pratique) mais moins générales (i.e. moins souvent vérifiées).

Remarque 2.5.24 Dans le cas des contraintes d'inégalité, on peut s'inspirer de la notion de cas régulier (introduite à la Remarque 2.5.8 pour les contraintes d'égalité) afin de donner une condition très simple qui entraîne la condition de qualification de la Définition 2.5.14. En effet, pour $u \in K$ les contraintes inactives ne "jouent" pas et seules sont à prendre en compte les contraintes actives $i \in I(u)$ qui sont justement des contraintes d'égalité en ce point ! On vérifie alors sans peine que la condition suivante (qui dit que u est un point régulier pour les contraintes d'égalité $F_i(u) = 0$ pour $i \in I(u)$)

$$(F_i'(u))_{i\in I(u)} \text{ est une famille libre} \tag{2.55}$$

entraîne (2.44), c'est-à-dire que les contraintes sont qualifiées. En effet, il suffit de prendre $\overline{w} = \sum_{i\in I(u)} \alpha_i F_i'(u)$ tel que $\langle F_j'(u), \overline{w}\rangle = -1$ pour tout $j \in I(u)$ (l'existence des coefficients α_i découle de l'inversibilité de la matrice $(\langle F_i'(u), F_j'(u)\rangle)_{ij}$. Il est clair cependant que (2.44) n'implique pas (2.55). •

Remarque 2.5.25 Dans le cas des contraintes combinées d'égalité et d'inégalité, on peut aussi s'inspirer de la notion de cas régulier pour donner une condition plus simple qui implique la condition de qualification de la Définition 2.5.22. Cette condition "forte" (c'est-à-dire moins souvent vérifiée) de qualification est

$$\big(G_i'(u)\big)_{1\le i\le N} \bigcup (F_i'(u))_{i\in I(u)} \text{ est une famille libre.} \tag{2.56}$$

On vérifie facilement que (2.56) entraîne (2.53), c'est-à-dire que les contraintes sont qualifiées. •

Remarque 2.5.26 Revenant au cas des contraintes d'inégalité, supposées convexes, une autre condition de qualification possible est la suivante. On suppose qu'il existe $\overline{v} \in V$ tel que l'on ait, pour tout $i \in \{1, \dots, M\}$,

$$\begin{array}{l} \text{les fonctions } F_i \text{ sont convexes et,} \\ \quad \text{ou bien} \quad F_i(\overline{v}) < 0 \ , \\ \quad \text{ou bien} \quad F_i(\overline{v}) = 0 \quad \text{et} \quad F_i \quad \text{est affine.} \end{array} \tag{2.57}$$

L'hypothèse (2.57) entraîne que les contraintes sont qualifiées en $u \in K$ au sens de la Définition 2.5.14. En effet, si $i \in I(u)$ et si $F_i(\overline{v}) < 0$, alors, d'après la condition de convexité (2.20)

$$\langle F_i'(u), \overline{v} - u\rangle = F_i(u) + \langle F_i'(u), \overline{v} - u\rangle \le F_i(\overline{v}) < 0 \ .$$

D'autre part, si $i \in I(u)$ et si $F_i(\overline{v}) = 0$, alors F_i est affine et

$$\langle F_i'(u), \overline{v} - u\rangle = F_i(\overline{v}) - F_i(u) = 0 \ ,$$

et la Définition 2.5.14 de qualification des contraintes est satisfaite avec $\overline{w} = \overline{v} - u$. L'avantage de l'hypothèse (2.57) est de ne pas nécessiter de connaître le point de minimum u ni de calculer les dérivées des fonctions $F_1, \dots, F_M$. •

2.6 Point-selle, théorème de Kuhn et Tucker, dualité

Nous avons vu après la Définition 2.5.9 du Lagrangien $\mathcal{L}$ comment il est possible d'interpréter le couple (u, λ) (point de minimum, multiplicateur de Lagrange) comme un **point stationnaire** de ce Lagrangien. Nous allons dans cette section préciser la nature de ce point stationnaire comme **point-selle** et montrer comment cette formulation permet de caractériser un minimum (ce qui veut dire que, sous certaines hypothèses, nous verrons que les conditions **nécessaires** de stationnarité du Lagrangien sont aussi **suffisantes**). Nous explorerons brièvement la **théorie de la dualité** qui en découle.

Outre l'intérêt théorique de cette caractérisation, son intérêt pratique du point de vue des algorithmes numériques sera illustré au Chapitre 3. Signalons enfin que la notion de point-selle joue un rôle fondamental dans la **théorie des jeux**.

2.6.1 Point-selle

De manière abstraite, V et Q étant deux espaces de Hilbert réels, un Lagrangien $\mathcal{L}$ est une application de $V \times Q$ (ou d'une partie $U \times P$ de $V \times Q$) dans $\mathbb{R}$. Dans le cadre du Théorème 2.5.6 sur les contraintes d'égalité (ou plutôt de la Définition 2.5.9), nous avions $U = V$, $P = Q = \mathbb{R}^M$ et $\mathcal{L}(v, q) = J(v) + q \cdot F(v)$. La situation est un peu différente dans le cadre du Théorème 2.5.16 sur les contraintes d'inégalité, où pour le même Lagrangien $\mathcal{L}(v, q) = J(v) + q \cdot F(v)$ il faut prendre $U = V$, $Q = \mathbb{R}^M$ et $P = (\mathbb{R}_+)^M$.

Donnons maintenant la définition d'un point-selle, souvent appelé également min-max ou col (voir la Figure 2.4).

Définition 2.6.1 *On dit que $(u, p) \in U \times P$ est un point-selle de $\mathcal{L}$ sur $U \times P$ si*

$$\forall\, q \in P \qquad \mathcal{L}(u, q) \leq \mathcal{L}(u, p) \leq \mathcal{L}(v, p) \quad \forall\, v \in U\ . \tag{2.58}$$

Le résultat suivant montre le lien entre cette notion de point-selle et les problèmes de minimisation avec contraintes d'égalité (2.36) ou contraintes d'inégalité (2.43) étudiés dans la section précédente. Pour simplifier, nous utiliserons de nouveau des inégalités entre vecteurs, notant parfois $q \geq 0$ au lieu de $q \in (\mathbb{R}_+)^M$.

Proposition 2.6.2 *On suppose que les fonctions $J, F_1, \ldots, F_M$ sont continues sur V, et que l'ensemble K est défini par (2.36) ou (2.43). On note $P = \mathbb{R}^M$ dans le cas de contraintes d'égalité (2.36) et $P = (\mathbb{R}_+)^M$ dans le cas de contraintes d'inégalité (2.43). Soit U un ouvert de V contenant K. Pour $(v, q) \in U \times P$, on pose $\mathcal{L}(v, q) = J(v) + q \cdot F(v)$.*

Supposons que (u, p) soit un point-selle de $\mathcal{L}$ sur $U \times P$. Alors $u \in K$ et u est un minimum global de J sur K. De plus, si J et $F_1, \ldots, F_M$ sont dérivables en u, on a

$$J'(u) + \sum_{i=1}^{M} p_i F_i'(u) = 0\ . \tag{2.59}$$

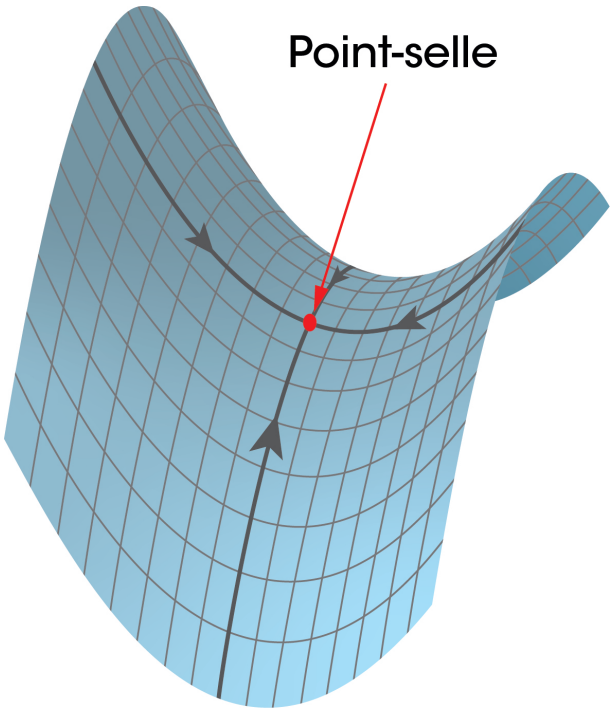

FIGURE 2.4 – Point sell ou col pour un Lagrangien.

Démonstration. Écrivons la condition de point-selle

$$\forall\, q \in P \quad J(u) + q \cdot F(u) \leq J(u) + p \cdot F(u) \leq J(v) + p \cdot F(v) \quad \forall\, v \in U\ . \tag{2.60}$$

Examinons d'abord le cas de contraintes d'égalité. Puisque $P = \mathbb{R}^M$, la première inégalité dans (2.60) montre que $F(u) = 0$, i.e. $u \in K$. Il reste alors $J(u) \leq J(v) + p \cdot F(v) \quad \forall\, v \in U$, qui montre bien (en prenant $v \in K$) que u est un minimum global de J sur K.

Dans le cas de contraintes d'inégalité, on a $P = (\mathbb{R}_+)^M$ et la première inégalité de (2.60) montre maintenant que $F(u) \leq 0$ et que $p \cdot F(u) = 0$. Ceci prouve encore que $u \in K$, et permet de déduire facilement de la deuxième inégalité que u est un minimum global de J sur K.

Enfin, si J et $F_1, \ldots, F_M$ sont dérivables en u, la deuxième inégalité de (2.60) montre que u est un point de minimum sans contrainte de $J + p \cdot F$ dans l'ouvert U, ce qui implique que la dérivée s'annule en u, $J'(u) + p \cdot F'(u) = 0$ (cf. la Remarque 2.5.2). □

2.6.2 Théorème de Kuhn et Tucker

Nous revenons au problème de minimisation sous contraintes d'inégalité pour lequel l'ensemble K est donné par (2.43), c'est-à-dire

$$K = \{v \in V\ , \quad F_i(v) \leq 0 \quad \text{pour} \quad 1 \leq i \leq m\}\ . \tag{2.61}$$

Le Théorème 2.5.16 a donné une condition nécessaire d'optimalité. Dans cette sous-section nous allons voir que cette condition est aussi **suffisante** si les contraintes et

la fonction coût sont **convexes**. En effet, la Proposition 2.6.2 affirme que, si (u, p) est un point-selle du Lagrangien, alors u réalise le minimum de J sur K. Pour un problème de minimisation convexe avec des contraintes d'inégalités convexes, nous allons établir une réciproque de ce résultat, c'est-à-dire que, si u réalise le minimum de J sur K, alors il existe $p \in (\mathbb{R}_+)^M$ tel que (u, p) soit point-selle du Lagrangien. On suppose désormais que $J, F_1, \ldots, F_M$ sont convexes continues sur V.

Remarque 2.6.3 Comme $J, F_1, \ldots, F_M$ sont convexes continues, K est convexe fermé et l'existence d'un minimum global de J sur K est assuré par le Théorème 2.3.9 dès que K est non vide et que la condition "infinie à l'infini" (2.9) est vérifiée. •

Le théorème de Kuhn et Tucker (appelé aussi parfois théorème de Karush, Kuhn et Tucker) affirme que, dans le cas convexe, la condition nécessaire d'optimalité du Théorème 2.5.16 est en fait une condition **nécessaire et suffisante**.

Théorème 2.6.4 (de Kuhn et Tucker) *On suppose que les fonctions J, $F_1, \ldots, F_M$ sont convexes continues sur V et dérivables sur l'ensemble K (2.61). On introduit le Lagrangien $\mathcal{L}$ associé*

$$\mathcal{L}(v, q) = J(v) + q \cdot F(v) \quad \forall\, (v, q) \in V \times (\mathbb{R}_+)^M \ .$$

Soit $u \in K$ un point de K où les contraintes sont qualifiées au sens de la Définition 2.5.14. Alors u est un minimum global de J sur K si et seulement si il existe $p \in (\mathbb{R}_+)^M$ tel que (u, p) soit un point-selle du Lagrangien $\mathcal{L}$ sur $V \times (\mathbb{R}_+)^M$ ou, de manière équivalente, tel que

$$F(u) \leq 0 \ , \ p \geq 0 \ , \ p \cdot F(u) = 0 \ , \ J'(u) + \sum_{i=1}^{M} p_i F_i'(u) = 0 \ . \tag{2.62}$$

Démonstration. Si u est un minimum de J sur K, on peut appliquer le Théorème 2.5.16, qui donne exactement la condition d'optimalité (2.62), d'où l'on déduit facilement que (u, p) est point-selle de $\mathcal{L}$ sur $V \times (\mathbb{R}_+)^M$ (en utilisant le fait que $J(v) + p \cdot F(v)$ est convexe). Réciproquement, si (u, p) est point-selle, on a déjà montré à la Proposition 2.6.2 que u est un minimum global de J sur K. □

Remarque 2.6.5 Le Théorème 2.6.4 de Kuhn et Tucker ne s'applique qu'aux contraintes d'inégalité, et pas aux contraintes d'égalité, en général. Cependant, il est bon de remarquer que des contraintes **d'égalité affines** $Av = b$ peuvent s'écrire sous la forme de contraintes d'inégalité (affines donc convexes) $Av - b \leq 0$ et $b - Av \leq 0$. C'est une évidence qui permet cependant d'appliquer le Théorème 2.6.4 de Kuhn et Tucker à un problème de minimisation avec contraintes d'égalité affines. •

L'exercice suivant permet d'interpréter les multiplicateurs de Lagrange p_i comme la sensibilité de la valeur minimale de J aux variations des contraintes F_i : en économie, ces coefficients mesurent des prix ou des coûts marginaux , en mécanique des forces de liaison correspondant à des contraintes cinématiques, etc...

Exercice 2.6.1 On considère le problème d'optimisation, dit perturbé

$$\inf_{F_i(v)\leq u_i,\ 1\leq i\leq m} J(v), \tag{2.63}$$

avec $u = (u_1, \ldots, u_m) \in \mathbb{R}^m$. On se place sous les hypothèses du Théorème 2.6.4 de Kuhn et Tucker. On note $m^*(u)$ la valeur minimale du problème perturbé (2.63).

1. Montrer que si $p \in \mathbb{R}^m$ est le multiplicateur de Lagrange pour le problème non perturbé (c'est-à-dire (2.63) avec $u = 0$), alors

$$m^*(u) \geq m^*(0) - p \cdot u \ . \tag{2.64}$$

2. Déduire de (2.64) que si $u \mapsto m^*(u)$ est dérivable, alors

$$p_i = -\frac{\partial m^*}{\partial u_i}(0).$$

Interpréter ce résultat (cf. l'Exemple 1.2.5 en économie).

2.6.3 Dualité

Donnons un bref aperçu de la théorie de la dualité pour les problèmes d'optimisation. Nous l'appliquerons au problème de minimisation convexe avec contraintes d'inégalité de la sous-section précédente. Nous avons associé à ce problème de minimisation un problème de recherche d'un point-selle (u, p) pour le Lagrangien $\mathcal{L}(v, q) = J(v) + q \cdot F(v)$. Mais nous allons voir que, à l'existence d'un point-selle (u, p) du Lagrangien, on peut associer inversement non pas un mais **deux** problèmes d'optimisation (plus précisément, un problème de minimisation et un problème de maximisation), qui seront dits **duaux** l'un de l'autre. Nous expliquerons ensuite sur deux exemples simples en quoi l'introduction du **problème dual** peut être utile pour la résolution du problème d'origine, dit **problème primal** (par opposition au dual).

Revenons un instant au cadre général de la Définition 2.6.1.

Définition 2.6.6 *Soit V et Q deux espaces de Hilbert réels, et $\mathcal{L}$ un Lagrangien défini sur une partie $U \times P$ de $V \times Q$. On suppose qu'il existe un point-selle (u, p) de $\mathcal{L}$ sur $U \times P$*

$$\forall\, q \in P \quad \mathcal{L}(u, q) \leq \mathcal{L}(u, p) \leq \mathcal{L}(v, p) \quad \forall\, v \in U \ . \tag{2.65}$$

Pour $v \in U$ et $q \in P$, posons

$$\mathcal{J}(v) = \sup_{q\in P} \mathcal{L}(v, q) \qquad \mathcal{G}(q) = \inf_{v\in U} \mathcal{L}(v, q) \ . \tag{2.66}$$

On appelle problème primal le problème de minimisation

$$\inf_{v\in U} \mathcal{J}(v) \ , \tag{2.67}$$

et problème dual le problème de maximisation

$$\sup_{q\in P} \mathcal{G}(q) \ . \tag{2.68}$$

Remarque 2.6.7 Bien sûr, sans hypothèses supplémentaires, il peut arriver que $\mathcal{J}(v) = +\infty$ pour certaines valeurs de v ou que $\mathcal{G}(q) = -\infty$ pour certaines valeurs de q. Mais l'existence supposée du point-selle (u, p) dans la Définition 2.6.6 nous assure que les **domaines** de $\mathcal{J}$ et $\mathcal{G}$ (i.e. les ensembles $\{v \in U\ ,\ \mathcal{J}(v) < +\infty\}$ et $\{q \in P\ ,\ \mathcal{G}(q) > -\infty\}$ sur lesquels ces fonctions sont bien définies) ne sont pas vides, puisque (2.65) montre que $\mathcal{J}(u) = \mathcal{G}(p) = \mathcal{L}(u, p)$. Les problèmes primal et dual ont donc bien un sens. Le résultat suivant montre que ces deux problèmes sont étroitement liés au point-selle (u, p). •

Théorème 2.6.8 (de dualité) *Le couple (u, p) est un point-selle de $\mathcal{L}$ sur $U \times P$ si et seulement si*

$$\mathcal{J}(u) = \min_{v \in U} \mathcal{J}(v) = \max_{q \in P} \mathcal{G}(q) = \mathcal{G}(p)\ . \tag{2.69}$$

Remarque 2.6.9 Par la Définition (2.66) de $\mathcal{J}$ et $\mathcal{G}$, (2.69) est équivalent à

$$\mathcal{J}(u) = \min_{v \in U} \left(\sup_{q \in P} \mathcal{L}(v, q) \right) = \max_{q \in P} \left(\inf_{v \in U} \mathcal{L}(v, q) \right) = \mathcal{G}(p)\ . \tag{2.70}$$

Si le sup et l'inf sont atteints dans (2.70) (c'est-à-dire qu'on peut les écrire max et min, respectivement), on voit alors que (2.70) traduit la possibilité d'échanger l'ordre du min et du max appliqués au Lagrangien $\mathcal{L}$. Ce fait (qui est faux si $\mathcal{L}$ n'admet pas de point selle) explique le nom de min-max qui est souvent donné à un point-selle. •

Démonstration. Soit (u, p) un point-selle de $\mathcal{L}$ sur $U \times P$. Notons $\mathcal{L}^* = \mathcal{L}(u, p)$. Pour $v \in U$, il est clair d'après (2.66) que $\mathcal{J}(v) \geq \mathcal{L}(v, p)$, d'où $\mathcal{J}(v) \geq \mathcal{L}^*$ d'après (2.65). Comme $\mathcal{J}(u) = \mathcal{L}^*$, ceci montre que $\mathcal{J}(u) = \inf_{v \in U} \mathcal{J}(v) = \mathcal{L}^*$. On montre de la même façon que $\mathcal{G}(p) = \sup_{q \in P} \mathcal{G}(q) = \mathcal{L}^*$.

Réciproquement, supposons que (2.69) a lieu et posons $\mathcal{L}^* = \mathcal{J}(u)$. La définition (2.66) de $\mathcal{J}$ montre que

$$\mathcal{L}(u, q) \leq \mathcal{J}(u) = \mathcal{L}^* \quad \forall\, q \in P\ . \tag{2.71}$$

De même, on a aussi :

$$\mathcal{L}(v, p) \geq \mathcal{G}(p) = \mathcal{L}^* \quad \forall\, v \in U\ , \tag{2.72}$$

et on déduit facilement de (2.71)-(2.72) que $\mathcal{L}(u, p) = \mathcal{L}^*$, ce qui montre que (u, p) est point-selle. □

Remarque 2.6.10 Même si le Lagrangien $\mathcal{L}$ n'admet pas de point selle sur $U \times P$, on a tout de même l'inégalité élémentaire suivante, dite de **dualité faible**

$$\inf_{v \in U} \left(\sup_{q \in P} \mathcal{L}(v, q) \right) \geq \sup_{q \in P} \left(\inf_{v \in U} \mathcal{L}(v, q) \right). \tag{2.73}$$

En effet, pour tout $v \in U$ et $q \in P$, $\mathcal{L}(v, q) \geq \inf_{v' \in U} \mathcal{L}(v', q)$, donc $\sup_{q \in P} \mathcal{L}(v, q) \geq \sup_{q \in P} \inf_{v' \in U} \mathcal{L}(v', q)$, et puisque ceci est vrai pour tout $v \in V$, $\inf_{v \in V} \sup_{q \in P} \mathcal{L}(v, q) \geq \sup_{q \in P} \inf_{v' \in U} \mathcal{L}(v', q)$, ce qui donne (2.73). La différence (positive) entre les deux membres de l'inégalité (2.73) est appelée **saut de dualité**. •

Exercice 2.6.2 Donner un exemple de Lagrangien pour lequel l'inégalité (2.73) est stricte avec ses deux membres finis.

Exercice 2.6.3 Soit U (respectivement P) un convexe compact non vide de V (respectivement Q). On suppose que le Lagrangien est tel que $v \to \mathcal{L}(v,q)$ est convexe sur U pour tout $q \in P$, et $q \to \mathcal{L}(v,q)$ est concave sur P pour tout $v \in U$. Montrer alors l'existence d'un point selle de $\mathcal{L}$ sur $U \times P$.

Application

Nous appliquons ce résultat de dualité au problème précédent de minimisation convexe avec contraintes d'inégalité convexes

$$\inf_{v \in V\ ,\ F(v) \leq 0} J(v) \tag{2.74}$$

avec J et $F = (F_1, \ldots, F_M)$ convexes sur V. On introduit le Lagrangien

$$\mathcal{L}(v,q) = J(v) + q \cdot F(v) \quad \forall\, (v,q) \in V \times (\mathbb{R}_+)^M \ .$$

Dans ce cadre, on voit facilement que, pour tout $v \in V$,

$$\mathcal{J}(v) = \sup_{q \in (\mathbb{R}_+)^M} \mathcal{L}(v,q) = \begin{cases} J(v) & \text{si } F(v) \leq 0 \\ +\infty & \text{sinon ,} \end{cases} \tag{2.75}$$

ce qui montre que le problème primal $\inf_{v \in V} \mathcal{J}(v)$ est exactement le problème d'origine (2.74) ! D'autre part, la fonction $\mathcal{G}(q)$ du problème dual est bien définie par (2.66), car (2.66) est ici un problème de minimisation convexe. De plus, $\mathcal{G}(q)$ est une fonction concave car elle est l'infimum de fonctions affines (voir l'Exercice 2.3.2). Par conséquent, le problème dual

$$\sup_{q \in (\mathbb{R}_+)^M} \mathcal{G}(q) \ ,$$

est un problème de maximisation concave **plus simple** que le problème primal (2.74) car les contraintes sont linéaires ! Cette particularité est notamment exploitée dans des algorithmes numériques (cf. l'algorithme d'Uzawa). Une simple combinaison des Théorèmes de Kuhn et Tucker 2.6.4 et de dualité 2.6.8 nous donne le résultat suivant.

Corollaire 2.6.11 *On suppose que les fonctions $J, F_1, \ldots, F_M$ sont convexes et dérivables sur V. Soit $u \in V$ tel que $F(u) \leq 0$ et les contraintes sont qualifiées en u au sens de la Définition 2.5.14. Alors, si u est un minimum global de $\mathcal{J}$ sur V, il existe $p \in (\mathbb{R}_+)^M$ tel que*

1. *p est un maximum global de $\mathcal{G}$ sur $(\mathbb{R}_+)^M$,*
2. *(u,p) est un point-selle du Lagrangien $\mathcal{L}$ sur $V \times (\mathbb{R}_+)^M$,*
3. *$(u,p) \in V \times (\mathbb{R}_+)^M$ vérifie la condition d'optimalité nécessaire et suffisante*

$$F(u) \leq 0 \ ,\ p \geq 0 \ ,\ p \cdot F(u) = 0 \ ,\ J'(u) + p \cdot F'(u) = 0 \ . \tag{2.76}$$

L'application la plus courante du Corollaire 2.6.11 est la suivante. Supposons que le problème dual de maximisation est plus facile à résoudre que le problème primal (c'est le cas en général car ses contraintes sont plus simples). Alors pour calculer la solution u du problème primal on procède en deux étapes. Premièrement, on calcule la solution p du problème dual. Deuxièmement, on dit que (u, p) est un point selle du Lagrangien, c'est-à-dire que l'on calcule u, solution du problème de minimisation **sans contrainte**

$$\min_{v \in V} \mathcal{L}(v, p) \ .$$

Précisons qu'avec les hypothèses faites il n'y a pas a priori d'unicité des solutions pour tous ces problèmes. Précisons aussi que pour obtenir l'existence du minimum u dans le Corollaire 2.6.11 il suffit d'ajouter une hypothèse de forte convexité ou de comportement infini à l'infini sur J.

Remarque 2.6.12 Pour illustrer le Corollaire 2.6.11 et l'intérêt de la dualité, nous considérons un problème de minimisation quadratique dans $\mathbb{R}^N$ avec contraintes d'inégalité affines

$$\min_{v \in \mathbb{R}^N,\ F(v) = Bv - c \leq 0} \left\{ J(v) = \frac{1}{2} Av \cdot v - b \cdot v \right\} \ , \tag{2.77}$$

où A est une matrice $N \times N$ symétrique définie positive, $b \in \mathbb{R}^N$, B une matrice $M \times N$ et $c \in \mathbb{R}^M$. Le Lagrangien est donné par

$$\mathcal{L}(v, q) = \frac{1}{2} Av \cdot v - b \cdot v + q \cdot (Bv - c) \quad \forall\, (v, q) \in \mathbb{R}^N \times (\mathbb{R}_+)^M \ . \tag{2.78}$$

Nous avons déjà fait dans (2.75) le calcul de $\mathcal{J}$, et dit que le problème primal est exactement (2.77). Examinons maintenant le problème dual. Pour $q \in (\mathbb{R}_+)^M$, le problème

$$\min_{v \in \mathbb{R}^N} \mathcal{L}(v, q)$$

a une solution unique puisque $v \to \mathcal{L}(v, q)$ est une fonction fortement convexe. Cette solution vérifie $\frac{\partial \mathcal{L}}{\partial v}(v, q) = Av - b + B^* q = 0$, soit $v = A^{-1}(b - B^* q)$. On obtient donc

$$\mathcal{G}(q) = \mathcal{L}\big(A^{-1}(b - B^* q), q\big) \ ,$$

et le problème dual s'écrit finalement

$$\sup_{q \geq 0} \left(-\frac{1}{2} q \cdot BA^{-1}B^* q + (BA^{-1}b - c) \cdot q - \frac{1}{2} A^{-1} b \cdot b \right) \ . \tag{2.79}$$

Certes, la fonctionnelle à maximiser dans (2.79) n'a pas une allure particulièrement sympathique. Il s'agit encore d'un problème avec fonctionnelle quadratique et contraintes affines. Cependant, le Corollaire 2.6.11 nous assure qu'il a une solution. On peut voir d'ailleurs que cette solution n'est pas forcément unique (sauf si la matrice B est de rang M car la matrice $BA^{-1}B^*$ est alors définie positive).

Mais l'avantage important du problème dual (2.79) vient du fait que les contraintes ($q \geq 0$) s'expriment sous une forme particulièrement simple, bien plus simple que pour le problème primal ; et nous verrons à la Section 3.3 que cet avantage peut être utilisé pour mettre au point un algorithme de calcul de la solution du problème primal. •

Exercice 2.6.4 On reprend l'Exemple 1.2.8, déjà étudié à l'Exercice 2.5.4. En mécanique il est bien connu que la minimisation de l'énergie "en déplacements" $J(u)$ de l'Exercice 2.5.4 est équivalent à la minimisation d'une autre énergie, dite **complémentaire** dont la signification physique est tout aussi importante que celle de $J(u)$. Cette énergie complémentaire est définie en terme d'un champ de contraintes (mécaniques) $\tau(x)$ par

$$G(\tau) = \frac{1}{2}\int_0^L |\tau|^2 dx. \tag{2.80}$$

Elle s'accompagne d'une contrainte sur τ qui doit être **statiquement admissible**, c'est-à-dire vérifier $-\tau' = f$ dans $(0, L)$. Autrement dit, on considère le problème de minimisation sous contrainte

$$\inf_{-\tau'=f \text{ dans } (0,L)} \left\{ G(\tau) = \frac{1}{2}\int_0^L |\tau|^2 dx \right\}. \tag{2.81}$$

Pour τ et v, deux fonctions définies de $(0, L)$ dans $\mathbb{R}$, on introduit le Lagrangien correspondant

$$\mathcal{L}(\tau, v) = \frac{1}{2}\int_0^L |\tau|^2 dx + \int_0^L v(\tau' + f) dx.$$

Montrer que la fonction duale $\mathcal{D}(v)$ correspondante n'est rien d'autre que l'opposée de l'énergie $-J(u)$ de l'Exercice 2.5.4. En admettant que $-J(u)$ admette un point de maximum u et que (2.81) admette un point de minimum σ, montrer que (σ, u) est un point selle du Lagrangien et que $\sigma = u'$.

Terminons par un exercice récréatif qui montre la relation entre les problèmes de point-selle ou min-max et la théorie des jeux.

Exercice 2.6.5 Soit une matrice rectangulaire

$$A = \begin{pmatrix} 1 & 0 & 4 & 2 & 3 & 5 \\ -3 & 2 & -1 & 2 & -5 & 2 \\ -4 & 2 & -2 & 0 & -1 & 2 \\ -2 & 4 & -1 & 6 & -2 & 2 \\ -1 & 2 & -6 & 3 & -1 & 1 \end{pmatrix}.$$

On suppose que deux joueurs s'affrontent sur le jeu suivant : le premier choisit une ligne i, le deuxième une colonne j, sans qu'ils ne connaissent le choix de l'autre. Une fois révélé leurs choix, le gain (ou la perte, selon le signe) du premier joueur est déterminé par le coefficient a_{ij} de la matrice A (le gain de l'autre joueur est $-a_{ij}$). Montrer que la stratégie de minimisation du risque pour chacun des joueurs conduit à un problème de min-max que l'on résoudra. Le jeu est-il équitable avec cette matrice A ?

Chapitre 3

ALGORITHMES D'OPTIMISATION

3.1 Introduction

L'objet de ce chapitre est de présenter et analyser quelques algorithmes permettant de calculer, ou plus exactement d'**approcher** les solutions de problèmes d'optimisation. Un point commun à tous ces algorithmes est qu'ils s'inspirent des conditions d'optimalité étudiées au chapitre précédent et qu'en particulier ils utilisent la connaissance des dérivées des fonctions objectifs et des contraintes. Tous les algorithmes présentés ici sont effectivement utilisés en pratique pour résoudre sur ordinateur des problèmes concrets d'optimisation.

Ces algorithmes sont aussi tous de nature itérative : à partir d'une donnée initiale u^0, chaque méthode construit une suite $(u^n)_{n\in\mathbb{N}}$ dont nous montrerons qu'elle converge, sous certaines hypothèses, vers la solution u du problème d'optimisation considéré. Après avoir montré la **convergence de ces algorithmes** (c'est-à-dire, la convergence de la suite (u^n) vers u quel que soit le choix de la donnée initiale u^0), nous dirons aussi un mot de leur vitesse de convergence.

Dans toute cette section nous supposerons que la fonction objectif à minimiser J est α-convexe différentiable. Cette hypothèse d'α-convexité est assez forte, mais nous verrons plus loin qu'elle est cruciale pour les démonstrations de convergence des algorithmes. L'application des algorithmes présentés ici à la minimisation de fonctions convexes qui ne sont pas fortement convexes peut soulever quelques petites difficultés, sans parler des **grosses** difficultés qui apparaissent lorsque l'on cherche à approcher le minimum d'une fonction non convexe ! Typiquement, ces algorithmes peuvent, au mieux converger vers un minimum local (voire vers un simple point critique), très loin d'un minimum global, au pire ne pas converger, diverger ou osciller entre plusieurs limites.

Remarque 3.1.1 Nous nous limitons aux seuls algorithmes déterministes et nous ne disons rien des algorithmes de type stochastique (recuit simulé, algorithmes génétiques, etc.). Outre le fait que leur analyse fait appel à la théorie des probabilités (que nous n'abordons pas dans ce cours), leur utilisation est très différente. Pour sché-

matiser simplement, disons que les algorithmes déterministes sont les plus efficaces pour la minimisation de fonctions convexes, tandis que les algorithmes stochastiques permettent d'approcher des minima **globaux** (et pas seulement locaux) de fonctions non convexes (à un prix toutefois assez élevé en pratique). •

3.2 Algorithmes de type gradient (cas sans contraintes)

Commençons par étudier la résolution pratique de problèmes d'optimisation en l'absence de contraintes. Soit J une fonction α-convexe différentiable définie sur l'espace de Hilbert réel V, on considère le problème sans contrainte

$$\inf_{v\in V} J(v) \ . \tag{3.1}$$

D'après le Théorème 2.3.8 il existe une unique solution u, caractérisée d'après la Remarque 2.5.2 par l'équation d'Euler

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

3.2.1 Algorithme de gradient à pas optimal

L'algorithme de gradient consiste à "se déplacer" d'une itérée u^n en suivant la ligne de plus grande pente associée à la fonction coût $J(v)$. La direction de descente correspondant à cette ligne de plus grande pente issue de u^n est donnée par l'opposé du gradient $J'(u^n)$. En effet, si l'on cherche u^{n+1} sous la forme

$$u^{n+1} = u^n - \mu^n w^n \ , \tag{3.2}$$

avec $\mu^n > 0$ petit et w^n unitaire dans V, un développement de Taylor à l'ordre 1,

$$J(u^{n+1}) = J(u^n) - \mu^n \langle J'(u^n), w^n\rangle + o(\mu^n),$$

montre que le choix de la direction $w_n = \dfrac{J'(u^n)}{\|J'(u^n)\|}$ permet de trouver la plus petite valeur de $J(u^{n+1})$ si on néglige le terme de reste (c'est-à-dire en l'absence d'autres informations comme les dérivées supérieures ou les itérées antérieures).

Cette remarque simple nous conduit, parmi les méthodes du type (3.2) qui sont appelées "méthodes de descente", à l'algorithme de **gradient à pas optimal**, dans lequel on résout une succession de problème de minimisation à une seule variable réelle (même si V n'est pas de dimension finie). A partir de u^0 quelconque dans V, on construit la suite (u^n) définie par

$$u^{n+1} = u^n - \mu^n \, J'(u^n) \ , \tag{3.3}$$

où $\mu^n \in \mathbb{R}$ est choisi à chaque étape tel que

$$J(u^{n+1}) = \inf_{\mu\in R} J\big(u^n - \mu \, J'(u^n)\big) \ . \tag{3.4}$$

Notons que l'on n'a pas normalisé le vecteur gradient dans (3.3). Cet algorithme converge comme l'indique le résultat suivant.

Théorème 3.2.1 *On suppose que J est α-convexe différentiable et que J' est Lipschitzien sur tout borné de V, c'est-à-dire que*

$$\forall\, M > 0,\ \exists\, C_M > 0,\ \|v\| + \|w\| \leq M \ \Rightarrow\ \|J'(v) - J'(w)\| \leq C_M \|v - w\|\,. \quad (3.5)$$

Alors l'algorithme de gradient à pas optimal converge : quel que soit u^0, la suite (u^n) définie par (3.3) et (3.4) converge vers la solution u de (3.1).

Démonstration. La fonction $f(\mu) = J\big(u^n - \mu\, J'(u^n)\big)$ est fortement convexe et dérivable sur $\mathbb{R}$ (si $J'(u^n) \neq 0$; sinon, on a déjà convergé, $u^n = u$!). Le problème de minimisation (3.4) a donc bien une solution unique, caractérisée par la condition $f'(\mu) = 0$, ce qui s'écrit aussi

$$\langle J'(u^{n+1}),\, J'(u^n)\rangle = 0\,. \quad (3.6)$$

Ceci montre que deux "directions de descente" consécutives sont orthogonales.

Puisque (3.6) implique que $\langle J'(u^{n+1}), u^{n+1} - u^n\rangle = 0$, on déduit de l'$\alpha$-convexité de J que

$$J(u^n) - J(u^{n+1}) \geq \frac{\alpha}{2}\|u^n - u^{n+1}\|^2\,, \quad (3.7)$$

ce qui prouve que la suite $J(u^n)$ est décroissante. Comme elle est minorée par $J(u)$, elle converge et (3.7) montre que $u^{n+1} - u^n$ tend vers 0. D'autre part, l'α-convexité de J et le fait que la suite $J(u^n)$ est bornée montrent que la suite (u^n) est bornée : il existe une constante M telle que

$$\|u^n\| \leq M\,.$$

Écrivant (3.5) pour $v = u^n$ et $w = u^{n+1}$ et utilisant (3.6), on obtient

$$\|J'(u^n)\|^2 \leq \|J'(u^n)\|^2 + \|J'(u^{n+1})\|^2 = \|J'(u^n) - J'(u^{n+1})\|^2 \leq C_M^2 \|u^{n+1} - u^n\|^2,$$

ce qui prouve que $J'(u^n)$ tend vers 0. L'α-convexité de J donne alors

$$\alpha\|u^n - u\|^2 \leq \langle J'(u^n) - J'(u), u^n - u\rangle = \langle J'(u^n), u^n - u\rangle \leq \|J'(u^n)\|\ \|u^n - u\|\,,$$

qui implique $\alpha\|u^n - u\| \leq \|J'(u^n)\|$, d'où l'on déduit la convergence de l'algorithme. □

Remarque 3.2.2 Il est utile de noter l'intérêt pratique de la dernière inégalité de cette démonstration : outre la preuve de la convergence, elle donne une majoration aisément calculable de l'erreur $u^n - u$. •

Remarque 3.2.3 Si J est de classe C^2 sur V, alors la condition (3.5) du théorème est automatiquement vérifiée. •

Remarque 3.2.4 Plaçons nous dans le cas où $V = \mathbb{R}^N$. Si P est une matrice symétrique définie positive de taille $N \times N$, alors $PJ'(u^n)$ est aussi une direction de descente car un développement de Taylor pour la récurrence

$$u^{n+1} = u^n - \mu^n P J'(u^n)$$

conduit à

$$J(u^{n+1}) = J(u^n) - \mu^n \langle PJ'(u^n), J'(u^n)\rangle + o(\mu^n) < J(u^n)$$

si $\mu^n > 0$ est petit et $J'(u^n) \neq 0$. Il peut être intéressante d'utiliser un tel **préconditionnement** du gradient pour améliorer la convergence. Par exemple, si les ordres de grandeur des différentes composantes de u, et donc de $J'(u)$, sont très différents, on peut utiliser une telle matrice P diagonale pour remettre à la même échelle ces composantes. Plus généralement, le choix d'une telle matrice revient à changer le produit scalaire de V avec lequel on identifie le gradient. De ce point de vue, cette approche se généralise aux espaces de Hilbert V de dimension infinie : on change la direction de descente si on change le produit scalaire de V. •

Remarque 3.2.5 La recherche du pas optimal $\mu^n \in \mathbb{R}$ dans (3.4) est un problème très classique, appelé recherche linéaire ou en ligne. Il existe de très nombreux algorithmes, rigoureux ou heuristiques, pour sa résolution (voir [22]). •

3.2.2 Algorithme de gradient à pas fixe

L'algorithme de gradient à pas fixe consiste simplement en la construction d'une suite u^n définie par

$$u^{n+1} = u^n - \mu\, J'(u^n)\ , \tag{3.8}$$

où μ est un paramètre positif fixé. Cette méthode est donc plus simple que l'algorithme de gradient à pas optimal, puisqu'on fait à chaque étape l'économie de la résolution de (3.4). Le résultat suivant montre sous quelles hypothèses on peut choisir le paramètre μ pour assurer la convergence.

Théorème 3.2.6 *On suppose que J est α-convexe différentiable et que J' est Lipschitzien sur V, c'est-à-dire qu'il existe une constante $C > 0$ telle que*

$$\|J'(v) - J'(w)\| \leq C\|v - w\| \quad \forall\, v, w \in V\ . \tag{3.9}$$

Alors, si $0 < \mu < 2\alpha/C^2$, l'algorithme de gradient à pas fixe converge : quel que soit u^0, la suite (u^n) définie par (3.8) converge vers la solution u de (3.1).

Démonstration. Posons $v^n = u^n - u$. Comme $J'(u) = 0$, on a $v^{n+1} = v^n - \mu\big(J'(u^n) - J'(u)\big)$, d'où il vient

$$\begin{aligned} \|v^{n+1}\|^2 &= \|v^n\|^2 - 2\mu\langle J'(u^n) - J'(u), u^n - u\rangle + \mu^2 \big\|J'(u^n) - J'(u)\big\|^2 \\ &\leq \big(1 - 2\alpha\mu + C^2\mu^2\big)\ \|v^n\|^2\ , \end{aligned} \tag{3.10}$$

d'après (3.9) et l'α-convexité. Si $0 < \mu < 2\alpha/C^2$, il est facile de voir que $1 - 2\alpha\mu + C^2\mu^2 \in]0, 1[$, et la convergence se déduit de (3.10). De manière équivalente, la même démonstration montre que l'application $v \mapsto v - \mu J'(v)$ est strictement contractante lorsque $0 < \mu < 2\alpha/C^2$, donc elle admet un unique point fixe (qui n'est autre que u) vers lequel converge la suite u^n. □

Remarque 3.2.7 L'hypothèse (3.9) est vérifiée si la fonction J est de classe C^2 et que $J''(u)(w,w) \leq C\|w\|^2$ pour tout $u, w \in V$.

L'algorithme de gradient à pas fixe est plus simple que celui à pas optimal puisqu'il ne nécessite pas de résoudre une optimisation uni-dimensionnelle (3.4) à chaque itération (on parle d'une recherche linéaire ou en ligne). Par contre, on ne connait en général pas de bonne estimation des constantes α et C et on ne sait pas a priori si un choix de pas de descente $\mu > 0$ est suffisamment petit. Il faut noter aussi que, pour l'algorithme de gradient à pas fixe, à la différence du gradient à pas optimal, la suite $J(u^n)$ n'est pas nécessairement monotone. •

Remarque 3.2.8 La démonstration du Théorème 3.2.6 permet d'obtenir une estimation de la **vitesse de convergence** de cet algorithme. Cette vitesse permet de fixer le nombre d'itérations n nécessaires pour rendre l'erreur $\|u^n - u\|$ inférieure à une tolérance ϵ fixée a priori. En effet, l'inégalité (3.10) montre que la convergence de l'algorithme de gradient à pas fixe est au moins géométrique, puisque

$$\|u^n - u\| \leq \gamma^n \|u^0 - u\| \quad \text{avec} \quad \gamma = \sqrt{1 - 2\alpha\mu + \mu^2 C^2}.$$

Sous réserve d'une analyse plus poussée conduisant à une meilleure estimation, on peut optimiser le choix du pas de descente μ dans l'intervalle $]0, 2\alpha/C^2[$ qui minimise la valeur de γ. On trouve facilement que le pas optimal est $\mu_{opt} = \alpha/C^2$ qui conduit à

$$\gamma_{opt} = \sqrt{1 - \frac{\alpha^2}{C^2}}.$$

Comme la quantité $\|u^n - u\|^{1/n}$ a une limite finie, égale à γ, avec $0 < \gamma < 1$, lorsque n tend vers $+\infty$, on dit que la convergence est géométrique. On peut montrer par des exemples ad hoc que l'algorithme de gradient à pas fixe ne peut pas converger plus vite que cette convergence géométrique. On verra par la suite d'autres algorithmes où on peut améliorer la constante γ_{opt} dans la convergence géométrique (par exemple le gradient conjugué), voire où on peut améliorer la convergence géométrique qui devient quadratique (exemple de la méthode de Newton). •

Exercice 3.2.1 On se place sous les hypothèses du Théorème 3.2.6 et on reprend les notations de la Remarque 3.2.8. En utilisant la définition (3.8) de l'algorithme, montrer qu'il existe une constante C_0 telle que

$$\|J'(u^n)\| \leq C_0 \gamma^n,$$

et, en utilisant la convexité de J au point u^n,

$$0 \leq J(u^n) - J(u) \leq C_0 \gamma^{2n} \quad \text{avec} \quad \gamma = \sqrt{1 - 2\alpha\mu + \mu^2 C^2}.$$

Exercice 3.2.2 Pour $V = \mathbb{R}^2$ et $J(x,y) = ax^2 + by^2$ avec $a, b > 0$, montrer que l'algorithme de gradient à pas optimal converge en une seule itération si $a = b$ ou si $x^0 y^0 = 0$, et que la convergence est géométrique dans les autres cas. Étudier aussi la convergence de l'algorithme de gradient à pas fixe : pour quelles valeurs du paramètre μ la convergence se produit-elle, pour quelle valeur est-elle la plus rapide ?

Exercice 3.2.3 Soit A une matrice symétrique définie positive de taille $N \times N$, soit $b \in \mathbb{R}^N$. On considère l'algorithme du gradient à pas fixe pour la fonction définie sur $\mathbb{R}^N$ par

$$J(x) = \frac{1}{2}Ax \cdot x - b \cdot x\,.$$

Récrire l'algorithme sous la forme $(x^n - x^*) = B(x^{n-1} - x^*)$ où $x^* = A^{-1}b$ et B est une matrice que l'on précisera. En étudiant le rayon spectral de B montrer que l'on peut obtenir une meilleure vitesse de convergence de l'algorithme en optimisant le choix du pas de descente μ, à savoir

$$\|x^n - x^*\| \leq \gamma_B^n \|x^0 - x^*\| \quad \text{avec} \quad \gamma_B = \sqrt{\frac{\lambda_N - \lambda_1}{\lambda_N + \lambda_1}},$$

où $0 < \lambda_1 \leq \lambda_N$ sont respectivement la plus petite et la plus grande valeur propre de A. Comparer avec γ_{opt} de la Remarque 3.2.8 (identifier les valeurs C et α pour la fonction $J(x)$).

3.2.3 Autres algorithmes du premier ordre

Il existe de nombreux autres algorithmes de descente qui à chaque itération n'utilise que l'information du gradient. Typiquement ils différent des algorithmes de gradient précédents car la nouvelle itération u^{n+1} ne dépend pas que de la solution courante u^n mais aussi de la précédente u^{n-1} (on parle alors de méthode à deux pas). Nous donnons quelques exemples simples ici.

Algorithme de la boule pesante

L'algorithme de la boule pesante trouve son origine dans interprétation d'une récurrence comme un schéma de discrétisation explicite en temps. Par exemple, l'algorithme de gradient à pas fixe (3.8) peut se voir comme un schéma explicite pour l'équation différentielle

$$\dot{u}(t) = -J'(u(t)),$$

pour peu qu'on considère le pas de descente μ comme un pas de temps (voir [1] pour des détails sur cette notion). Cette équation différentielle ordinaire modélise la trajectoire d'un point qui descendrait le long du potentiel $J(u)$ (on peut vérifier que $J(u(t))$ est décroissant le long de la trajectoire ; c'est ce qu'on appelle une fonction de Lyapunov). Si la fonction J n'est pas convexe et présente des minima locaux, la solution $u(t)$ peut converger vers un tel minimum local. Une idée est de changer l'équation différentielle, pour y rajouter un terme d'inertie, qui permettrait au point matériel solution de cette nouvelle équation de sortir d'un minimum local en remontant la pente grâce à l'inertie acquise. Autrement dit, l'équation différentielle de cette boule pesante serait

$$m\ddot{u}(t) + \dot{u}(t) = -J'(u(t)),$$

où $m \geq 0$ serait la masse de cette boule. Par analogie avec une discrétisation explicite en temps, on obtiendrait

$$u^{n+1} = u^n - \mu\, J'(u^n) + \nu(u^n - u^{n-1}), \tag{3.11}$$

où $\mu > 0$ et $\nu > 0$ sont deux paramètres positifs. Bien sûr, il faut joindre à cette récurrence pour $n \geq 1$ deux initialisations $u^1, u^0 \in V$.

Théorème 3.2.9 *On suppose que J est convexe, de classe C^2 tel que, pour $0 < \alpha \leq C$,*

$$\alpha\,\mathrm{Id} \leq J''(u) \leq C\,\mathrm{Id}.$$

Alors, si u^1, u^0 sont suffisamment proche de l'unique point de minimum u de J, pour $0 \leq \nu < 1$ et $0 < \mu < 2(1+\nu)/C$, l'algorithme de la boule pesante converge et, pour $\nu = (\frac{\sqrt{C}-\sqrt{\alpha}}{\sqrt{C}+\sqrt{\alpha}})^2$ et $\mu = 4/(\sqrt{C}+\sqrt{\alpha})^2$, on a

$$\|u^n - u\| \leq C\gamma^n \|u^0 - u\| \quad \textit{avec} \quad \gamma = \frac{\sqrt{C}-\sqrt{\alpha}}{\sqrt{C}+\sqrt{\alpha}}.$$

Remarque 3.2.10 Il est intéressant de comparer la vitesse de convergence du Théorème 3.2.9 avec celle de l'algorithme du gradient à pas fixe, fournie à la Remarque 3.2.8, qui s'écrit

$$\|u^n - u\| \leq \gamma_{opt}^n \|u^0 - u\| \quad \text{avec} \quad \gamma_{opt} = \sqrt{1 - \frac{\alpha^2}{C^2}}.$$

On voit donc, comme $0 < \alpha/C < 1$, que la méthode de la boule pesante avec ces paramètres converge asymptotiquement plus vite. Malheureusement, les valeurs de α et C ne sont souvent pas connues en pratique. •

Remarque 3.2.11 Nesterov [21] a proposé un autre algorithme qui accélère la convergence de la méthode du gradient

$$\begin{cases} v^n = u^n + \dfrac{a^n - 1}{a^{n+1}}(u^n - u^{n-1}) \\ u^{n+1} = v^n - \mu\, J'(v^n), \end{cases} \tag{3.12}$$

avec $a^n = 1 + n/2$ et un pas fixe $\mu > 0$ et que l'on initialise avec $u^0 = u^{-1} \in V$. On peut améliorer cet algorithme en changeant la formule pour le coefficient de relaxation a^n, du moins si l'on connait les valeurs de constantes α et C telles que $\alpha\,\mathrm{Id} \leq J''(v) \leq C\,\mathrm{Id}$ (ce qui n'est pas le cas le plus souvent en pratique). Avec une valeur optimale de a^n, l'algorithme de Nesterov converge avec le même taux que l'algorithme de la boule pesante. •

Algorithme du gradient conjugué

La méthode du gradient conjugué est une méthode du premier ordre où la direction de descente dans (3.2) n'est pas le gradient mais une combinaison du

gradient et de la direction de descente précédente (c'est donc encore une méthode multi-pas). On introduit donc une suite supplémentaire $p^n \in V$ pour la direction de descente. L'algorithme du gradient conjugué s'écrit

$$\left\{ \begin{array}{l} u^{n+1} = u^n - \mu^n\, p^n \\ p^{n+1} = J'(u^{n+1}) + \beta^n p^n \end{array} \right. \tag{3.13}$$

où $\mu^n \in \mathbb{R}$ est le pas optimal qui minimise $\mu \to J(u^n - \mu\, p^n)$, et $\beta^n \in \mathbb{R}$ est un coefficient que l'on peut calculer suivant deux formules différentes. La première formule est dite de Polak-Ribière

$$\beta^n = \frac{\langle J'(u^{n+1}), J'(u^{n+1}) - J'(u^n)\rangle}{\|J'(u^n)\|^2},$$

tandis que la seconde est appelée Fletcher-Rieves

$$\beta^n = \frac{\|J'(u^{n+1})\|^2}{\|J'(u^n)\|^2}.$$

Expliquons d'où viennent ces formules en quelques mots. Tout d'abord, la recherche linéaire pour trouver le pas μ^n est classique et réminiscente de l'algorithme du gradient à pas optimal. La formule pour β^n s'inspire du cas où $J(u)$ est quadratique. Si $J(u)$ est une fonction linéaire quelconque, l'algorithme du gradient conjugué fonctionnera néanmoins approximativement comme dans le cas quadratique au voisinage d'un point de minimum non dégénée (c'est-à-dire où la Hessienne est définie positive). Nous allons donc formellement supposer que la Hessienne $J''(u)$ est indépendante de u et égale à une matrice constante A. L'idée essentielle est de demander à ce que les directions de descente p^n et p^{n+1} soient conjuguées par rapport à la matrice A, c'est-à-dire que $p^{n+1} \cdot Ap^n = 0$. Comme expliqué ci-dessous (voir l'Exercice 3.2.4), cette propriété est à la source de la convergence de l'algorithme du gradient conjugué pour une fonction quadratique et on essaye donc de la reproduire dans le cas général.

On souhaite donc que la valeur de β^n soit telle que

$$p^{n+1} \cdot Ap^n = 0.$$

On remplace p^{n+1} par $J'(u^{n+1}) + \beta^n p^n$ pour obtenir

$$\beta^n = -\frac{J'(u^{n+1}) \cdot Ap^n}{Ap^n \cdot p^n},$$

puis, comme $p^n = (u^n - u^{n+1})/\mu^n$, on a

$$\mu^n Ap^n = Au^n - Au^{n+1} = J'(u^n) - J'(u^{n+1})$$

puisque la Hessienne est constante et égale à A. Or la condition d'optimalité pour μ^n dit que $J'(u^{n+1}) \cdot p^n = 0$ tandis que $J'(u^n) \cdot p^n = J'(u^n) \cdot (J'(u^n) + \beta_{n-1} p^{n-1})$ et $J'(u^n) \cdot p^{n-1} = 0$ à cause de l'optimalité de μ^{n-1}. Donc

$$Ap^n \cdot p^n = \|J'(u^n)\|^2/\mu^n.$$

Par ailleurs,

$$J'(u^{n+1}) \cdot Ap^n = J'(u^{n+1}) \cdot (J'(u^n) - J'(u^{n+1}))/\mu^n,$$

d'où l'on déduit la formule de Polak-Ribière. Un calcul similaire conduit à la formule de Fletcher-Rieves. Remarquons que ces deux formules ne sont pas égales dans le cas général mais qu'elles coincident pour une fonction $J(u)$ quadratique.

En fait, la méthode du gradient conjugué a d'abord été inventée pour résoudre des systèmes linéaires dont la matrice est symétrique, définie positive, c'est-à-dire pour minimiser sur $\mathbb{R}^n$ une fonction quadratique

$$J(x) = \frac{1}{2}Ax \cdot x - b \cdot x$$

avec $b \in \mathbb{R}^n$ et A une matrice symétrique définie positive. C'est dans ce cas quadratique que l'algorithme du gradient conjugué est le plus efficace puisqu'on peut même démontrer qu'il converge exactement en au plus n itérations ! Dans ce cas quadratique, les formules (3.13) se simplifient en

$$\text{pour } k \geq 0 \quad \begin{cases} x_{k+1} = x_k - \mu_k p_k \\ r_{k+1} = r_k - \mu_k A p_k \\ p_{k+1} = r_{k+1} + \beta_k p_k \end{cases} \tag{3.14}$$

avec une initialisation $x_0 \in \mathbb{R}^n$, $p_0 = r_0 = Ax_0 - b$ et

$$\mu_k = \frac{\|r_k\|^2}{Ap_k \cdot p_k} \text{ et } \beta_k = \frac{\|r_{k+1}\|^2}{\|r_k\|^2}.$$

On vérifie facilement que α_k est précisément l'unique point de minimum du polynôme quadratique $\mu \to J(x_k - \mu p_k)$. On peut vérifier par récurrence que la suite r_k coincide avec le gradient $J'(x_k)$ et donc que β_k est donné par la formule de Fletcher-Rieves.

Exercice 3.2.4 Soit A une matrice symétrique définie positive. Soit $x_0 \in \mathbb{R}^n$. Pour la récurrence (3.14) démontrer que $r_k = Ax_k - b$. Montrer que la suite p_k vérifie $Ap_k \cdot p_j = 0$ pour $0 \leq j < k$ (on dit que la suite p_k est conjuguée par rapport à A). Vérifier que les formules de Fletcher-Rieves et Polak-Ribière coincident dans ce cas.

Le fait que la suite des directions de descente p_k soit conjuguée par rapport à A donne son nom à l'algorithme. C'est une propriété cruciale pour l'efficacité de la méthode car elle assure que la nouvelle direction de descente p_k "travaille" de manière orthogonale (au sens du produit scalaire $\langle x, y \rangle_A = Ax \cdot y$) par rapport aux directions précédentes. Le nouveau gain de minimisation ne nuit pas aux gains passés (pas d'effet de convergence en zig-zag) et c'est la raison du résultat de convergence exacte en moins de n itérations. Néanmoins, la convergence approchée peut avoir lieu en beaucoup moins d'itérations comme le montre le résultat suivant que nous admettrons.

Proposition 3.2.12 *Soit A une matrice symétrique réelle définie positive. Soit x la solution exacte du système $Ax = b$. Soit $(x_k)_{k\geq 0}$ la suite de solutions approchées du gradient conjugué. Alors*

$$\|x_k - x\|_2 \leq 2\sqrt{\text{cond}(A)} \left(\frac{\sqrt{\text{cond}(A)} - 1}{\sqrt{\text{cond}(A)} + 1} \right)^k \|x_0 - x\|_2,$$

où $\text{cond}(A) = \lambda_n/\lambda_1$, avec $\lambda_1 \leq \lambda_n$ la plus petite et la plus grande valeur propre de A, est appelé le conditionnement de A.

Le conditionnement d'une matrice A vérifie $\text{cond}(A) \geq 1$ et, en général, il peut être très grand pour de nombreuses applications. Dans ce cas, on a

$$\frac{\sqrt{\text{cond}(A)} - 1}{\sqrt{\text{cond}(A)} + 1} < \sqrt{\frac{\text{cond}(A) - 1}{\text{cond}(A) + 1}},$$

c'est-à-dire que la Proposition 3.2.12 améliore la vitesse de convergence (toujours géométrique) de la méthode du gradient à pas fixe (voir la Remarque 3.2.8 et l'Exercice 3.2.3). En pratique, la méthode du gradient conjugué converge plus vite que celle du gradient.

Remarque 3.2.13 Comme la vitesse de convergence de la méthode du gradient conjugué dépend du conditionnement de la matrice A, une idée pour améliorer sa rapidité est de préconditionner le système linéaire $Ax = b$ en le pré-multipliant par une matrice C^{-1} telle que le conditionnement de $(C^{-1}A)$ soit plus petit que celui de A. En pratique on choisit une matrice C "proche" de A mais plus facile à inverser (voir [1] pour des détails). •

Algorithme de sous-gradient

Les méthodes de gradient peuvent se généraliser au cas des fonctions qui ne sont pas différentiables mais sont convexes. Cette généralisation s'appelle la **méthode de sous-gradient**. Il nous faut d'abord définir la notion de **sous-gradient** pour les fonctions convexes. Par souci de simplicité on se limite à la dimension finie mais le cas des espaces de Hilbert n'est pas vraiment plus difficile.

Définition 3.2.14 *Si J est une fonction convexe de $\mathbb{R}^n$ dans $\mathbb{R}$, on appelle* **sous-différentiel** *de J en x l'ensemble*

$$\partial J(x) = \{p \in \mathbb{R}^n \mid J(y) - J(x) \geq p \cdot (y - x), \qquad \forall y \in \mathbb{R}^n\} \ . \tag{3.15}$$

Les éléments de $\partial J(x)$ sont appelés **sous-gradients**.

Il résulte aussitôt de cette définition que le sous-différentiel $\partial J(x)$ est un convexe fermé de $\mathbb{R}^n$. D'autre part, si J est dérivable au point x, on vérifie aisément que le sous-différentiel est un singleton, $\partial J(x) = \{J'(x)\}$.

Lemme 3.2.15 *Soit J une fonction convexe de $\mathbb{R}^n$ dans $\mathbb{R}$. Si $x \in \mathbb{R}^n$ vérifie $0 \in \partial J(x)$, alors x est un point de minimum de J.*

Démonstration. La démonstration est évidente par définition du sous-différentiel. Remarquons que la condition $0 \in \partial J(x)$ généralise au cas convexe non-différentiable la condition d'optimalité usuelle $J'(x) = 0$ (lorsque J est dérivable). □

Exercice 3.2.5 Soit $J(x) = |x|$ définie de $\mathbb{R}$ dans $\mathbb{R}$. Calculer son sous-gradient $\partial J(0)$.

Une classe importante de fonctions convexes, qui ne sont pas différentiables mais admettent un sous-gradient que l'on sait calculer, est celle des fonctions qui sont des maxima de fonctions convexes différentiables. C'est une situation courante puisque, par exemple, on sait, d'après l'Exercice 2.3.2, que toute fonction convexe est le supremum des fonctions affines qui la minorent. Plus généralement, pour un espace (non vide) de paramètres Λ, on définit

$$\mathcal{J}(x) = \sup_{\lambda \in \Lambda} J_\lambda(x), \tag{3.16}$$

où chaque fonction $J_\lambda(x)$ est convexe et différentiable sur $\mathbb{R}^n$. Pour simplifier, on va supposer que l'ensemble des paramètres Λ est fini. On calcul alors aisément un sous-gradient de $\mathcal{J}$.

Lemme 3.2.16 *Supposons Λ fini, soit $\mathcal{J}$ la fonction définie par* (3.16) *avec des fonctions $J_\lambda(x)$ convexes et différentiables, et pour tout $x \in \mathbb{R}^n$, posons $\Gamma(x) = \{\lambda \in \Lambda \mid J_\lambda(x) = \mathcal{J}(x)\}$. Alors, pour tout $\lambda \in \Gamma(x)$, $J'_\lambda(x)$ est un sous-gradient de $\mathcal{J}$ au point x.*

Démonstration. Pour tout $\lambda \in \Gamma(x)$, et pour tout $y \in \mathbb{R}^n$, on a $\mathcal{J}(y) - \mathcal{J}(x) \geq J_\lambda(y) - J_\lambda(x) \geq J'_\lambda(x) \cdot (y - x)$, par convexité de J_λ, ce qui démontre le résultat. □

Remarque 3.2.17 Le Lemme 3.2.16 est une version faible du résultat suivant sur les sous-différentiels de maxima de fonctions convexes. Sous les mêmes hypothèses, pour tout $x \in \mathbb{R}^n$, on a en fait

$$\partial \mathcal{J}(x) = \mathrm{co}\Big(\bigcup_{\lambda \in \Gamma(x)} \partial J_\lambda(x) \Big) \ . \tag{3.17}$$

•

Remarque 3.2.18 Les fonctions non régulières du type (3.16) se retrouvent dans au moins deux contextes importants. D'une part, elles sont utilisées dans les méthodes de relaxations Lagrangiennes pour l'optimisation combinatoire ou en variables entières (voir [5]). D'autre part, elles correspondent à l'approche, dite **du pire des cas**, en optimisation robuste. Supposons que l'on veuille optimiser un système décrit par une variable x. Mais la description de ce système est entachée d'incertitudes, d'erreurs ou de données inconnues, caractérisées par λ. On peut identifier chaque valeur de λ à un scenario possible de fonctionnement du système. Si l'on veut optimiser pour tous les scenarios possibles, l'approche du pire des cas consiste à minimiser

le maximum par rapport à la variable λ. Notons que c'est une approche, assez pessimiste (on prévoit le pire !), et qu'elle est parfois remplacée par une approche **en moyenne** où on optimise la moyenne de $J_\lambda(x)$ sur Λ (ou bien une somme pondérée de la moyenne et de la variance). •

L'**algorithme de sous-gradient** pour minimiser la fonction convexe $\mathcal{J}$ consiste, pour une initialisation $x_0 \in \mathbb{R}^n$, à construire la suite

$$x_{k+1} = x_k - \frac{\rho_k}{\|p_k\|} p_k \ , \tag{3.18}$$

où p_k est un sous-gradient quelconque de $\mathcal{J}$ au point x_k, et où $\rho_k > 0$ est une suite de réels strictement positifs telle que

$$\rho_k \to 0, \quad \sum_{i\in\mathbb{N}} \rho_i = +\infty, \quad \sum_{i\in\mathbb{N}} \rho_i^2 < +\infty \ . \tag{3.19}$$

évidemment, la valeur x_{k+1} n'est bien définie que si $p_k \neq 0$. Lorsque $p_k = 0$, l'algorithme s'arrête : x_k est alors le minimum de $\mathcal{J}$ en vertu du Lemme 3.2.15. Un exemple de suite de pas ρ_k vérifiant (3.19) est $\rho_k = 1/\sqrt{k+1}$.

Proposition 3.2.19 *Sous les hypothèses du Lemme 3.2.16 et avec des pas vérifiant* (3.19)*, l'algorithme du sous-gradient converge.*

Démonstration. On note x_* l'unique point de minimum de $\mathcal{J}$ et on développe la norme suivante

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &= \|x_k - x_*\|^2 - 2\frac{\rho_k}{\|p_k\|} p_k \cdot (x_k - x_*) + \rho_k^2 \\ &\leq \|x_k - x_*\|^2 - 2\frac{\rho_k}{\|p_k\|} \left(\mathcal{J}(x_k) - \mathcal{J}(x_*)\right) + \rho_k^2 \end{aligned}$$

car p_k est un sous gradient de $\mathcal{J}$ au point x_k. On somme ces inégalités et une minoration évidente conduit à

$$\|x_{i+1} - x_*\|^2 + 2 \min_{0\leq k\leq i} \left(\mathcal{J}(x_k) - \mathcal{J}(x_*)\right) \sum_{k=0}^{i} \frac{\rho_k}{\|p_k\|} \leq \|x_0 - x_*\|^2 + \sum_{k=0}^{i} \rho_k^2.$$

En utilisant, l'hypothèse que $\mathcal{J}$ est globalement L-Lipschitzienne et (3.19) on en déduit

$$2 \min_{0\leq k\leq i} \left(\mathcal{J}(x_k) - \mathcal{J}(x_*)\right) \leq L \frac{\|x_0 - x_*\|^2 + \sum_{k=0}^{\infty} \rho_k^2}{\sum_{k=0}^{i} \rho_k},$$

ce qui démontre le résultat puisque la série des ρ_k diverge. La vitesse de convergence est très lente puisque pour $\rho_k = 1/\sqrt{k+1}$ on trouve que le membre de droite tends vers zéro comme $1/\sqrt{i}$, ce qui est bien plus lent que la convergence géométrique pour l'algorithme du gradient (voir la Remarque 3.2.8). □

Remarque 3.2.20 Dans l'algorithme de sous-gradient (3.18) on normalise le sous-gradient p_k. En effet, comme l'a montré l'exemple de l'Exercice 3.2.5, la norme d'un sous-gradient peut-être très variable et la valeur de cette norme ne donne aucun indication sur l'optimalité ou non du point où on l'a calculé. •

Gradient stochastique

Il s'agit d'un algorithme qui est dédié à une classe particulière de problèmes d'optimisation, très utile pour l'apprentissage machine, dont une situation typique est présentée dans l'Exemple 1.2.6. Dans ce contexte, il est naturel de se restreindre à la dimension finie $V = \mathbb{R}^d$. La spécificité de ce type de problèmes est que la fonction à minimiser est une somme (ou une moyenne) d'un très grand nombre n de fonctions. Par exemple, en apprentissage machine n est le nombre de données à partir desquelles on veut apprendre les paramètres $x \in \mathbb{R}^d$ du modèle explicatif. Autrement dit, étant données n fonctions f_i, on considère

$$\inf_{x\in\mathbb{R}^d} F(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x). \tag{3.20}$$

Rappelons que l'algorithme du gradient "classique" (qu'on appelle dans ce contexte, algorithme de *batch*) construit la suite

$$x^{k+1} = x^k - \frac{\mu^k}{n}\sum_{i=1}^{n} f_i'(x^k)$$

où $\mu^k > 0$ est un pas de descente (fixe ou optimal). Par contraste, **l'algorithme du gradient stochastique** n'utilise qu'une seule dérivée de fonction f_i par itération. Autrement dit, à partir d'une initialisation $x^0 \in \mathbb{R}^d$, on construit la suite

$$x^{k+1} = x^k - \mu^k f_{i_k}'(x^k), \tag{3.21}$$

où $\mu^k > 0$ est un pas de descente et i_k est un indice tiré aléatoirement (indépendamment des précédents tirages) et uniformément dans l'ensemble $\{1, ..., n\}$.

Proposition 3.2.21 *On suppose que $F(x)$ est α-convexe et qu'il existe une constante C, indépendante de n, telle que pour tout $x \in \mathbb{R}^d$*

$$\frac{1}{n}\sum_{i=1}^{n} \|f_i'(x)\|^2 \leq C(1 + \|x - x^*\|^2) \tag{3.22}$$

où x^* désigne l'unique point de minimum de $F(x)$. Si la suite des pas vérifie

$$\mu^k = \frac{1}{k+1},$$

alors l'algorithme du gradient stochastique (3.21) *converge.*

Remarque 3.2.22 Les hypothèses sur F sont vérifiées dans le cas de l'Exemple 1.2.6 (car chacune des dérivées $f_i'(x)$ est bornée) ou bien si chaque fonction f_i est quadratique et fortement convexe, uniformément par rapport à i. •

Démonstration. On réécrit (3.21) sous la forme

$$x^{k+1} - x^* = x^k - x^* - \mu^k f'_{i_k}(x^k),$$

dont on calcule la norme au carré

$$\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 - 2\mu^k (x^k - x^*) \cdot f'_{i_k}(x^k) + (\mu^k)^2 \|f'_{i_k}(x^k)\|^2. \tag{3.23}$$

On prend l'espérance de (3.23) par rapport à la seule variable aléatoire qui gouverne le choix de l'indice i_k, en notant que x^k ne dépend pas de cette variable aléatoire, pour obtenir

$$\mathbb{E}\Big(\|x^{k+1} - x^*\|^2\Big) = \|x^k - x^*\|^2 - 2\mu^k (x^k - x^*) \cdot F'(x^k) + \frac{(\mu^k)^2}{n} \sum_{i=1}^{n} \|f'_i(x^k)\|^2$$

car, par l'hypothèse de tirage uniforme de i_k, on a

$$\mathbb{E}\Big(f'_{i_k}(x^k)\Big) = \frac{1}{n} \sum_{i=1}^{n} f'_i(x^k) = F'(x^k) \quad \text{et} \quad \mathbb{E}\Big(\|f'_{i_k}(x^k)\|^2\Big) = \frac{1}{n} \sum_{i=1}^{n} \|f'_i(x^k)\|^2.$$

Par la forte convexité de F, voir (2.24), on en déduit

$$\|x^{k+1} - x^*\|^2 \leq \big(1 - 2\mu^k \alpha\big) \|x^k - x^*\|^2 + \frac{(\mu^k)^2}{n} \sum_{i=1}^{n} \|f'_i(x^k)\|^2.$$

Utilisant l'hypothèse (3.22), il vient

$$\|x^{k+1} - x^*\|^2 \leq \big(1 - 2\mu^k \alpha + C(\mu^k)^2\big) \|x^k - x^*\|^2 + C(\mu^k)^2.$$

On choisit le pas de descente $0 < \mu^k < 2\alpha/C$ de manière à ce que le coefficient d'amplification ρ^k soit strictement plus petit que 1

$$0 < \rho^k = 1 - 2\mu^k \alpha + C(\mu^k)^2 < 1.$$

Une récurrence facile montre alors que

$$\|x^{k+1} - x^*\|^2 \leq \Pi^k \|x^0 - x^*\|^2 + C\Pi^k \sum_{i=0}^{k} \frac{(\mu^i)^2}{\Pi^i}, \tag{3.24}$$

avec

$$\Pi^k = \prod_{j=0}^{k} \rho^j.$$

On voit immédiatement que pour démontrer la convergence de l'algorithme, non seulement le pas μ^k ne doit pas être trop grand pour que Π^k tende vers zéro mais il doit aussi lui-même tendre vers zéro pour que la série dans (3.24) converge. C'est

une situation similaire à celle de l'algorithme du sous-gradient (voir la Proposition 3.2.19). On vérifie que le choix $\mu^k = 1/(k+1)$ conduit à

$$\rho^k = 1 - \frac{2\alpha}{k+1} + \mathcal{O}(k^{-2}), \quad \Pi^k = \mathcal{O}(k^{-2\alpha}), \quad \Pi^k \sum_{i=0}^{k} \frac{(\mu^i)^2}{\Pi^i} = \mathcal{O}(k^{-1}),$$

ce qui prouve la convergence. □

Comme pour l'algorithme du sous-gradient la convergence est particulièrement lente (algébrique au lieu de géométrique pour le gradient à pas fixe). Du coup, il n'est pas clair que cet algorithme soit préférable à celui du gradient. Mais il ne faut pas oublier que si n est très grand, le coût d'une itération de gradient stochastique est approximativement n plus fois faible que le coût d'une itération du gradient puisqu'on n'évalue qu'une seule dérivée f_i' au lieu de n (pour les deux algorithmes on calcule aussi régulièrement la fonction objectif F pour vérifier sa décroissance et la bonne convergence des itérations). Par ailleurs, il existe des stratégies heuristiques du choix du pas de descente μ^k qui peuvent améliorer en pratique le choix théorique ci-dessus. En particulier, l'algorithme du gradient stochastique est souvent plus rapide lors des premières itérations que l'algorithme du gradient et surtout plus insensible au "bruit" (qu'il soit numérique ou dans les données).

Remarque 3.2.23 Entre l'algorithme du gradient stochastique (3.21) et l'algorithme classique du gradient (dit "batch") on peut proposer un algorithme de **mini-batch** qui sélectionne une collection de $m \geq 1$ indices

$$x^{k+1} = x^k - \mu^k \frac{1}{m} \sum_{i \in \mathcal{I}_k} f_i'(x^k), \tag{3.25}$$

où $\mu^k > 0$ est un pas de descente et $\mathcal{I}_k$ est une collection de m indices distincts tirés aléatoirement (indépendamment des précédents tirages) et uniformément dans l'ensemble $\{1, ..., n\}$. •

3.3 Algorithmes de type gradient (cas avec contraintes)

On étudie maintenant la résolution de problèmes d'optimisation avec contraintes

$$\inf_{v \in K} J(v) \ , \tag{3.26}$$

où J est une fonction α-convexe différentiable définie sur K, sous-ensemble convexe fermé non vide de l'espace de Hilbert réel V. Le Théorème 2.3.8 assure alors l'existence et l'unicité de la solution u de (3.26), caractérisée d'après le Théorème 2.5.1 par la condition

$$\langle J'(u), v - u \rangle \geq 0 \quad \forall\, v \in K \ . \tag{3.27}$$

Selon les algorithmes étudiés ci-dessous, nous serons parfois amenés à préciser des hypothèses supplémentaires sur l'ensemble K.

3.3.1 Algorithme de gradient à pas fixe avec projection

L'algorithme de gradient à pas fixe s'adapte au cas du problème (3.26) avec contraintes à partir de la remarque suivante. Pour tout réel $\mu > 0$, (3.27) s'écrit

$$\langle u - \big(u - \mu J'(u)\big), v - u\rangle \geq 0 \quad \forall\, v \in K\ . \tag{3.28}$$

Notons P_K l'opérateur de projection sur l'ensemble convexe K, défini au Théorème 8.1.3 de projection sur un convexe (voir la Remarque 8.1.4). Alors, d'après ce théorème, (3.28) n'est rien d'autre que la caractérisation de u comme la projection orthogonale de $u - \mu J'(u)$ sur K. Autrement dit,

$$u = P_K\big(u - \mu J'(u)\big) \quad \forall\, \mu > 0\ . \tag{3.29}$$

Il est facile de voir que (3.29) est en fait équivalent à (3.27), et caractérise donc la solution u de (3.26). L'algorithme de **gradient à pas fixe avec projection** (ou plus simplement de gradient projeté) est alors défini par l'itération

$$u^{n+1} = P_K\big(u^n - \mu J'(u^n)\big)\ , \tag{3.30}$$

où μ est un paramètre positif fixé.

Théorème 3.3.1 *On suppose que J est α-convexe différentiable et que J' est Lipschitzien sur V (de constante C, voir (3.9)). Alors, si $0 < \mu < 2\alpha/C^2$, l'algorithme de gradient à pas fixe avec projection converge : quel que soit $u^0 \in K$, la suite (u^n) définie par (3.30) converge vers la solution u de (3.26).*

Démonstration. La démonstration reprend celle du Théorème 3.2.6 où l'on a montré que l'application $v \mapsto v - \mu J'(v)$ est strictement contractante lorsque $0 < \mu < 2\alpha/C^2$, c'est-à-dire que

$$\exists\, \gamma \in]0,1[\quad , \quad \big\|\big(v - \mu J'(v)\big) - \big(w - \mu J'(w)\big)\big\| \leq \gamma \|v - w\|\ .$$

Puisque la projection P_K est faiblement contractante d'après (8.2), l'application $v \mapsto P_K\big(v - \mu J'(v)\big)$ est strictement contractante, ce qui prouve la convergence de la suite (u^n) définie par (3.30) vers la solution u de (3.26). □

Exercice 3.3.1 Soit $V = \mathbb{R}^N$ et $K = \{x \in \mathbb{R}^N$ tel que $\sum_{i=1}^N x_i = 1\}$. Expliciter l'opérateur de projection orthogonale P_K et interpréter dans ce cas la formule (3.29) en terme de multiplicateur de Lagrange.

3.3.2 Algorithme d'Uzawa

Le résultat précédent montre que la méthode de gradient à pas fixe avec projection est applicable à une large classe de problèmes d'optimisation convexe avec contraintes. Mais cette conclusion est largement un leurre du point de vue pratique, car l'opérateur de projection P_K n'est pas connu explicitement en général : la projection d'un élément $v \in V$ sur un convexe fermé quelconque de V peut être très difficile à déterminer !

Une exception importante concerne, en dimension finie (pour $V = \mathbb{R}^M$), les sous-ensembles K de la forme

$$K = \prod_{i=1}^{M} [a_i, b_i] \tag{3.31}$$

(avec éventuellement $a_i = -\infty$ ou $b_i = +\infty$ pour certains indices i). En effet, il est alors facile de voir que, si $x = (x_1, x_2, \ldots, x_M) \in \mathbb{R}^M$, $y = P_K(x)$ a pour composantes

$$y_i = \min\left(\max\left(a_i, x_i\right), b_i\right) \quad \text{pour} \quad 1 \leq i \leq M \ , \tag{3.32}$$

autrement dit, il suffit juste de "tronquer" les composantes de x. Cette propriété simple, jointes aux remarques sur la dualité énoncée dans la Section 2.6, va nous conduire à un nouvel algorithme. En effet, même si le problème primal fait intervenir un ensemble K des solutions admissibles sur lequel la projection P_K ne peut être déterminée explicitement, le problème dual sera fréquemment posé sur un ensemble de la forme (3.31), typiquement sur $(\mathbb{R}_+)^M$. Dans ce cas, le problème dual peut être résolu par la méthode du gradient à pas fixe avec projection, et la solution du problème primal pourra ensuite être obtenue en résolvant un problème de minimisation **sans contrainte**. Ces remarques sont à la base de l'algorithme d'Uzawa, qui est en fait une méthode de recherche de point-selle.

Considérons le problème de minimisation convexe

$$\inf_{F(v) \leq 0} J(v) \ , \tag{3.33}$$

où J est une fonctionnelle convexe définie sur V et F une fonction convexe de V sur $\mathbb{R}^M$. Sous les hypothèses du Théorème de Kuhn et Tucker 2.6.4, la résolution de (3.33) revient à trouver un point-selle (u, p) du Lagrangien

$$\mathcal{L}(v, q) = J(v) + q \cdot F(v) \ , \tag{3.34}$$

sur $V \times (\mathbb{R}_+)^M$. A partir de la Définition 2.6.1 du point-selle

$$\forall\, q \in (\mathbb{R}_+)^M \quad \mathcal{L}(u, q) \leq \mathcal{L}(u, p) \leq \mathcal{L}(v, p) \quad \forall\, v \in V \ , \tag{3.35}$$

on déduit que $(p - q) \cdot F(u) \geq 0$ pour tout $q \in (\mathbb{R}_+)^M$, d'où on tire, pour tout réel $\mu > 0$,

$$(p - q) \cdot \big(p - \big(p + \mu F(u)\big)\big) \leq 0 \quad \forall\, q \in (\mathbb{R}_+)^M \ ,$$

ce qui, d'après (8.1), montre que

$$p = P_{\mathbb{R}_+^M}(p + \mu F(u)) \quad \forall\, \mu > 0 \ , \tag{3.36}$$

$P_{\mathbb{R}_+^M}$ désignant la projection de $\mathbb{R}^M$ sur $(\mathbb{R}_+)^M$.

Au vu de cette propriété et de la seconde inégalité dans (3.35), nous pouvons introduire **l'algorithme d'Uzawa** : à partir d'un élément quelconque $p^0 \in (\mathbb{R}_+)^M$, on construit les suites (u^n) et (p^n) déterminées par les itérations

$$\begin{aligned} &\mathcal{L}(u^n, p^n) = \inf_{v \in V} \mathcal{L}(v, p^n) \ , \\ &p^{n+1} = P_{\mathbb{R}_+^M}(p^n + \mu F(u^n)) \ , \end{aligned} \tag{3.37}$$

μ étant un paramètre positif fixé. On peut interpréter l'algorithme d'Uzawa en disant qu'alternativement il minimise le Lagrangien par rapport à v avec q fixé et il maximise (par un seul pas de l'algorithme du gradient projeté) ce même Lagrangien par rapport à q avec v fixé. Une autre manière de voir l'algorithme d'Uzawa est la suivante : il prédit une valeur du multiplicateur de Lagrange q et effectue une minimisation sans contrainte du Lagrangien par rapport à v, puis il corrige la prédiction de q en l'augmentant si la contrainte est violée et en le diminuant sinon. Nous verrons une troisième interprétation de l'algorithme d'Uzawa dans le cadre de la théorie de la dualité ci-dessous.

Théorème 3.3.2 *On suppose que J est α-convexe différentiable, que F est convexe et Lipschitzienne de V dans $\mathbb{R}^M$, c'est-à-dire qu'il existe une constante C telle que*

$$\|F(v) - F(w)\| \leq C\|v - w\| \quad \forall\, v, w \in V \;, \tag{3.38}$$

et qu'il existe un point-selle (u, p) du Lagrangien (3.34) sur $V \times (\mathbb{R}_+)^M$. Alors, si $0 < \mu < 2\alpha/C^2$, l'algorithme d'Uzawa converge : quel que soit l'élément initial p^0, la suite (u^n) définie par (3.37) converge vers la solution u du problème (3.33).

Démonstration. Rappelons d'abord que l'existence d'une solution u de (3.33) découle de celle du point-selle (u, p) (voir la Proposition 2.6.2), alors que son unicité est une conséquence de l'α-convexité de J. De même, p^n étant fixé, le problème de minimisation dans (3.37) a bien une solution unique u^n. D'après l'Exercice 2.5.6, les inéquations d'Euler satisfaites par u et u^n s'écrivent

$$\langle J'(u), v - u\rangle + p \cdot \big(F(v) - F(u)\big) \geq 0 \quad \forall\, v \in V \;, \tag{3.39}$$

$$\langle J'(u^n), v - u^n\rangle + p^n \cdot \big(F(v) - F(u^n)\big) \geq 0 \quad \forall\, v \in V \;. \tag{3.40}$$

Prenant successivement $v = u^n$ dans (3.39) et $v = u$ dans (3.40) et additionnant, on obtient

$$\langle J'(u) - J'(u^n), u^n - u\rangle + (p - p^n) \cdot \big(F(u^n) - F(u)\big) \geq 0 \;,$$

d'où en utilisant l'α-convexité de J et en posant $r^n = p^n - p$

$$r^n \cdot \big(F(u^n) - F(u)\big) \leq -\alpha\|u^n - u\|^2 \;. \tag{3.41}$$

D'autre part, la projection $P_{\mathbb{R}^M_+}$ étant faiblement contractante d'après (8.2), en soustrayant (3.36) à (3.37) on obtient

$$\|r^{n+1}\| \leq \|r^n + \mu(F(u^n) - F(u))\| \;,$$

soit

$$\|r^{n+1}\|^2 \leq \|r^n\|^2 + 2\mu r^n \cdot \big(F(u^n) - F(u)\big) + \mu^2\|F(u^n) - F(u)\|^2 \;.$$

Utilisant (3.38) et (3.41), il vient

$$\|r^{n+1}\|^2 \leq \|r^n\|^2 + \big(C^2\mu^2 - 2\mu\alpha\big)\|u^n - u\|^2 \;.$$

Si $0 < \mu < 2\alpha/C^2$, on peut trouver $\beta > 0$ tel que $C^2\mu^2 - 2\mu\alpha < -\beta$, d'où

$$\beta\|u^n - u\|^2 \leq \|r^n\|^2 - \|r^{n+1}\|^2 \ . \tag{3.42}$$

Ceci montre alors que la suite $\|r^n\|^2$ est décroissante : le membre de droite de (3.42) tend donc vers 0, ce qui entraîne que u^n tend vers u. □

Ainsi, l'algorithme d'Uzawa permet d'approcher la solution de (3.33) en remplaçant ce problème avec contraintes par une suite de problèmes de minimisation sans contraintes (3.37). A chaque itération, la détermination de p^n est élémentaire, puisque d'après (3.32) l'opérateur de projection $P_{\mathbb{R}_+^M}$ est une simple troncature à zéro des composantes négatives. Il faut aussi noter que le Théorème 3.3.2 ne dit rien de la convergence de la suite (p^n). En fait, cette convergence n'est pas assurée sous les hypothèses du théorème, qui n'assurent d'ailleurs pas l'unicité de l'élément $p \in (\mathbb{R}_+)^M$ tel que (u, p) soit point-selle (voir la Remarque 2.6.12 et l'Exercice 3.3.2 ci-dessous).

Il reste à faire le lien entre l'algorithme d'Uzawa et la théorie de la dualité, comme nous l'avons déjà annoncé. Rappelons d'abord que le problème dual de (3.33) s'écrit

$$\sup_{q \geq 0} \mathcal{G}(q) \ , \tag{3.43}$$

où, par définition

$$\mathcal{G}(q) = \inf_{v \in V} \mathcal{L}(v, q) \ , \tag{3.44}$$

et que le multiplicateur de Lagrange p est une solution du problème dual (3.43). En fait, sous des hypothèses assez générales, on peut montrer que $\mathcal{G}$ est différentiable et que le gradient $\mathcal{G}'(q)$ est précisément égal à $F(u_q)$, où u_q est l'unique solution du problème de minimisation (3.44). En effet, on a

$$\mathcal{G}(q) = J(u_q) + q \cdot F(u_q),$$

et en dérivant formellement par rapport à q

$$\mathcal{G}'(q) = F(u_q) + \langle J'(u_q) + q \cdot F'(u_q), u_q' \rangle = F(u_q),$$

à cause de la condition d'optimalité pour u_q. On voit alors que **l'algorithme d'Uzawa n'est autre que la méthode du gradient à pas fixe avec projection appliquée au problème dual** puisque la deuxième équation de (3.37) peut s'écrire $p^{n+1} = P_{\mathbb{R}_+^M}\big(p^n + \mu\mathcal{G}'(p^n)\big)$ (le changement de signe par rapport à (3.30) vient du fait que le problème dual (3.43) est un problème de maximisation et non de minimisation). Le lecteur vérifiera très facilement cette assertion dans le cas particulier étudié à l'exercice suivant.

Exercice 3.3.2 Appliquer l'algorithme d'Uzawa au problème de la Remarque 2.6.12 (fonctionnelle quadratique et contraintes affines en dimension finie). Si la matrice B est de rang M, ce qui assure l'unicité de p d'après la Remarque 2.6.12, montrer que la suite p^n converge vers p.

Remarque 3.3.3 Une variante, plus simple, de l'algorithme d'Uzawa est **l'algorithme d'Arrow-Hurwicz** qui s'interprète lui aussi comme un algorithme de point selle. Simplement, au lieu de minimiser exactement en v à chaque itération de (3.37), l'algorithme d'Arrow-Hurwicz effectue un seul pas d'une méthode de gradient à pas fixe $\nu > 0$. Concrètement, à partir d'éléments quelconques $p^0 \in (\mathbb{R}_+)^M$ et $u^0 \in V$, on construit les suites (u^n) et (p^n) déterminées par les itérations

$$\begin{aligned} u^{n+1} &= u^n - \nu \left(J'(u^n) + p^n \cdot F'(u^n) \right) , \\ p^{n+1} &= P_{\mathbb{R}_+^M}(p^n + \mu F(u^{n+1})) , \end{aligned} \tag{3.45}$$

$\mu > 0, \nu > 0$ étant deux paramètres positif fixés. Autrement dit, (3.45) recherche un point selle en alternant un pas de minimisation en v et un pas de maximisation en q. •

3.3.3 Pénalisation des contraintes

Nous décrivons brièvement un autre moyen d'approcher un problème de minimisation avec contraintes par une suite de problèmes de minimisation sans contraintes ; c'est la procédure de **pénalisation** des contraintes. Nous évitons de parler ici de "méthode" ou "d'algorithme" car la pénalisation des contraintes n'est pas une méthode à proprement parler. La résolution effective des problèmes sans contraintes que nous allons construire doit être réalisée à l'aide de l'un des algorithmes de la Sous-section 3.2. Cette résolution peut d'ailleurs soulever des difficultés, car le problème "pénalisé" (3.47) est souvent "mal conditionné".

Nous nous plaçons pour simplifier dans le cas où $V = \mathbb{R}^N$, et nous considérons de nouveau le problème de minimisation convexe

$$\inf_{F(v) \leq 0} J(v) , \tag{3.46}$$

où J est une fonction convexe continue de $\mathbb{R}^N$ dans $\mathbb{R}$ et F une fonction convexe continue de $\mathbb{R}^N$ dans $\mathbb{R}^M$.

Pour $\varepsilon > 0$, nous introduisons alors le problème sans contraintes

$$\inf_{v \in \mathbb{R}^N} \left(J(v) + \frac{1}{\varepsilon} \sum_{i=1}^{M} \left[\max \left(F_i(v), 0 \right) \right]^2 \right) , \tag{3.47}$$

dans lequel ont dit que les contraintes $F_i(v) \leq 0$ sont "pénalisées". On peut alors énoncer le résultat suivant, qui montre que, pour ε petit, le problème (3.47) "approche bien" le problème (3.46).

Proposition 3.3.4 *On suppose que J est continue, strictement convexe, et infinie à l'infini, que les fonctions F_i sont convexes et continues pour $1 \leq i \leq M$, et que l'ensemble*

$$K = \left\{ v \in \mathbb{R}^N \quad , \quad F_i(v) \leq 0 \quad \forall\, i \in \{1, \ldots, M\} \right\}$$

est non vide. En notant u l'unique solution de (3.46) et, pour $\varepsilon > 0$, u_ε l'unique solution de (3.47), on a alors

$$\lim_{\varepsilon \to 0} u_\varepsilon = u \ .$$

Démonstration. L'ensemble K étant convexe fermé, l'existence et l'unicité de u découlent du Théorème 2.2.1 et de la stricte convexité de J. De plus, la fonction $G(v) = \sum_{i=1}^{M} [\max(F_i(v), 0)]^2$ est continue et convexe puisque la fonction de $\mathbb{R}$ dans $\mathbb{R}$ qui à x associé $\max(x, 0)^2$ est convexe et croissante. On en déduit que la fonctionnelle $J_\varepsilon(v) = J(v) + \varepsilon^{-1} G(v)$ est strictement convexe, continue, et infinie à l'infini puisque $G(v) \geq 0$, ce qui implique l'existence et l'unicité de u_ε. Comme $G(u) = 0$, on peut écrire

$$J_\varepsilon(u_\varepsilon) = J(u_\varepsilon) + \frac{G(u_\varepsilon)}{\varepsilon} \leq J_\varepsilon(u) = J(u) \ . \tag{3.48}$$

Ceci montre que

$$J(u_\varepsilon) \leq J_\varepsilon(u_\varepsilon) \leq J(u) \ , \tag{3.49}$$

et donc que u_ε est borné d'après la condition "infinie à l'infini". On peut donc extraire de la famille (u_ε) une suite (u_{ε_k}) qui converge vers une limite u_* lorsque ε_k tend vers 0. On a alors $0 \leq G(u_{\varepsilon_k}) \leq \varepsilon_k (J(u) - J(u_{\varepsilon_k}))$ d'après (3.48). Passant à la limite, on obtient $G(u_*) = 0$, qui montre que $u_* \in K$. Comme (3.49) implique que $J(u_*) \leq J(u)$, on a alors $u_* = u$, ce qui conclut la démonstration, toutes les suites extraites (u_{ε_k}) convergeant vers la même limite u. □

Exercice 3.3.3 En plus des hypothèses de la Proposition 3.3.4, on suppose que les fonctions J et $F_1, \ldots, F_M$ sont continûment différentiables. On note de nouveau $I(u)$ l'ensemble des contraintes actives en u, et on suppose que les contraintes sont qualifiées en u au sens de la Définition 2.5.14. Enfin, on suppose que les vecteurs $\left(F_i'(u)\right)_{i \in I(u)}$ sont linéairement indépendants, ce qui assure l'unicité des multiplicateurs de Lagrange $\lambda_1, \ldots, \lambda_M$ tels que $J'(u) + \sum_{i=1}^{M} \lambda_i F_i'(u) = 0$, avec $\lambda_i = 0$ si $i \notin I(u)$. Montrer alors que, pour tout indice $i \in \{1, \ldots, M\}$

$$\lim_{\varepsilon \to 0} \left[\frac{2}{\varepsilon} \max(F_i(u_\varepsilon), 0) \right] = \lambda_i \ .$$

Remarque 3.3.5 Nous verrons à la Sous-section 4.2.3 une autre méthode de pénalisation par introduction de fonctions "barrières". •

3.3.4 Algorithme du Lagrangien augmenté

Nous concluons cette sous-section en présentant l'algorithme du Lagrangien augmenté, qui est une méthode combinant les avantages de l'algorithme d'Uzawa et de la méthode de pénalisation. En particulier, il est nettement plus robuste que la méthode de pénalisation dont il évite le caractère "mal conditionné".

L'idée de la méthode est d'introduire un Lagrangien **augmenté** d'un terme de pénalisation des contraintes. Pour fixer les idées, plaçons nous en dimension finie,

$V = \mathbb{R}^N$, et considérons le problème de minimisation avec contraintes d'égalité

$$\inf_{F(v)=0} J(v) \ , \tag{3.50}$$

où J est une fonction de $\mathbb{R}^N$ dans $\mathbb{R}$ et F une fonction de $\mathbb{R}^N$ dans $\mathbb{R}^M$, toutes les deux supposées de classe C^1.

Pour $v \in \mathbb{R}^N$, $\lambda \in \mathbb{R}^M$ et $\mu > 0$, nous introduisons le **Lagrangien augmenté**

$$\mathcal{L}_{aug}(v, \lambda, \mu) = J(v) + \lambda \cdot F(v) + \frac{\mu}{2}\|F(v)\|^2, \tag{3.51}$$

où $\|F(v)\|$ est la norme Euclidienne dans $\mathbb{R}^M$ du vecteur $F(v)$. Autrement dit, on additionne une pénalisation quadratique de la contrainte au Lagrangien habituel (3.34). Dans (3.51), λ est bien sûr un multiplicateur de Lagrange pour la contrainte, tandis que μ est un coefficient de pénalisation. Dans la sous-section précédente ce coefficient était noté $1/\varepsilon$. Ce changement de notation n'est pas anodin car, ici, il ne sera pas nécessaire en pratique de faire tendre μ vers l'infini pour que la contrainte soit satisfaite.

Le principe de l'algorithme est de trouver un point selle de (3.51). A λ et μ fixés, on minimise en v le Lagrangien augmenté par une des méthodes proposées pour la minimisation sans contrainte. A v et μ fixés, on maximise en λ par un algorithme itératif dont on va donner maintenant le principe. A l'itération n on note λ^n la valeur du multiplicateur de Lagrange et v^n le point de minimum de $v \to \mathcal{L}_{aug}(v, \lambda^n, \mu)$. Ecrivons la condition d'optimalité pour v^n

$$J'(v^n) + \lambda^n \cdot F'(v^n) + \mu F(v^n) \cdot F'(v^n) = 0. \tag{3.52}$$

Si le même vecteur v^n était solution de (3.50), on aurait la condition d'optimalité du Théorème 2.5.6

$$J'(v^n) + \lambda^* \cdot F'(v^n) = 0, \tag{3.53}$$

où λ^* est un multiplicateur de Lagrange optimal. En comparant (3.52) et (3.53) on obtient (en supposant que les contraintes sont qualifiées en v^n)

$$\lambda^* = \lambda^n + \mu F(v^n). \tag{3.54}$$

De (3.54) on tire deux informations. D'une part, si la suite λ^n tends vers λ^*, il n'est pas nécessaire de faire tendre le coefficient de pénalisation μ vers l'infini pour avoir la satisfaction progressive de la contrainte $F(v) = 0$. D'autre part, (3.54) suggère une formule de mise à jour du multiplicateur de Lagrange

$$\lambda^{n+1} = \lambda^n + \mu F(v^n).$$

Cette formule est très semblable à celle de l'algorithme d'Uzawa, voir (3.45), sauf qu'ici le pas μ n'est pas petit. Au total **l'algorithme du Lagrangien augmenté** est le suivant : on choisit $\mu > 0$, on initialise $\lambda_0 \in \mathbb{R}^M$ et on construit deux suites v^n et λ^n par

$$\mathcal{L}_{aug}(v^n, \lambda^n, \mu) = \min_{v \in \mathbb{R}^N} \mathcal{L}_{aug}(v, \lambda^n, \mu),$$

$$\lambda^{n+1} = \lambda^n + \mu F(v^n).$$

Par ailleurs, de temps en temps, et un nombre limité de fois (voir le chapitre 17 de [22]), on augmente la valeur du coefficient de pénalisation μ. Mais, comme le rappelle le résultat ci-dessous, il n'est pas nécessaire de faire tendre μ vers l'infini pour converger.

Lemme 3.3.6 *On suppose que les fonctions J et F sont de classe C^2. Soit v^* un point de minimum local de* (3.50) *où les contraintes sont qualifiées au sens que la matrice $F'(v^*)$ est de rang M. Soit λ^* un multiplicateur de Lagrange pour lequel la condition d'optimalité de* (3.50) *est vérifiée. On suppose que la condition suffisante d'optimalité d'ordre 2 est satisfaite en v^*, à savoir*

$$\left(J''(v^*) + \lambda^* \cdot F''(v^*)\right)(w,w) > 0 \quad \forall\, w \in K(v^*) = \mathrm{Ker}F'(v^*), w \neq 0. \tag{3.55}$$

Alors, il existe $\mu_0 > 0$ tel que, pour tout $\mu \geq \mu_0$, v^* est un point de minimum local de $v \to \mathcal{L}_{aug}(v, \lambda^*, \mu)$ (sans contrainte).

Remarque 3.3.7 L'intérêt du Lemme 3.3.6 est de montrer que, si l'on connait la valeur du multiplicateur de Lagrange optimal, alors la minimisation, sans contrainte et avec une pénalisation **finie** de la contrainte, du Lagrangien augmenté conduit à une solution du problème d'origine (3.50). Cela justifie, en quelque sorte, le fait qu'il n'est pas nécessaire de faire tendre vers l'infini le coefficient de pénalisation μ pour que l'algorithme du Lagrangien augmenté converge, du moins si on a une bonne estimation du multiplicateur de Lagrange. •

Démonstration. Si v^* un point de minimum local de (3.50) et que les contraintes sont qualifiées, alors la condition d'optimalité du premier ordre du Théorème 2.5.6 donne

$$J'(v^*) + \lambda^* \cdot F'(v^*) = 0 \quad \text{et} \quad F(v^*) = 0,$$

ce qui implique que $\mathcal{L}'_{aug}(v^*, \lambda^*, \mu) = 0$. Calculons la dérivée seconde du Lagrangien augmenté

$$\mathcal{L}''_{aug}(v^*, \lambda^*, \mu)(w,w) = \mathcal{L}''(v^*, \lambda^*)(w,w) + \mu \|F'(v^*)w\|^2, \tag{3.56}$$

avec

$$\mathcal{L}''(v^*, \lambda^*)(w,w) = J''(v^*)(w,w) + \lambda^* \cdot F''(v^*)(w,w)$$

et où il n'y a pas de dérivée seconde dans le terme de pénalisation car $F(v^*) = 0$. L'hypothèse (3.55) montre que le premier terme à droite dans (3.56) est une forme quadratique définie positive sur $\mathrm{Ker}F'(v^*)$, tandis que le second terme est une forme quadratique définie positive sur le sous-espace de dimension M engendré par les colonnes de $F'(v^*)$ (qui est aussi l'orthogonal de $\mathrm{Ker}F'(v^*)$). Notons que sur ce sous-espace de dimension fini $\mathcal{L}''_{aug}(v*, \lambda^*, \mu)(w,w)$ peut prendre des valeurs négatives mais qu'on peut les minorer par $-\mu_0\|w\|^2$ où $-\mu_0$ est la plus petite valeur propre de la matrice $M \times M$ qui représente cette forme quadratique sur ce sous-espace. Il suffit alors de choisir $\mu > \mu_0$ pour que la somme des deux termes soit strictement positif pour $w \neq 0$. Au total, la dérivée seconde du Lagrangien augmenté est définie positive en v^* qui est un point critique. C'est donc un point de minimum local. □

3.4 Méthode de Newton

3.4.1 Cas de la dimension finie

Pour simplifier la présentation, on se place en dimension finie $V = \mathbb{R}^N$. On veut résoudre le problème de minimisation d'une fonction régulière $J(v)$ de $\mathbb{R}^N$ dans $\mathbb{R}$, sans contraintes. On sait que les éventuels points de minimum de $J(v)$ se trouvent parmi les zéros de la dérivée $J'(v)$. Le principe de la méthode de Newton est de chercher les zéros de la dérivée $J'(v)$. Remarquons tout de suite un inconvénient de ce principe : ces zéros peuvent aussi correspondre à des points de maximum ou des points selle et la méthode de Newton ne permet pas de faire le tri entre minima, maxima ou simples points stationnaires. Evidemment, si la fonction J est convexe, on sait qu'elle n'a que des points de minimum.

Rentrons dans les détails et notons $F = J'$. On suppose que F une fonction de classe C^2 de $\mathbb{R}^N$ dans $\mathbb{R}^N$. Soit u un zéro régulier de F c'est-à-dire que

$$F(u) = 0 \quad \text{et} \quad F'(u) \text{ matrice inversible.}$$

Une formule de Taylor au voisinage de v nous donne

$$F(u) = F(v) + F'(v)(u - v) + \mathcal{O}\left(\|u - v\|^2\right),$$

c'est-à-dire

$$u = v - \left(F'(v)\right)^{-1} F(v) + \mathcal{O}\left(\|v - u\|^2\right).$$

La méthode de Newton consiste à résoudre de façon itérative cette équation en négligeant le reste. Pour un choix initial $u^0 \in \mathbb{R}^N$, on calcule

$$u^{n+1} = u^n - \left(F'(u^n)\right)^{-1} F(u^n) \quad \text{pour} \quad n \geq 0. \tag{3.57}$$

Rappelons que l'on ne calcule pas l'inverse de la matrice $F'(u^n)$ dans (3.57) mais que l'on résout un système linéaire. Du point de vue de l'optimisation, la méthode de Newton s'interprète de la manière suivante. Soit J une fonction de classe C^3 de $\mathbb{R}^N$ dans $\mathbb{R}$, et soit u un minimum local de J. En notant $F = J'$, la méthode précédente permet de résoudre la condition nécessaire d'optimalité $J'(u) = 0$. Plus précisément, on peut aussi voir la méthode de Newton comme une méthode de minimisation. A cause du développement de Taylor

$$J(w) = J(v) + J'(v) \cdot (w - v) + \frac{1}{2} J''(v)(w - v) \cdot (w - v) + \mathcal{O}\left(\|w - v\|^3\right), \tag{3.58}$$

on peut approcher $J(w)$ au voisinage de v par une fonction quadratique. La méthode de Newton consiste alors à minimiser cette approximation quadratique et à itérer. Le minimum de la partie quadratique du terme de droite de (3.58) est donné par $w = v - \left(J''(v)\right)^{-1} J'(v)$ si la matrice $J''(v)$ est définie positive. On retrouve alors la formule itérative (3.57).

L'avantage principal de la méthode de Newton est sa convergence bien plus rapide que les méthodes précédentes.

Proposition 3.4.1 *Soit F une fonction de classe C^2 de $\mathbb{R}^N$ dans $\mathbb{R}^N$, et u un zéro régulier de F (i.e. $F(u) = 0$ et $F'(u)$ inversible). Il existe un réel $\epsilon > 0$ et une constante $0 < C < 1/\epsilon$ tels que, si u^0 est assez proche de u au sens où $\|u - u^0\| \leq \epsilon$, la méthode de Newton définie par (3.57) converge, c'est-à-dire que la suite (u^n) converge vers u, au sens où*

$$\|u^{n+1} - u\| \leq C\|u^n - u\|^2 \quad \text{et} \quad \|u^n - u\| \leq C^{-1}(C\epsilon)^{2^n}. \tag{3.59}$$

Remarque 3.4.2 La convergence de la méthode de Newton, dite **quadratique**, est extrêmement rapide (bien plus que la méthode du gradient qui converge simplement géométriquement d'après la Remarque 3.2.8). Le nombre de chiffres significatifs dans l'approximation u^n de la solution u double à chaque itération. Bien sûr, cette convergence rapide a un prix car, à chaque itération de la méthode de Newton (3.57), il faut résoudre un système linéaire, ce qui est coûteux en temps de calcul. De plus, la convergence rapide donnée par (3.59) n'a lieu que si F est de classe C^2, et si u^0 est assez proche de u, hypothèses bien plus restrictives que celles que nous avons utilisées jusqu'à présent. Effectivement, même dans des cas très simples dans $\mathbb{R}$, la méthode de Newton peut diverger pour certaines données initiales u^0 ; il faut noter aussi que la convergence quadratique (3.59) ne se produit qu'au voisinage d'un zéro régulier, comme le montre l'application de la méthode de Newton à la fonction $F(x) = \|x\|^2$ dans $\mathbb{R}^N$, pour laquelle la convergence n'est que géométrique. Par ailleurs, si on applique la méthode de Newton pour la minimisation d'une fonction J comme expliqué ci-dessus, il se peut que la méthode converge vers un maximum ou un col de J, et non pas vers un minimum, car elle ne fait que rechercher les zéros de J'. La méthode de Newton n'est donc pas supérieure en tout point aux algorithmes précédents, mais la propriété de convergence locale quadratique (3.59) la rend cependant particulièrement intéressante. •

Démonstration. Par continuité de F' et de F'' il existe $\delta > 0$ tel que F' est inversible en tout point v de la boule de centre u et de rayon δ et, de plus, il existe $C_1, C_2 > 0$ tels que

$$\| \left(F'(v)\right)^{-1} \| \leq C_1, \quad \|F''(v)\| \leq C_2, \quad \text{pour tout } \|v - u\| \leq \delta.$$

Supposons que toutes les itérées jusqu'à u^n sont restées proches de u, au sens où $\|u - u^n\| \leq \delta$, donc $F'(u^n)$ est inversible. Comme $F(u) = 0$, on déduit de (3.57)

$$u^{n+1} - u = u^n - u - \left(F'(u^n)\right)^{-1} \left(F(u^n) - F(u)\right)$$

qui, par développement de Taylor autour de u^n, devient

$$u^{n+1} - u = \frac{1}{2} \left(F'(u^n)\right)^{-1} \left(\int_0^1 F''(u^n + s(u - u^n))\, ds \right) (u^n - u) \cdot (u^n - u).$$

On peut majorer pour obtenir

$$\|u^{n+1} - u\| \leq \frac{1}{2} C_1 C_2 \|u^n - u\|^2.$$

qui n'est rien d'autre que la première partie de (3.59) avec $C = C_1C_2/2$. On choisit $0 < \epsilon < \min(\delta, C^{-1})$ et $\|u - u^0\| \leq \epsilon$. On vérifie par récurrence que $\|u - u^n\| \leq \epsilon \leq \delta$ (donc toutes les itérées restent proches de u). En prenant le logarithme de l'inégalité précédente, on a

$$\log \|u^{n+1} - u\| \leq \log C + 2 \log \|u^n - u\|,$$

d'où l'on déduit

$$\log \|u^n - u\| \leq 2^n \log \|u^0 - u\| + (2^n - 1) \log C$$

c'est-à-dire

$$\|u^n - u\| \leq C^{-1} \Big(C\|u^0 - u\| \Big)^{2^n} \leq C^{-1} (C\epsilon)^{2^n}$$

qui converge vers zéro puisque $C\epsilon < 1$. □

Remarque 3.4.3 Un inconvénient majeur de la méthode de Newton est la nécessité de connaître le Hessien $J''(v)$ (ou la matrice dérivée $F'(v)$). Lorsque le problème est de grande taille ou bien si J n'est pas facilement deux fois dérivable, on peut modifier la méthode de Newton pour éviter de calculer cette matrice $J''(v) = F'(v)$. Les méthodes, dites de quasi-Newton, proposent de calculer de façon itérative aussi une approximation S^n de $(F'(u^n))^{-1}$. On remplace alors la formule (3.57) par

$$u^{n+1} = u^n - S^n F(u^n) \quad \text{pour} \quad n \geq 0.$$

En général on calcule S^n par une formule de récurrence du type

$$S^{n+1} = S^n + C^n$$

où C^n est une matrice de rang 1 qui dépend de $u^n, u^{n+1}, F(u^n), F(u^{n+1})$, choisie de manière à ce que $S^n - (F'(u^n))^{-1}$ converge vers 0. Pour plus de détails sur ces méthodes de quasi-Newton nous renvoyons à [6] et [12]. •

On peut adapter la méthode de Newton à la minimisation d'une fonction J avec des contraintes d'égalité. Soit J une fonction de classe C^3 de $\mathbb{R}^N$ dans $\mathbb{R}$, $G = (G_1, ..., G_M)$ une fonction de classe C^3 de $\mathbb{R}^N$ dans $\mathbb{R}^M$ (avec $M \leq N$), et soit u un minimum local de

$$\min_{v \in \mathbb{R}^N \,,\; G(v)=0} J(v) \,. \tag{3.60}$$

Si les vecteurs $(G_1'(u), ..., G_M'(u))$ sont linéairement indépendants, la condition nécessaire d'optimalité du Théorème 2.5.6 est

$$J'(u) + \sum_{i=1}^{M} \lambda_i G_i'(u) = 0 \,, \quad G_i(u) = 0 \quad 1 \leq i \leq M \,. \tag{3.61}$$

où les $\lambda_1, \ldots, \lambda_M \in \mathbb{R}$ sont les multiplicateurs de Lagrange. On peut alors résoudre le système (3.61) de $(N + M)$ équations à $(N + M)$ inconnues $(u, \lambda) \in \mathbb{R}^{N+M}$ par une méthode de Newton. On pose donc

$$F(u, \lambda) = \begin{pmatrix} J'(u) + \lambda \cdot G'(u) \\ G(u) \end{pmatrix},$$

dont la matrice dérivée est

$$F'(u,\lambda) = \begin{pmatrix} J''(u) + \lambda \cdot G''(u) & (G'(u))^* \\ G'(u) & 0 \end{pmatrix}.$$

On peut alors appliquer l'algorithme de Newton (3.57) à cette fonction $F(u,\lambda)$ si la matrice $F'(u,\lambda)$ est inversible. Nous allons voir que cette condition est "naturelle" au sens où elle correspond à une version un peu plus forte de la condition d'optimalité d'ordre 2 de la Proposition 2.5.12. La matrice $F'(u,\lambda)$ est inversible si elle est injective. Soit (w,μ) un élément de son noyau

$$\begin{cases} J''(u)w + \lambda \cdot G''(u)w + (G'(u))^*\mu = 0 \\ G_i'(u) \cdot w = 0 \text{ pour } 1 \leq i \leq M \end{cases}$$

On en déduit que $w \in \mathrm{Ker}G'(u) = \bigcap_{i=1}^M \mathrm{Ker}G_i'(u)$ et $(J''(u) + \lambda \cdot G''(u))w \in \mathrm{Im}(G'(u))^*$. Or $\mathrm{Im}(G'(u))^* = [\mathrm{Ker}G'(u)]^\perp$. Par conséquent, si on suppose que

$$(J''(u) + \lambda \cdot G''(u))\,(w,w) > 0 \quad \forall\, w \in \mathrm{Ker}G'(u), w \neq 0\,, \tag{3.62}$$

la matrice $F'(u,\lambda)$ est inversible. On remarque que (3.62) est l'inégalité stricte dans la condition d'optimalité d'ordre 2 de la Proposition 2.5.12. Il est donc naturel de faire l'hypothèse (3.62) qui permet d'utiliser l'algorithme de Newton. On peut ainsi démontrer la convergence de cette méthode (voir [6]). Il est intéressant d'interpréter cet algorithme comme une méthode de minimisation. On introduit le Lagrangien $\mathcal{L}(v,\mu) = J(v) + \mu \cdot G(v)$, ses dérivées par rapport à v, $\mathcal{L}'$ et $\mathcal{L}''$, et on vérifie que l'équation

$$(u^{n+1},\lambda^{n+1}) = (u^n,\lambda^n) - \left(F'(u^n,\lambda^n)\right)^{-1} F(u^n,\lambda^n)$$

est la condition d'optimalité pour que u^{n+1} soit un point de minimum du problème quadratique à contraintes affines

$$\min_{\substack{w\in\mathbb{R}^N \\ G(u^n)+G'(u^n)\cdot(w-u^n)=0}} Q^n(w)\,, \tag{3.63}$$

avec

$$Q^n(w) = \left(\mathcal{L}(u^n,\lambda^n) + \mathcal{L}'(u^n,\lambda^n)\cdot(w-u^n) + \frac{1}{2}\mathcal{L}''(u^n,\lambda^n)(w-u^n)\cdot(w-u^n)\right)\,,$$

et λ^{n+1} est le multiplicateur de Lagrange associé au point de minimum de (3.63). On remarque que dans (3.63) on a effectué un développement de Taylor à l'ordre deux en w sur le Lagrangien $\mathcal{L}(w,\lambda^n)$ et on a linéarisé la contrainte $G(w)$ autour du point u^n.

Remarque 3.4.4 Dans (3.63) on a utilisé une approximation quadratique du Lagrangien et non pas de la fonction J. On pourrait essayer de se contenter d'une méthode itérative de résolution de l'approximation quadratique à contraintes affines suivante

$$\min_{\substack{w\in\mathbb{R}^N \\ G(v)+G'(v)\cdot(w-v)=0}} \left(J(v) + J'(v)\cdot(w-v) + \frac{1}{2}J''(v)(w-v)\cdot(w-v)\right)\,. \tag{3.64}$$

Malheureusement la méthode basée sur (3.64) peut ne pas converger ! En particulier, il n'est pas évident que le Hessien $J''(v)$ soit définie positif sur l'espace des contraintes (c'est le Hessien du Lagrangien qui est positif comme l'affirme la condition d'optimalité d'ordre 2 de la Proposition 2.5.12). •

3.4.2 Cas de la dimension infinie

Expliquons rapidement comment la méthode de Newton s'applique à la minimisation de fonctions définies sur un espace de Hilbert V de dimension infinie. Considérons une fonction convexe, de classe C^2, $J(v)$ de V dans $\mathbb{R}$. On repart du développement de Taylor à l'ordre 2 (3.58) et, pour une initialisation $u^0 \in V$, on calcule u^{n+1} comme un point de minimum de l'approximation quadratique

$$J^n(w) = J(u^n) + J'(u^n)(w - u^n) + \frac{1}{2}J''(u^n)\Big((w - u^n), (w - u^n)\Big), \tag{3.65}$$

où $V \times V \ni (w_1, w_2) \to J''(u^n)(w_1, w_2)$ est une forme bilinéaire symétrique. Si on suppose de plus qu'elle est coercive et continue, c'est-à-dire, s'il existe deux constantes $0 < \nu < M$ telles que

$$\nu\|w\|^2 \leq J''(u^n)(w, w) \leq M\|w\|^2,$$

alors la fonction $J^n(w)$, définie par (3.65), est fortement convexe, donc admet un unique point de minimum u^{n+1} caractérisé par la condition d'optimalité du premier ordre

$$(J^n)'(u^{n+1})(w) = 0 \quad \text{pour tout } w \in V,$$

que l'on peut réécrire : trouver $u^{n+1} \in V$, solution de

$$J''(u^n)\Big((u^{n+1} - u^n), w\Big) = -J'(u^n)(w) \quad \text{pour tout } w \in V. \tag{3.66}$$

L'équation (3.66) n'est rien d'autre qu'une formulation variationnelle sur V qu'on peut résoudre à l'aide du lemme de Lax-Milgram grâce à l'hypothèse de coercivité de $J''(u^n)$ (voir [1]). C'est typiquement de cette façon que sont résolues les équations aux dérivées partielles non-linéaires.

3.5 Méthodes d'approximations successives

Considérons un problème général d'optimisation sous contraintes d'égalité

$$\inf_{F(v)=0} J(v), \tag{3.67}$$

où $J(v)$ et $(F_1(v), ..., F_M(v)) = F(v)$ sont des fonctions régulières de $\mathbb{R}^N$ dans $\mathbb{R}$. Les remarques qui suivent s'appliquent de la même manière aux problèmes avec contraintes d'inégalité, moyennant des modifications évidentes.

Si l'application directe des algorithmes d'optimisation ci-dessus est trop compliquée ou coûteuse pour (3.67), une stratégie courante consiste à remplacer ce dernier

par un problème approché obtenu par développement de Taylor des fonctions J et F. L'idée sous-jacente est qu'il est plus facile de résoudre le problème approché que le problème exact. Ces approximations n'ayant qu'un caractère local, il faut itérer cette stratégie en faisant un nouveau développement de Taylor au point de minimum obtenu sur le précédent problème approché.

La première méthode, dite de **programmation linéaire successive** (ou séquentielle), consiste à remplacer les fonctions J et F par des approximations affines (cette méthode est connue aussi sous l'acronyme SLP pour l'anglais "sequential linear programming"). Etant donné une initialisation $v^0 \in V$ (ne vérifiant pas nécessairement la contrainte $F(v) = 0$), on calcule une suite de solutions approchées v^n, $n \geq 1$, définies comme les solutions de

$$\inf_{F(v^{n-1})+F'(v^{n-1})\cdot(v-v^{n-1})=0} \Big\{ J(v^{n-1}) + J'(v^{n-1}) \cdot (v - v^{n-1}) \Big\}, \tag{3.68}$$

qui n'est rien d'autre qu'un programme linéaire, comme étudié dans la Section 4.2 et pour lequel on dispose d'algorithmes extrêmement efficaces. Une difficulté immédiate dans la résolution de (3.68) est que sa valeur minimum puisse être $-\infty$ et qu'il n'y ait pas de solution optimale. (Notons que, sous une condition de qualification standard, $(F_1'(v^{n-1}), ..., F_M'(v^{n-1}))$ famille libre de $\mathbb{R}^N$, l'ensemble admissible de (3.68) n'est pas vide.) C'est pourquoi, en pratique, cette méthode s'accompagne d'une contrainte supplémentaire, dite de **région de confiance**, qui prend la forme

$$\|v - v^{n-1}\| \leq \delta, \tag{3.69}$$

où $\delta > 0$ est un paramètre qui définit la taille du voisinage de v^{n-1} dans lequel (3.68) est une bonne approximation de (3.67). La norme dans (3.69) peut-être soit la norme $\|v\|_\infty = \max_{1\leq i\leq N} |v_i|$, soit la norme $\|v\|_1 = \sum_{i=1}^N |v_i|$, ce qui dans les deux cas préserve le fait que le problème approché est un programme linéaire. Ce dernier a alors nécessairement au moins une solution optimale puisque l'ensemble admissible est désormais borné.

Une deuxième méthode, dite de **programmation quadratique séquentielle**, consiste à remplacer la fonction J par une approximation quadratique et F par une approximation affine (cette méthode est connue aussi sous l'acronyme SQP pour l'anglais "sequential quadratic programming"). Etant donné une initialisation $v^0 \in V$ (ne vérifiant pas nécessairement la contrainte $F(v) = 0$), on calcule une suite de solutions approchées v^n, $n \geq 1$, définies comme les solutions de

$$\inf_{F(v^{n-1})+F'(v^{n-1})\cdot(v-v^{n-1})=0} \Big\{ J(v^{n-1}) + J'(v^{n-1})\cdot(v-v^{n-1}) + \frac{1}{2} Q^{n-1}(v-v^{n-1})\cdot(v-v^{n-1}) \Big\}, \tag{3.70}$$

où Q^{n-1} est une matrice symétrique de taille N. Si Q^{n-1} est définie positive, alors on sait résoudre explicitement le problème (3.70) (voir l'Exercice 2.5.10). Le point crucial dans cette méthode SQP est que Q^{n-1} **n'est pas** la Hessienne de la fonction objectif $J''(v^{n-1})$ mais est la Hessienne du Lagrangien

$$Q^{n-1} = J''(v^{n-1}) + \lambda^{n-1} \cdot F''(v^{n-1}),$$

où λ^{n-1} est le multiplicateur de Lagrange dans la condition d'optimalité pour v^{n-1} (solution à l'itération précédente). En effet, ce qui importe n'est pas l'approximation de J par son développement de Taylor à l'ordre 2 dans tout $\mathbb{R}^N$ mais seulement sur la variété définit par la contrainte $F(v) = 0$. Concrêtement, on écrit

$$J(v) \approx J(v^{n-1}) + J'(v^{n-1}) \cdot (v - v^{n-1}) + \frac{1}{2} J''(v^{n-1})(v - v^{n-1}) \cdot (v - v^{n-1}) \quad (3.71)$$

et

$$0 = F(v) \approx F(v^{n-1}) + F'(v^{n-1}) \cdot (v - v^{n-1}) + \frac{1}{2} F''(v^{n-1})(v - v^{n-1}) \cdot (v - v^{n-1}), \quad (3.72)$$

puis on multiplie (3.72) par le multiplicateur de Lagrange λ^{n-1} et on somme le résultat à (3.71), ce qui donne exactement la fonction objectif de (3.70) (à une constante près, en utilisant la contrainte linéaire). La Proposition 2.5.12 (condition d'optimalité du 2ème ordre) montre que la matrice Q^{n-1} est positive si v^{n-1} est le point de minimum de (3.67), ce qui entraine que (3.70) admet au moins une solution optimale. Par contre, la matrice $J''(v^{n-1})$ n'a aucune raison d'être positive en général. Néanmoins, si v^{n-1} n'est pas un point de minimum, il peut être nécessaire de recourir à nouveau à une contrainte de région de confiance du type de (3.69). Pour plus de détails nous renvoyons à [22].

Chapitre 4

PROGRAMMATION LINÉAIRE

4.1 Introduction

Ce chapitre est consacré à la **programmation linéaire** qui permet de résoudre efficacement les problèmes d'optimisation continue où les contraintes et le critère s'expriment linéairement en fonction des variables (voir l'Exemple 1.2.1). Ce type de problème est extrêmement fréquent dans ce qu'il est convenu d'appeler la RO ou **recherche opérationnelle**. Dans RO, le mot "opérationnel" s'est d'abord entendu au sens propre : la RO est née, en grande partie, des problèmes de planification qui se sont posés pendant la seconde guerre mondiale et peu après. Ainsi G. Dantzig, l'inventeur de l'algorithme du simplexe, était conseiller pour l'armée de l'air américaine, et la planification du pont aérien sur Berlin en 1948 est une application célèbre de la programmation linéaire (voir [23] pour plus de détails).

La RO ne se limite pas du tout à la programmation linéaire et, bien au contraire, emprunte des outils à plusieurs domaines scientifiques : optimisation (continue et combinatoire), probabilités (algorithmes stochastiques), théorie des jeux, mathématiques discrètes, théorie des graphes, l'informatique, d'une part via la théorie de la complexité, et d'autre part via la programmation par contraintes. Une part importante de l'activité en RO relève par ailleurs de l'art du praticien (modélisation, conception d'heuristiques, etc.). Nous nous contentons ici de donner un éclairage très partiel sur les apports de l'optimisation, à travers la programmation linéaire, dans ce vaste domaine. En particulier, la Section 4.3 donnera des méthodes pour résoudre des programmes linéaires où l'on se restreint à des solutions entières et pas réelles. Bien sûr, la Section 4.2 est une présentation de la programmation linéaire classique. Rappelons que celle-ci est aussi une brique de base dans de nombreux algorithmes d'optimisation qui linéarisent les problèmes à résoudre.

Pour une introduction à la partie mathématisée de la RO nous renvoyons le lecteur vers l'ouvrage [5] issu d'un cours de troisième année à l'Ecole Polytechnique.

4.2 Programmation linéaire

4.2.1 Définitions et propriétés

On veut résoudre le problème suivant, dit **programme linéaire sous forme standard**,

$$\inf_{x\in\mathbb{R}^n \text{tel que } Ax=b,\ x\geq 0} c\cdot x, \tag{4.1}$$

où A est une matrice de taille $m\times n$, $b\in\mathbb{R}^m$, $c\in\mathbb{R}^n$, et la contrainte $x\geq 0$ signifie que toutes les composantes de x sont positives ou nulles. Dans tout ce qui suit on supposera que $m\leq n$ et que le rang de A est exactement m. En effet, si $\mathrm{rg}(A)<m$, certaines lignes de A sont liées et deux possibilités se présentent : soit les contraintes (correspondantes à ces lignes) sont incompatibles, soit elles sont redondantes et on peut donc éliminer les lignes inutiles.

Le problème (4.1) semble être un cas particulier de programme linéaire puisque les contraintes d'inégalités sont seulement du type $x\geq 0$. Il n'en est rien, et tout programme linéaire du type

$$\inf_{x\in\mathbb{R}^n \text{tel que } Ax\geq b,\ A'x=b'} c\cdot x.$$

peut se mettre sous la forme standard (4.1) quitte à changer la taille des données. En effet, remarquons tout d'abord que les contraintes d'égalité $A'x=b'$ sont évidemment équivalentes aux contraintes d'inégalité $A'x\leq b'$ et $A'x\geq b'$. On peut donc se restreindre au cas suivant (qui ne contient que des contraintes d'inégalité)

$$\inf_{x\in\mathbb{R}^n \text{tel que } Ax\geq b} c\cdot x. \tag{4.2}$$

Dans (4.2) on peut remplacer la contrainte d'inégalité en introduisant de nouvelles variables, dites **d'écarts**, $\lambda\in\mathbb{R}^m$. La contrainte d'inégalité $Ax\geq b$ est alors équivalente à $Ax=b+\lambda$ avec $\lambda\geq 0$. Ainsi (4.2) est équivalent à

$$\inf_{(x,\lambda)\in\mathbb{R}^{(n+m)} \text{tel que } Ax=b+\lambda,\ \lambda\geq 0} c\cdot x. \tag{4.3}$$

Finalement, si on décompose chaque composante de x en partie positive et négative, c'est-à-dire si on pose $x=x^+-x^-$ avec $x^+=\max(0,x)$ et $x^-=-\min(0,x)$, on obtient que (4.2) est équivalent à

$$\inf_{(x^+,x^-,\lambda)\in\mathbb{R}^{(2n+m)} \text{tel que } Ax^+-Ax^-=b+\lambda,\ x^+\geq 0, x^-\geq 0, \lambda\geq 0} c\cdot(x^+-x^-). \tag{4.4}$$

qui est bien sous forme standard (mais avec plus de variables). Il n'y a donc aucune perte de généralité à étudier le programme linéaire standard (4.1).

Nous avons déjà donné une motivation concrète de la programmation linéaire au début du Chapitre 1 (voir l'Exemple 1.2.1). Considérons pour l'instant un exemple simple qui va nous permettre de comprendre quelques aspects essentiels d'un programme linéaire

$$\min_{\substack{x_1\geq 0, x_2\geq 0, x_3\geq 0\\ 2x_1+x_2+3x_3=6}} x_1+4x_2+2x_3\ . \tag{4.5}$$

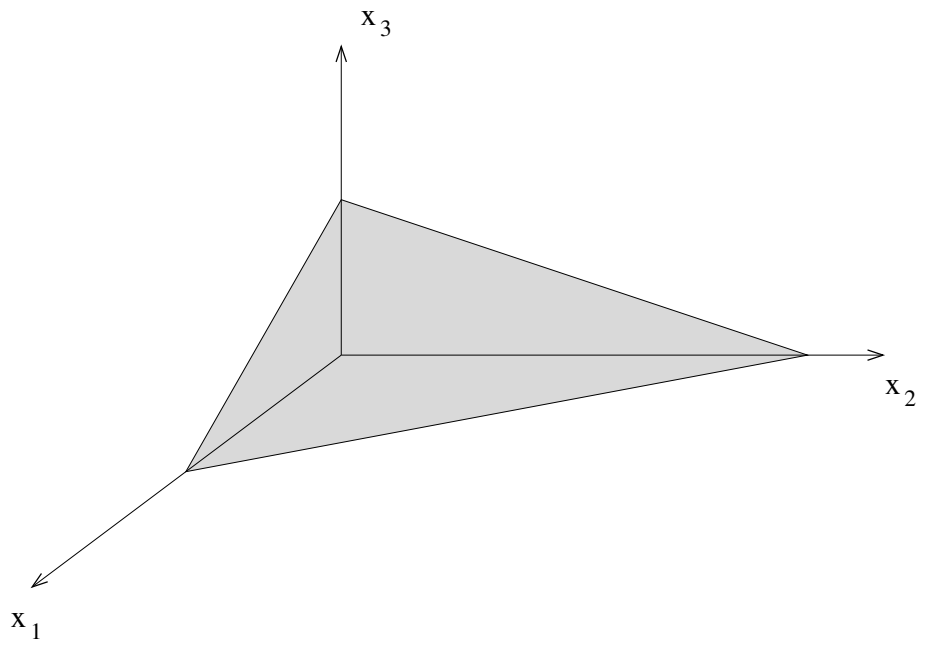

FIGURE 4.1 – Ensemble admissible pour l'exemple (4.5).

Sur la Figure 4.1 nous avons tracé l'ensemble des (x_1, x_2, x_3) qui vérifient les contraintes : c'est un triangle plan T. C'est un fermé compact de $\mathbb{R}^3$, donc la fonction continue $x_1+4x_2+2x_3$ y atteint son minimum que l'on note M. Pour déterminer ce minimum on peut considérer la famille de plans parallèles $x_1 + 4x_2 + 2x_3 = c$ paramétrée par c. En augmentant la valeur de c à partir de $-\infty$, on "balaie" l'espace $\mathbb{R}^3$ jusqu'à atteindre le triangle T, et le minimum M est obtenu lorsque le plan "touche" ce triangle. Autrement dit, tout point de minimum de (4.5) est sur le bord du triangle T. Une autre façon de le voir est de dire que la fonction $x_1 + 4x_2 + 2x_3$ a un gradient non nul dans T donc ses extréma se trouvent sur le bord de T. Pour l'exemple (4.5) le point de minimum (unique) est le sommet $(0, 3, 0)$ de T. Nous verrons qu'il s'agit d'un fait général : un point de minimum (s'il existe) peut toujours se trouver en un des sommets de l'ensemble géométrique des vecteurs x qui vérifient les contraintes. Il "suffit" alors d'énumérer tous les sommets afin de trouver le minimum : c'est précisément ce que fait (de manière intelligente) l'algorithme du simplexe que nous verrons dans la prochaine sous-section.

Pour établir cette propriété en toute généralité pour le programme linéaire standard (4.1), nous avons besoin de quelques définitions qui permettent de préciser le vocabulaire.

Définition 4.2.1 *L'ensemble X_{ad} des vecteurs de $\mathbb{R}^n$ qui satisfont les contraintes de (4.1), c'est-à-dire*

$$X_{ad} = \{x \in \mathbb{R}^n \text{ tel que } Ax = b,\ x \geq 0\},$$

est appelé ensemble des **solutions admissibles**. *On appelle sommet ou point extrémal de X_{ad} tout point $\overline{x} \in X_{ad}$ qui ne peut pas se décomposer en une combinai-*

son convexe (non triviale) de deux autres points de X_{ad}, c'est-à-dire que, s'il existe $y, z \in X_{ad}$ et $\theta \in]0,1[$ tels que $\overline{x} = \theta y + (1-\theta)z$, alors $y = z = \overline{x}$.

Remarque 4.2.2 Le vocabulaire de l'optimisation est trompeur pour les néophytes. On appelle solution (admissible) un vecteur qui satisfait les contraintes. Par contre, un vecteur qui atteint le minimum de (4.1) est appelé **solution optimale** (ou point de minimum). •

On vérifie facilement que l'ensemble X_{ad} est un **polyèdre** (éventuellement vide). (Rappelons qu'un polyèdre est une intersection finie de demi-espaces de $\mathbb{R}^n$.) Ses points extrémaux sont donc les sommets de ce polyèdre. Lorsque X_{ad} est vide, par convention on note que

$$\inf_{x \in \mathbb{R}^n \text{tel que } Ax=b,\ x \geq 0} c \cdot x = +\infty.$$

Lemme 4.2.3 *Il existe au moins une solution optimale (ou point de minimum) du programme linéaire standard (4.1) si et seulement si la valeur du minimum est finie*

$$-\infty < \inf_{x \in \mathbb{R}^n \text{tel que } Ax=b,\ x \geq 0} c \cdot x < +\infty.$$

Démonstration. Soit $(x^k)_{k \geq 1}$ une suite minimisante de (4.1). On introduit la matrice $\mathcal{A}$ définie par

$$\mathcal{A} = \begin{pmatrix} c^* \\ A \end{pmatrix}.$$

La suite $\mathcal{A}x^k$ appartient au cône suivant

$$C = \left\{ \sum_{i=1}^{n} x_i \mathcal{A}_i \text{ avec } x_i \geq 0 \right\},$$

où les $\mathcal{A}_i$ sont les colonnes de la matrice $\mathcal{A}$. D'après le Lemme de Farkas 2.5.18 le cône C est fermé, ce qui implique que

$$\lim_{k \to +\infty} \mathcal{A}x^k = \begin{pmatrix} z_0 \\ b \end{pmatrix} \in C,$$

donc il existe $\overline{x} \geq 0$ tel que

$$\begin{pmatrix} z_0 \\ b \end{pmatrix} = \begin{pmatrix} c \cdot \overline{x} \\ A\overline{x} \end{pmatrix},$$

et le minimum est atteint en $\overline{x}$. □

Définition 4.2.4 *On appelle* **base** *associée à (4.1) une base de $\mathbb{R}^m$ formée de m colonnes de A. On note B cette base qui est une sous-matrice de A, carrée d'ordre m inversible. Après permutation de ses colonnes on peut écrire A sous la forme (B, N)*

où N est une matrice de taille $m \times (n-m)$. De la même façon on peut décomposer x en (x_B, x_N) de sorte qu'on a

$$Ax = Bx_B + Nx_N.$$

Les composantes du vecteur x_B sont appelées variables de base et celles de x_N variables hors base. Une **solution basique** *(ou de base) est un vecteur $x \in X_{ad}$ tel que $x_N = 0$. Si en plus l'une des composantes de x_B est nulle, on dit que la solution basique est* **dégénérée**.

La notion de solution basique correspond à celle de sommet de X_{ad}.

Lemme 4.2.5 *Les sommets du polyèdre X_{ad} sont exactement les solutions basiques.*

Démonstration. Si $x \in X_{ad}$ est une solution basique, dans une certaine base de $\mathbb{R}^n$ on a $x = (x_1, ..., x_m, 0, ..., 0)$, $A = (B, N)$ avec $B = (b_1, ..., b_m)$, une base de $\mathbb{R}^m$ telle que $\sum_{i=1}^m x_i b_i = b$. Supposons qu'il existe $0 < \theta < 1$ et $y, z \in X_{ad}$ tels que $x = \theta y + (1-\theta)z$. Nécessairement, les $n-m$ dernières composantes de y et z sont nulles et, comme y et z appartiennent à X_{ad}, on a $\sum_{i=1}^m y_i b_i = b$ et $\sum_{i=1}^m z_i b_i = b$. Par unicité de la décomposition dans une base, on en déduit que $x = y = z$, et donc x est un sommet de X_{ad}.

Réciproquement, si x est un sommet de X_{ad}, on note k le nombre de ses composantes non nulles, et après un éventuel réarrangement on a $b = \sum_{i=1}^k x_i a_i$ où les (a_i) sont les colonnes de A. Pour montrer que x est une solution basique il suffit de prouver que la famille $(a_1, ..., a_k)$ est libre dans $\mathbb{R}^m$ (on obtient une base B en complétant cette famille). Supposons que ce ne soit pas le cas : il existe alors $y \neq 0$ tel que $\sum_{i=1}^k y_i a_i = 0$ et $(y_{k+1}, ..., y_n) = 0$. Comme les composantes $(x_1, ..., x_k)$ sont strictement positives, il existe $\epsilon > 0$ (petit) tel que $(x + \epsilon y) \in X_{ad}$ et $(x - \epsilon y) \in X_{ad}$. Le fait que $x = (x + \epsilon y)/2 + (x - \epsilon y)/2$ contredit le caractère extrémal de x, donc x est une solution basique. □

Le résultat fondamental suivant nous dit qu'il est suffisant de chercher une solution optimale parmi les sommets du polyèdre X_{ad}.

Proposition 4.2.6 *S'il existe une solution optimale du programme linéaire standard (4.1), alors il existe une solution optimale basique.*

Démonstration. La démonstration est très similaire à celle du Lemme 4.2.5. Soit $x \in X_{ad}$ une solution optimale de (4.1). On note k le nombre de ses composantes non nulles, et après un éventuel réarrangement on a

$$b = \sum_{i=1}^k x_i a_i,$$

où les (a_i) sont les colonnes de A. Si la famille $(a_1, ..., a_k)$ est libre dans $\mathbb{R}^m$, alors x est une solution optimale basique. Si $(a_1, ..., a_k)$ est lié, alors il existe $y \neq 0$ tel que

$$\sum_{i=1}^k y_i a_i = 0 \text{ et } (y_{k+1}, ..., y_n) = 0.$$

Comme les composantes $(x_1, ..., x_k)$ sont strictement positives, il existe $\epsilon > 0$ tel que $(x \pm \epsilon y) \in X_{ad}$. Comme x est un point de minimum, on a nécessairement

$$c \cdot x \leq c \cdot (x \pm \epsilon y),$$

c'est-à-dire $c \cdot y = 0$. On définit alors une famille de points $z_\epsilon = x + \epsilon y$ paramétrée par ϵ. En partant de la valeur $\epsilon = 0$, si on augmente ou on diminue ϵ on reste dans l'ensemble X_{ad} jusqu'à une valeur ϵ_0 au delà de laquelle la contrainte $z_\epsilon \geq 0$ est violée. Autrement dit, $z_{\epsilon_0} \in X_{ad}$ possède au plus $(k-1)$ composantes non nulles et est encore solution optimale. On répète alors l'argument précédent avec $x = z_{\epsilon_0}$ et une famille de $(k-1)$ colonnes (a_i). A force de diminuer la taille de cette famille, on obtiendra finalement une famille libre et une solution optimale basique. □

Remarque 4.2.7 En appliquant la Proposition 4.2.6 lorsque $c = 0$ (toute solution admissible est alors optimale), on voit grâce au Lemme 4.2.5 que dès que X_{ad} est non-vide, X_{ad} a au moins un sommet. Cette propriété n'a pas lieu pour des polyèdres généraux (considérer un demi-plan de $\mathbb{R}^2$). •

Exercice 4.2.1 Résoudre le programme linéaire suivant

$$\max_{x_1 \geq 0, x_2 \geq 0} x_1 + 2x_2$$

sous les contraintes

$$\begin{cases} -3x_1 + 2x_2 & \leq & 2, \\ -x_1 + 2x_2 & \leq & 4, \\ x_1 + x_2 & \leq & 5. \end{cases}$$

En pratique le nombre de sommets du polyèdre X_{ad} est gigantesque car il peut être exponentiel par rapport au nombre de variables. On le vérifie sur un exemple dans l'exercice suivant.

Exercice 4.2.2 Montrer que l'on peut choisir la matrice A de taille $m \times n$ et le vecteur $b \in \mathbb{R}^m$ de telle façon que X_{ad} soit le cube unité $[0,1]^{n-m}$ dans le sous-espace affine de dimension $n - m$ défini par $Ax = b$. En déduire que le nombre de sommets de X_{ad} est alors 2^{n-m}.

4.2.2 Algorithme du simplexe

L'algorithme du simplexe est dû à G. Dantzig dans les années 1940. Il consiste à parcourir les sommets du polyèdre des solutions admissibles jusqu'à ce qu'on trouve une solution optimale (ce qui est garanti si le programme linéaire admet effectivement une solution optimale). L'algorithme du simplexe ne se contente pas d'énumérer tous les sommets, il décroît la valeur de la fonction $c \cdot x$ en passant d'un sommet au suivant.

On considère le programme linéaire standard (4.1). Rappelons qu'un sommet (ou solution basique) de l'ensemble des solutions admissibles X_{ad} est caractérisé par

une base B (m colonnes libres de A). Après permutation de ses colonnes, on peut écrire

$$A = (B, N) \text{ et } x = (x_B, x_N),$$

de sorte qu'on a $Ax = Bx_B + Nx_N$. Toute solution admissible peut s'écrire $x_B = B^{-1}(b - Nx_N) \geq 0$ et $x_N \geq 0$. Le sommet associé à B est défini (s'il existe) par $\overline{x}_N = 0$ et $\overline{x}_B = B^{-1}b \geq 0$. Si on décompose aussi $c = (c_B, c_N)$ dans cette base, alors on peut comparer le coût d'une solution admissible quelconque x avec celui de la solution basique $\overline{x}$

$$c \cdot x - c \cdot \overline{x} = c_B \cdot B^{-1}(b - Nx_N) + c_N \cdot x_N - c_B \cdot B^{-1}b = (c_N - N^*(B^{-1})^* c_B) \cdot x_N. \quad (4.6)$$

On en déduit la condition d'optimalité suivante.

Proposition 4.2.8 *Supposons que la solution basique associée à B est non dégénérée, c'est-à-dire que $B^{-1}b > 0$. Une condition nécessaire et suffisante pour que cette solution basique associée à B soit optimale est que*

$$\tilde{c}_N = c_N - N^*(B^{-1})^* c_B \geq 0. \quad (4.7)$$

Le vecteur $\tilde{c}_N$ est appelé **vecteur des coûts réduits**.

Démonstration. Soit $\overline{x}$ une solution basique non dégénérée associée à B. Si $\tilde{c}_N \geq 0$, alors pour toute solution admissible x (4.6) implique que

$$c \cdot x - c \cdot \overline{x} = \tilde{c}_N \cdot x_N \geq 0 \ ,$$

puisque $x_N \geq 0$. Donc la condition (4.7) est suffisante pour que $\overline{x}$ soit optimal. Réciproquement, supposons qu'il existe une composante i de $\tilde{c}_N$ qui soit strictement négative, $(\tilde{c}_N \cdot e_i) < 0$. Pour $\epsilon > 0$ on définit alors un vecteur $x(\epsilon)$ par $x_N(\epsilon) = \epsilon e_i$ et $x_B(\epsilon) = B^{-1}(b - Nx_N(\epsilon))$. Par construction $Ax(\epsilon) = b$ et, comme $B^{-1}b > 0$, pour des valeurs suffisamment petites de ϵ on a $x(\epsilon) \geq 0$, donc $x(\epsilon) \in X_{ad}$. D'autre part, $x(0) = \overline{x}$ et, comme $\epsilon > 0$, on a

$$c \cdot x(\epsilon) = c \cdot x(0) + \epsilon(\tilde{c}_N \cdot e_i) < c \cdot \overline{x},$$

ce qui montre que $\overline{x}$ n'est pas optimal. Donc la condition (4.7) est nécessaire. □

Remarque 4.2.9 Dans le cadre de la Proposition 4.2.8, si la solution basique considérée est dégénérée, la condition (4.7) reste suffisante mais n'est plus nécessaire. •

On déduit de la Proposition 4.2.8 une méthode pratique pour décroître la valeur de la fonction coût $c \cdot x$ à partir d'une solution basique $\overline{x}$ (non dégénérée et non optimale). Comme $\overline{x}$ est non-optimale, il existe une composante du vecteur des coûts réduits $\tilde{c}_N$ telle que $\tilde{c}_N \cdot e_i < 0$. On définit alors $x(\epsilon)$ comme ci-dessus. Puisque le coût décroît linéairement avec ϵ, on a intérêt à prendre la plus grande valeur possible de ϵ telle que l'on reste dans X_{ad}. C'est le principe de l'algorithme du simplexe que nous présentons maintenant.

Algorithme du simplexe

— Initialisation (phase I) : on cherche une base initiale B^0 telle que la solution basique associée x^0 soit admissible

$$x^0 = \begin{pmatrix} (B^0)^{-1}b \\ 0 \end{pmatrix} \geq 0.$$

— Itérations (phase II) : à l'étape $k \geq 0$, on dispose d'une base B^k et d'une solution basique admissible x^k. On calcule le coût réduit $\tilde{c}_N^k = c_N^k - (N^k)^*(B^{k^{-1}})^* c_B^k$. Si $\tilde{c}_N^k \geq 0$, alors x^k est optimal et l'algorithme est fini. Sinon, il existe une variable hors-base d'indice i telle que $(\tilde{c}_N \cdot e_i) < 0$, et on note a_i la colonne correspondante de A. On pose

$$x^k(\epsilon) = (x_B^k(\epsilon), x_N^k(\epsilon)) \text{ avec } x_N^k(\epsilon) = \epsilon e_i \ , \ x_B^k(\epsilon) = (B^k)^{-1}(b - \epsilon a_i).$$

— Soit on peut choisir $\epsilon > 0$ aussi grand que l'on veut avec $x^k(\epsilon) \in X_{ad}$. Dans ce cas, le minimum du programme linéaire est $-\infty$.

— Soit il existe une valeur maximale $\epsilon^k \geq 0$ et un indice j tels que la j-ème composante de $x^k(\epsilon^k)$ s'annule. On obtient ainsi une nouvelle solution admissible basique

$$x^{k+1} = x^k(\epsilon^k),$$

correspondant à une nouvelle base B^{k+1} déduite de B^k en remplaçant sa j-ème colonne par la colonne a_i. La solution admissible x^{k+1} a un coût inférieur ou égal à celui de x^k.

Il reste un certain nombres de points pratiques à préciser dans l'algorithme du simplexe. Nous les passons rapidement en revue.

Dégénérescence et cyclage

On a toujours $c \cdot x^{k+1} \leq c \cdot x^k$, mais il peut y avoir égalité si la solution admissible basique x^k est dégénérée, auquel cas on trouve que $\epsilon^k = 0$ (si x^k n'est pas dégénérée, la démonstration de la Proposition 4.2.8 garantit une inégalité stricte). On a donc changé de base sans améliorer le coût : c'est le phénomène du cyclage qui peut empêcher l'algorithme de converger. Il existe des moyens de s'en prémunir, mais en pratique le cyclage n'apparaît jamais.

En l'absence de cyclage, l'algorithme du simplexe parcourt un sous-ensemble des sommets de X_{ad} en diminuant de façon stricte le coût. Comme il y a un nombre fini de sommets, l'algorithme doit nécessairement trouver un sommet optimal de coût minimal. On a donc démontré le résultat suivant.

Lemme 4.2.10 *Si toutes solutions admissibles basiques x^k produites par l'algorithme du simplexe sont non dégénérées, alors l'algorithme converge en un nombre fini d'étapes.*

A priori le nombre d'itérations de l'algorithme du simplexe peut être aussi grand que le nombre de sommets (qui est exponentiel par rapport au nombre de

variables n ; voir l'Exercice 4.2.2). Bien qu'il existe des exemples (académiques) où c'est effectivement le cas, en pratique cet algorithme converge en un nombre d'étapes qui est une fonction polynomiale de n.

Choix du changement de base

S'il y a plusieurs composantes du vecteur coût réduit $\tilde{c}_N^k$ strictement négatives, il faut faire un choix dans l'algorithme. Plusieurs stratégies sont possibles, mais en général on choisit la plus négative.

Initialisation

Comment trouver une solution admissible basique lors de l'initialisation ? (Rappelons que la condition d'admissibilité $x_B = B^{-1}b \geq 0$ n'est pas évidente en général.) Soit on en connaît une à cause de la structure du problème. Par exemple, pour le problème (4.4) qui possède m variables d'écart, $-\operatorname{Id}_m$ est une base de la matrice "globale" des contraintes d'égalité de (4.4). Si de plus $b \leq 0$, le vecteur $(x_+^0, x_-^0, \lambda^0) = (0, 0, -b)$ est alors une solution admissible basique pour (4.4).

Dans le cas général, on introduit une nouvelle variable $y \in \mathbb{R}^m$, un nouveau vecteur coût $k = (1, ..., 1)$ et un nouveau programme linéaire

$$\inf_{\substack{x\geq 0,\ y\geq 0\\ Ax+y=b}} k \cdot y\,, \tag{4.8}$$

où on a préalablement multiplié par -1 toutes les contraintes d'égalité correspondant à des composantes négatives de b de telles sortes que $b \geq 0$. Le vecteur $(x^0, y^0) = (0, b)$ est une solution admissible basique pour ce problème. S'il existe une solution admissible du programme linéaire original (4.1), alors il existe au moins une solution optimale de (4.8) et toutes les solutions optimales (x, y) vérifient nécessairement $y = 0$ et x est solution admissible de (4.1). En appliquant l'algorithme du simplexe à (4.8), on trouve ainsi une solution admissible basique pour (4.1) s'il en existe une. S'il n'en existe pas (c'est-à-dire si $X_{ad} = \emptyset$), on le détecte car le minimum de (4.8) est atteint par un vecteur (x, y) avec $y \neq 0$.

Inversion de la base

Tel que nous l'avons décrit l'algorithme du simplexe demande l'inversion à chaque étape de la base B^k, ce qui peut être très coûteux pour les problèmes de grande taille (avec beaucoup de contraintes puisque l'ordre de B^k est égal au nombre de contraintes). On peut tirer parti du fait que B^{k+1} ne diffère de B^k que par une colonne pour mettre au point une meilleur stratégie. En effet, si c'est la j-ème colonne

qui change, on a

$$B^{k+1} = B^k E^k \text{ avec } E^k = \begin{pmatrix} 1 & & l_1 & & \\ & \ddots & \vdots & 0 & \\ & 1 & \vdots & & \\ & & l_j & & \\ & & \vdots & 1 & \\ & 0 & \vdots & \ddots & \\ & & l_n & & 1 \end{pmatrix},$$

et E^k est facile à inverser

$$(E^k)^{-1} = \frac{1}{l_j} \begin{pmatrix} 1 & & -l_1 & & \\ & \ddots & \vdots & 0 & \\ & 1 & -l_{j-1} & & \\ & & 1 & & \\ & & -l_{j+1} & 1 & \\ & 0 & \vdots & \ddots & \\ & & -l_n & & 1 \end{pmatrix}.$$

On utilise donc la formule, sous forme factorisée,

$$(B^k)^{-1} = (E^{k-1})^{-1}(E^{k-2})^{-1} \cdots (E^0)^{-1}(B^0)^{-1}.$$

Exercice 4.2.3 Résoudre par l'algorithme du simplexe le programme linéaire

$$\min_{x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_4 \geq 0,\ x_5 \geq 0} x_1 + 2x_2$$

sous les contraintes

$$\begin{cases} -3x_1 + 2x_2 + x_3 = 2, \\ -x_1 + 2x_2 + x_4 = 4, \\ x_1 + x_2 + x_5 = 5. \end{cases}$$

Exercice 4.2.4 Résoudre par l'algorithme du simplexe le programme linéaire

$$\min_{x_1 \geq 0,\ x_2 \geq 0} 2x_1 - x_2$$

sous les contraintes $x_1 + x_2 \leq 1$ et $x_2 - x_1 \leq 1/2$ (on pourra s'aider d'un dessin et introduire des variables d'écart).

Exercice 4.2.5 Résoudre par l'algorithme du simplexe le programme linéaire

$$\min_{x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_4 \geq 0} 3x_3 - x_4$$

sous les contraintes

$$\begin{cases} x_1 - 3x_3 + 3x_4 = 6, \\ x_2 - 8x_3 + 4x_4 = 4. \end{cases}$$

4.2.3 Algorithmes de points intérieurs

Depuis les travaux de Khachian et Karmarkar au début des années 1980, une nouvelle classe d'algorithmes, dits de points intérieurs, est apparu pour résoudre des programmes linéaires. Le nom de cette classe d'algorithmes vient de ce qu'au contraire de la méthode du simplexe (qui, parcourant les sommets, reste sur le bord du polyèdre X_{ad}) ces algorithmes de points intérieurs évoluent à l'intérieur de X_{ad} et ne rejoignent son bord qu'à convergence. Nous allons décrire ici un de ces algorithmes que l'on appelle aussi **algorithme de trajectoire centrale**. Il y a deux idées nouvelles dans cette méthode : premièrement, on pénalise certaines contraintes à l'aide de potentiels ou fonctions "barrières" ; deuxièmement, on utilise une méthode de Newton pour passer d'une itérée à la suivante.

Décrivons cette méthode sur le programme linéaire standard

$$\inf_{x\in\mathbb{R}^n \text{tel que } Ax=b,\ x\geq 0} c\cdot x. \tag{4.9}$$

On définit un potentiel logarithmique pour $x > 0$

$$\pi(x) = -\sum_{i=1}^{n} \log x_i. \tag{4.10}$$

Pour un paramètre de pénalisation $\mu > 0$, on introduit le problème strictement convexe

$$\min_{x\in\mathbb{R}^n \text{tel que } Ax=b,\ x>0} \mu\pi(x) + c\cdot x. \tag{4.11}$$

Remarquons qu'en pratique la contrainte $x > 0$ n'en est pas une car elle n'est jamais active : quand on minimise (4.11) on ne peut pas "s'approcher" du bord de $x > 0$ sous peine de faire "exploser" le potentiel $\pi(x)$ vers $+\infty$.

Le principe de l'algorithme de trajectoire centrale est de minimiser (4.11) par une méthode de Newton pour des valeurs de plus en plus petites de μ. En effet, lorsque μ tend vers zéro, le problème pénalisé (4.11) tend vers le programme linéaire (4.9).

Exercice 4.2.6 Montrer que, si X_{ad} est borné non vide, (4.11) admet une unique solution optimale x^μ. Écrire les conditions d'optimalité et en déduire que, si (4.9) admet une unique solution optimale x^0, alors x^μ converge vers x^0 lorsque μ tend vers zéro.

4.2.4 Dualité

La théorie de la dualité (déjà évoquée lors de la Sous-section 2.6.3) est très utile en programmation linéaire. Considérons à nouveau le programme linéaire standard que nous appellerons primal (par opposition au dual)

$$\inf_{x\in\mathbb{R}^n \text{tel que } Ax=b,\ x\geq 0} c\cdot x, \tag{4.12}$$

où A est une matrice de taille $m \times n$, $b \in \mathbb{R}^m$, et $c \in \mathbb{R}^n$. Pour $p \in \mathbb{R}^m$, on introduit le Lagrangien de (4.12)

$$L(x,p) = c \cdot x + p \cdot (b - Ax), \tag{4.13}$$

où l'on a seulement "dualisé" les contraintes d'égalité. On introduit la fonction duale associée

$$G(p) = \min_{x \geq 0} L(x,p),$$

qui, après calcul, vaut

$$G(p) = \begin{cases} p \cdot b & \text{si } A^*p - c \leq 0 \\ -\infty & \text{sinon.} \end{cases} \tag{4.14}$$

Le problème dual de (4.12) est donc

$$\sup_{p \in \mathbb{R}^m \text{tel que } A^*p - c \leq 0} p \cdot b. \tag{4.15}$$

L'espace de solutions admissibles du problème dual (4.15) est noté

$$P_{ad} = \{p \in \mathbb{R}^m \text{ tel que } A^*p - c \leq 0\}.$$

Rappelons que l'espace de solutions admissibles de (4.12) est

$$X_{ad} = \{x \in \mathbb{R}^n \text{ tel que } Ax = b,\ x \geq 0\}.$$

Les programmes linéaires (4.12) et (4.15) sont dits en **dualité**. L'intérêt de cette notion vient du résultat suivant qui est un cas particulier du Théorème de dualité 2.6.11.

Théorème 4.2.11 *Si (4.12) ou (4.15) a une valeur optimale finie, alors il existe $\overline{x} \in X_{ad}$ solution optimale de (4.12) et $\overline{p} \in P_{ad}$ solution optimale de (4.15) qui vérifient*

$$\left(\min_{x \in \mathbb{R}^n \text{tel que } Ax = b,\ x \geq 0} c \cdot x \right) = c \cdot \overline{x} = \overline{p} \cdot b = \left(\max_{p \in \mathbb{R}^m \text{tel que } A^*p - c \leq 0} p \cdot b \right) \tag{4.16}$$

De plus, $\overline{x}$ et $\overline{p}$ sont solutions optimales de (4.12) et (4.15) si et seulement si elles vérifient les conditions d'optimalité de Kuhn et Tucker

$$A\overline{x} = b,\ \overline{x} \geq 0,\ A^*\overline{p} - c \leq 0,\ \overline{x} \cdot (c - A^*\overline{p}) = 0. \tag{4.17}$$

Si (4.12) ou (4.15) a une valeur optimale infinie, alors l'ensemble des solutions admissibles de l'autre problème est vide.

Remarque 4.2.12 Une conséquence immédiate du Théorème 4.2.11 de dualité est que, si $x \in X_{ad}$ et $p \in P_{ad}$ sont deux solutions admissibles de (4.12) et (4.15), respectivement, elles vérifient

$$c \cdot x \geq b \cdot p.$$

De même, si $\overline{x} \in X_{ad}$ et $\overline{p} \in P_{ad}$ vérifient

$$c \cdot \overline{x} = b \cdot \overline{p}$$

alors $\overline{x}$ est solution optimale de (4.12) et $\overline{p}$ de (4.15). Ces deux propriétés permettent de trouver facilement des bornes pour les valeurs optimales de (4.12) et (4.15), et de tester si un couple $(\overline{x}, \overline{p})$ est optimal. •

Démonstration. Supposons que X_{ad} et P_{ad} sont non vides. Soit $x \in X_{ad}$ et $p \in P_{ad}$. Comme $x \geq 0$ et $A^*p \leq c$, on a

$$c \cdot x \geq A^*p \cdot x = p \cdot Ax = p \cdot b,$$

puisque $Ax = b$. En particulier, cette inégalité implique que les valeurs optimales des deux problèmes, primal et dual, sont finies, donc qu'ils admettent des solutions optimales en vertu du Lemme 4.2.3. L'égalité (4.16) et la condition d'optimalité (4.17) sont alors une conséquence du Théorème de dualité 2.6.11.

Supposons maintenant que l'un des deux problèmes primal ou dual admet une valeur optimale finie. Pour fixer les idées, admettons qu'il s'agisse du problème dual (un argument symétrique fonctionne pour le problème primal). Alors, le Lemme 4.2.3 affirme qu'il existe une solution optimale $\overline{p}$ de (4.15). Si X_{ad} n'est pas vide, on se retrouve dans la situation précédente ce qui finit la démonstration. Montrons donc que X_{ad} n'est pas vide en utilisant encore le Lemme de Farkas 2.5.18. Pour $p \in \mathbb{R}^m$, on introduit les vecteurs de $\mathbb{R}^{m+1}$

$$\tilde{b} = \begin{pmatrix} b \\ -b \cdot \overline{p} \end{pmatrix} \quad \text{et} \quad \tilde{p} = \begin{pmatrix} p \\ 1 \end{pmatrix}.$$

On vérifie que $\tilde{b} \cdot \tilde{p} = b \cdot p - b \cdot \overline{p} \leq 0$, pour tout $p \in P_{ad}$. D'autre part, la condition $p \in P_{ad}$ peut se réécrire

$$\tilde{p} \in C = \left\{ \tilde{p} \in \mathbb{R}^{m+1} \text{ tel que } \tilde{p}_{m+1} = 1,\ \tilde{A}^*\tilde{p} \leq 0 \right\} \text{ avec } \tilde{A} = \begin{pmatrix} A \\ -c^* \end{pmatrix}.$$

Comme $\tilde{b} \cdot \tilde{p} \leq 0$ pour tout $\tilde{p} \in C$, le Lemme de Farkas 2.5.18 nous dit qu'il existe $\tilde{x} \in \mathbb{R}^n$ tel que $\tilde{x} \geq 0$ et $\tilde{b} = \tilde{A}\tilde{x}$, c'est-à-dire que $\tilde{x} \in X_{ad}$ qui n'est donc pas vide.

Finalement, supposons que la valeur optimale du problème primal est (4.12) $-\infty$. Si P_{ad} n'est pas vide, pour tout $x \in X_{ad}$ et tout $p \in P_{ad}$, on a $c \cdot x \geq b \cdot p$. En prenant une suite minimisante dans X_{ad} on obtient $b \cdot p = -\infty$, ce qui absurde. Donc P_{ad} est vide. Un raisonnement similaire montre que, si la valeur optimale de (4.12) est infinie, alors X_{ad} est vide. □

L'intérêt de la dualité pour résoudre le programme linéaire (4.12) est multiple. D'une part, selon l'algorithme choisi, il peut être plus facile de résoudre le problème dual (4.15) (qui a m variables et n contraintes d'inégalités) que le problème primal (4.12) (qui a n variables, m contraintes d'égalités et n contraintes d'inégalités). D'autre part, on peut construire des algorithmes numériques très efficaces pour la résolution de (4.12) qui utilisent les deux formes primale et duale du programme linéaire.

Exercice 4.2.7 Utiliser la dualité pour résoudre "à la main" (et sans calculs !) le programme linéaire

$$\min_{x_1\geq 0,\ x_2\geq 0,\ x_3\geq 0,\ x_4\geq 0} 8x_1 + 9x_2 + 4x_3 + 6x_4$$

sous les contraintes

$$\begin{cases} 4x_1 + x_2 + x_3 + 2x_4 \geq 1 \\ x_1 + 3x_2 + 2x_3 + x_4 \geq 1 \end{cases}$$

Exercice 4.2.8 Trouver le problème dual de (4.12) lorsqu'on dualise aussi la contrainte $x \geq 0$, c'est-à-dire qu'on introduit le Lagrangien

$$L(x, p, q) = c \cdot x + p \cdot (b - Ax) - q \cdot x$$

avec $q \in \mathbb{R}^n$ tel que $q \geq 0$. Comparer avec (4.15) et interpréter la nouvelle variable duale q. En déduire qu'il n'y a pas d'intérêt à "dualiser" aussi la contrainte $x \geq 0$.

Exercice 4.2.9 Vérifier que le problème dual de (4.15) est à nouveau (4.12).

Exercice 4.2.10 Soit $v \in \mathbb{R}^n$, $c \in \mathbb{R}^n$, A une matrice $m \times n$ et $b \in \mathbb{R}^m$. On considère le programme linéaire

$$\inf_{\substack{v\geq 0 \\ Av\leq b}} c \cdot v \ . \tag{4.18}$$

Montrer que le problème dual peut se mettre sous la forme suivante, avec $q \in \mathbb{R}^m$

$$\sup_{\substack{q\geq 0 \\ A^*q\leq c}} b \cdot q \ . \tag{4.19}$$

Soient v et q des solutions admissibles de (4.18) et (4.19), respectivement. Montrer que v et q sont des solutions optimales si, et seulement si,

$$(c - A^*q) \cdot v = 0 \quad \text{et} \quad (b - Ac) \cdot q \ . \tag{4.20}$$

Les deux égalités de (4.20) sont appelées **conditions des écarts complémentaires** (primales et duales, respectivement). Généraliser au cas où le problème primal comprend en outre des contraintes égalités.

4.3 Polyèdres entiers

Nous avons jusqu'ici traité de problèmes d'**optimisation continue** : la fonction à minimiser était différentiable, et l'ensemble des solutions admissibles était défini par l'intersection d'un nombre fini de contraintes inégalités, elles mêmes différentiables. L'**optimisation combinatoire**, au contraire, traite de problèmes pour lesquels l'ensemble des solutions admissibles est **discret**. La difficulté des problèmes combinatoires est d'une part, que l'on ne peut énumérer l'ensemble des solutions admissibles, qui est trop gros (de cardinal $n!$ dans le cas du problème d'affectation), et d'autre part, que la nature discrète de l'espace des solutions ne permet pas d'écrire directement des conditions d'optimalité à l'aide du calcul différentiel.

Nous allons cependant voir dans la suite de ce chapitre que, malgré les apparences, les méthodes de l'optimisation continue sont utiles en optimisation combinatoire. Considérons en effet le problème combinatoire typique

$$\sup_{x \in P \cap \mathbb{Z}^n} c \cdot x \ , \tag{4.21}$$

où $c \in \mathbb{R}^n$ et P est un polyèdre de $\mathbb{R}^n$. Par définition, un polyèdre P est défini comme l'intersection d'un nombre fini de demi-espaces fermés, c'est-à-dire que

$$P = \{x \in \mathbb{R}^n \mid Ax \leq b\} \tag{4.22}$$

avec $A \in \mathbb{R}^{m \times n}$ et $b \in \mathbb{R}^m$.

La formulation (4.21) montre bien la différence entre un problème combinatoire et un problème continu : si nous oublions la contrainte d'intégrité dans (4.21), nous obtenons

$$\sup_{x \in P} c \cdot x \ , \tag{4.23}$$

ce qui est un problème de programmation linéaire, parfois qualifié de problème continu **relâché, ou relaxé,** de (4.21). (De manière générale, on parle de problème relâché, ou relaxé, quand on oublie certaines contraintes.) Le problème relâché (4.23) peut se traiter efficacement par les méthodes de la section précédente : toute la difficulté de (4.21) vient de ce que nous nous restreignons aux points entiers du polyèdre P (par **point entier**, nous entendons point à coordonnées entières). Nous allons maintenant essentiellement caractériser les cas où la résolution du problème discret (4.21) est équivalente à celle de son relâché continu (4.23). Ces cas, qui peuvent sembler exceptionnels, sont en fait d'une grande importance pratique, car ils apparaissent naturellement dans un certain nombre de problèmes combinatoires concrets : plus courts chemins, affectation, et plus généralement problèmes de flots à coût minimum.

4.3.1 Points extrémaux de compacts convexes

La notion qui va permettre de relier problèmes combinatoires et problèmes discrets est celle de **point extrémal**, notion déjà rencontrée dans la Définition 4.2.1 pour un cas particulier.

Définition 4.3.1 *Soit K un ensemble convexe. Un point $x \in K$ est dit extrémal s'il ne peut pas s'écrire comme une combinaison convexe de deux autres points de K, autrement dit, si $x = (y+z)/2$ et $y, z \in K$, alors $y = z = x$. On note* $\operatorname{extr} K$ *l'ensemble des points extrémaux de K.*

Rappelons aussi que si X est un sous-ensemble de $\mathbb{R}^n$, on appelle **enveloppe convexe** de X, et l'on note $\operatorname{co} X$, le plus petit convexe contenant X, dont on vérifie qu'il est égal à l'ensemble des barycentres d'un nombre fini d'éléments de X. L'**enveloppe convexe fermée** de X, notée $\overline{\operatorname{co}}\, X$, est le plus petit convexe fermé contenant X. Il est égal à la fermeture de $\operatorname{co} X$. Le résultat suivant est fondamental.

Théorème 4.3.2 (Minkowski) *Un compact convexe K de $\mathbb{R}^n$ est enveloppe convexe de l'ensemble de ses points extrémaux, autrement dit $K = \operatorname{co} \operatorname{extr} K = \overline{\operatorname{co}} \operatorname{extr} K$.*

La preuve du théorème de Minkowski repose sur la notion d'hyperplan d'appui, introduite dans l'annexe sur les espaces de Hilbert (voir le Corollaire 8.1.13) : plus précisément, un hyperplan affine $H = \{y \in \mathbb{R}^n \mid c \cdot y = \alpha\}$, avec $c \in \mathbb{R}^n$, $c \neq 0$, et $\alpha \in \mathbb{R}$ est un **hyperplan d'appui** d'un convexe K, au point $x \in K$, si $\alpha = c \cdot x \leq c \cdot y$, pour tout $y \in K$. Nous utiliserons l'observation suivante.

Lemme 4.3.3 *Si H est un hyperplan d'appui d'un convexe $K \subset \mathbb{R}^n$, alors tout point extrémal de $H \cap K$ est point extrémal de K.*

Démonstration. Soit $H = \{y \in \mathbb{R}^n \mid c \cdot y = \alpha\}$ avec $c \in \mathbb{R}^n$, $c \neq 0$, et $\alpha \in \mathbb{R}$, un hyperplan d'appui de K. Si $x = (y + z)/2$ avec $y, z \in K$, et si $x \in K \cap H$, il vient $\alpha = c \cdot x = (c \cdot y + c \cdot z)/2$, et comme $\alpha \leq c \cdot y$ et $\alpha \leq c \cdot z$, on a nécessairement $\alpha = c \cdot y = c \cdot z$, donc $y, z \in K \cap H$. Si l'on suppose que x est un point extrémal de $K \cap H$, il vient donc $x = y = z$, ce qui montre que x est un point extrémal de K. □

Démonstration du théorème de Minkowski 4.3.2. Tout d'abord, puisque K est compact, il est fermé et donc $K = \operatorname{co} \operatorname{extr} K$ implique que $K = \overline{\operatorname{co}} \operatorname{extr} K$. On suppose que $K \neq \emptyset$ (sinon, le résultat est trivial). Rappelons que la dimension d'un convexe non-vide est par définition la dimension de l'espace affine qu'il engendre. On va montrer le théorème par récurrence sur la dimension de K. Quitte à remplacer $\mathbb{R}^n$ par un sous-espace affine, on peut supposer que K est de dimension n. Si $n = 0$, K est réduit à un point, et le théorème est vérifié. Supposons donc le théorème démontré pour les compacts convexes de dimension au plus $n - 1$, et montrons que tout point x de K est barycentre d'un nombre fini de point extrémaux de K. Si x est un point frontière de K, le Corollaire 8.1.13 fournit un hyperplan d'appui H de K en x. Comme $K \cap H$ est un compact convexe de dimension au plus $n - 1$, par hypothèse de récurrence, x est barycentre d'un nombre fini de points extrémaux de $K \cap H$, qui sont aussi des points extrémaux de K d'après le Lemme 4.3.3. Prenons maintenant un point quelconque x de K, et soit D une droite affine passant par x. L'ensemble $D \cap K$ est un segment de la forme $[y, z]$, où les points y, z sont des points frontières de K. D'après ce qui précède, y et z sont barycentres d'un nombre fini de points extrémaux de K. Comme x est lui même barycentre de y et z, le théorème est démontré. □

Nous appliquons maintenant le théorème de Minkowski au problème d'optimisation combinatoire (4.21). Dans ce cas, la fonction coût $J(x) = c \cdot x$ est linéaire, mais il sera plus clair de considérer plus généralement la **maximisation** de fonctions convexes, qui a des propriétés très différentes de la **minimisation** de fonctions convexes traitée aux Chapitres 2 et 3. Nous considérerons aussi un ensemble X arbitraire, au lieu de $P \cap \mathbb{Z}^n$.

Proposition 4.3.4 (Maximisation de fonction convexes) *Pour toute fonction convexe $J : \mathbb{R}^n \to \mathbb{R}$, et pour tout sous-ensemble $X \subset \mathbb{R}^n$,*

$$\sup_{x \in X} J(x) = \sup_{x \in \operatorname{co} X} J(x) = \sup_{x \in \overline{\operatorname{co}} X} J(x) \ , \tag{4.24}$$

et si X est borné,

$$\sup_{x\in X} J(x) = \sup_{x\in \operatorname{extr}\overline{\operatorname{co}}\, X} J(x) \ . \tag{4.25}$$

Démonstration. Si $y \in \operatorname{co} X$, on peut écrire $y = \sum_{1\le i\le k} \alpha_i x_i$, avec $x_i \in X$, $\alpha_i \ge 0$, et $\sum_{1\le j\le k} \alpha_j = 1$. Puisque J est convexe, on a $J(y) \le \sum_{1\le j\le k} \alpha_j J(x_j) \le \max_{1\le j\le k} J(x_j) \le \sup_{x\in X} J(x)$, et puisque ceci est vrai pour tout $y \in \operatorname{co} X$, on a $\sup_{x\in \operatorname{co} X} J(x) \le \sup_{x\in X} J(x)$. Par ailleurs, pour tout $z \in \overline{\operatorname{co}}\, X$, on peut écrire $z = \lim_{k\to\infty} y_k$, avec $y_k \in \operatorname{co} X$. Comme une fonction convexe propre $\mathbb{R}^n \to \mathbb{R}$ est nécessairement continue (cf. Lemme 2.3.5), on a $J(z) = \lim_{k\to\infty} J(y_k) \le \sup_{x\in \operatorname{co} X} J(x)$, et puisque ceci est vrai pour tout $z \in \overline{\operatorname{co}}\, X$, on a $\sup_{x\in \overline{\operatorname{co}}\, X} J(x) \le \sup_{x\in \operatorname{co} X} J(x)$. Les autres inégalités étant triviales, on a montré (4.24). Lorsque X est borné, $\overline{\operatorname{co}}\, X$ qui est aussi borné, est compact. D'après le théorème de Minkowski 4.3.2, $\overline{\operatorname{co}}\, X = \operatorname{co}\operatorname{extr}\overline{\operatorname{co}}\, X$, et en appliquant (4.24), $\sup_{x\in \operatorname{extr}\overline{\operatorname{co}}\, X} J(x) = \sup_{x\in \operatorname{co}\operatorname{extr}\overline{\operatorname{co}}\, X} J(x) = \sup_{x\in \overline{\operatorname{co}}\, X} J(x) = \sup_{x\in X} J(x)$, ce qui prouve (4.25). □

La Proposition 4.3.4 nous suggère de considérer l'enveloppe convexe de l'ensemble admissible $X = P \cap \mathbb{Z}^n$ de notre problème initial (4.21).

Définition 4.3.5 *On appelle* **enveloppe entière** *d'un polyèdre $P \subset \mathbb{R}^n$, l'enveloppe convexe de l'ensemble des points entiers de P, que l'on note $P_e = \operatorname{co}(P \cap \mathbb{Z}^n)$.*

Le terme "enveloppe entière" est traditionnel mais légèrement trompeur : d'ordinaire, une enveloppe est un objet plus gros, alors qu'ici $P_e \subset P$. En remplaçant X par $P \cap \mathbb{Z}^n$ dans la Proposition 4.3.4 on obtient le corollaire suivant.

Corollaire 4.3.6 *Si $J : \mathbb{R}^n \to \mathbb{R}$ est convexe, et si $P \subset \mathbb{R}^n$ est un polyèdre, alors*

$$\sup_{x\in P\cap\mathbb{Z}^n} J(x) = \sup_{x\in P_e} J(x) \ . \tag{4.26}$$

Ainsi, on peut toujours remplacer le problème discret (4.21) par un problème dont l'ensemble admissible est un convexe. Lorsque J est linéaire, le problème à droite de (4.26) est un programme linéaire classique : on a ainsi concentré la difficulté dans le calcul, ou l'approximation, du polyèdre P_e. Il y a un cas où tout devient facile.

Définition 4.3.7 *On dit qu'un polyèdre P est un* **polyèdre entier** *si $P = P_e$.*

Nous allons maintenant donner des conditions suffisantes (précises) pour qu'un polyèdre soit entier.

4.3.2 Matrices totalement unimodulaires

Rappelons qu'un polyèdre quelconque P, défini par (4.22), est plus général que le polyèdre X_{ad} des solutions admissibles du programme linéaire standard (cf. Définition 4.2.1) : en effet, X_{ad} est par définition inclus dans le cône positif de $\mathbb{R}^n$. Par ailleurs, nous avons déjà noté dans la Remarque 4.2.7 que X_{ad}, s'il est non-vide, a toujours des points extrémaux, ce qui n'est pas le cas pour un polyèdre quelconque (prendre un demi-espace). La caractérisation des points extrémaux de X_{ad} (Lemme 4.2.5) s'étend cependant de la manière suivante.

Lemme 4.3.8 *Un point extrémal du polyèdre P défini par* (4.22) *est nécessairement solution d'un système $A'x = b'$, où A' est une sous-matrice inversible formée de n lignes de A, et b' est le vecteur formé des composantes correspondantes de b.*

Démonstration. Soit x un point extrémal de P, et soit $I(x) = \{1 \leq i \leq m \mid A_i \cdot x = b_i\}$ (l'ensemble des contraintes actives en x), où A_i désigne la i-ème ligne de A. Si la famille $\{A_i\}_{i \in I(x)}$, n'est pas de rang n, on peut trouver un vecteur non nul y tel que $A_i \cdot y = 0$ pour tout $i \in I(x)$. Comme x est le milieu des points $x - \epsilon y$ et $x + \epsilon y$, qui sont bien des éléments de P si ϵ est assez petit, on contredit l'extrémalité de x. Ainsi, on peut trouver un sous ensemble $I' \subset I(x)$ de cardinal n tel que la matrice $n \times n$ dont les lignes sont les A_i, avec $i \in I'$, est inversible. Le système $A_i \cdot x = b_i, i \in I'$, caractérise alors x. □

Le Lemme 4.3.8 montre en particulier qu'un polyèdre n'a qu'un nombre fini de points extrémaux. Par ailleurs, il suggère d'étudier les cas où la solution d'un système linéaire est entière.

Proposition 4.3.9 *Soit $A \in \mathbb{Z}^{n \times n}$ une matrice inversible. Les assertions suivantes sont équivalentes :*

1. $\det A = \pm 1$ *;*
2. *pour tout $b \in \mathbb{Z}^n$, on a $A^{-1}b \in \mathbb{Z}^n$.*

Démonstration. L'implication 1⇒2 résulte aussitôt des formules de Cramer. Réciproquement, supposons que A vérifie l'assertion 2. Montrons d'abord que A^{-1} est à coefficients entiers. En prenant pour b le i-ème vecteur de la base canonique de $\mathbb{R}^n$, on voit que la i-ème colonne de A^{-1}, qui coïncide avec $A^{-1}b$, est à coefficients entiers. Comme ceci est vrai pour tout $1 \leq i \leq n$, on a $A^{-1} \in \mathbb{Z}^{n \times n}$. Donc $\det A^{-1} \in \mathbb{Z}$, et $1 = \det A \det A^{-1}$ montre que $\det A$ divise 1, c'est-à-dire que $\det A = \pm 1$. □

Définition 4.3.10 *On dit qu'une matrice $A \in \mathbb{Z}^{n \times n}$ est* **unimodulaire** *quand $\det A = \pm 1$, et qu'une matrice $D \in \mathbb{Z}^{m \times n}$ est* **totalement unimodulaire** *quand toute sous-matrice carrée extraite de D est de déterminant ± 1 ou 0.*

En prenant des sous-matrices 1×1, on voit en particulier que les coefficients d'une matrice totalement unimodulaire valent nécessairement ± 1 ou 0. L'introduction des matrices totalement unimodulaires est motivée par le résultat suivant.

Théorème 4.3.11 (Optimalité des solutions entières) *Soient $D \in \mathbb{Z}^{m \times n}$ une matrice totalement unimodulaire, $d_{\max} \in (\mathbb{Z} \cup \{+\infty\})^m$, $d_{\min} \in (\mathbb{Z} \cup \{-\infty\})^m$, $x_{\max} \in (\mathbb{Z} \cup \{+\infty\})^n$, et $x_{\min} \in (\mathbb{Z} \cup \{-\infty\})^n$. Alors, les points extrémaux du polyèdre*

$$Q = \{x \in \mathbb{R}^n \mid d_{\min} \leq Dx \leq d_{\max}, \quad x_{\min} \leq x \leq x_{\max}\} \tag{4.27}$$

sont nécessairement entiers. En particulier, si Q est borné, on a $Q = Q_e$, et pour toute fonction convexe J de $\mathbb{R}^n$ dans $\mathbb{R}$, on a

$$\max_{x \in Q} J(x) = \max_{x \in Q \cap \mathbb{Z}^n} J(x) \ . \tag{4.28}$$

Remarque 4.3.12 Le Théorème 4.3.11 est fondamental en pratique car il permet de s'affranchir de la contrainte de solutions entières pour calculer le maximum de J, tout en garantissant qu'il existe une solution optimale entière.

Le Théorème 4.3.11 ne dit surtout pas que toutes les solutions optimales sont entières. D'ailleurs, lorsque J est linéaire, il ne peut en être ainsi à moins que la solution optimale ne soit unique, car tout barycentre de solutions optimales d'un programme linéaire est solution optimale.

On notera que le Théorème 4.3.11 ne pose aucune condition sur J, hormis la convexité. En particulier, si $J(x) = c \cdot x$ est linéaire, le caractère entier des solutions optimales n'est pas directement relié au caractère entier du vecteur de coût c. •

Démonstration. On peut écrire

$$Q = \{x \in \mathbb{R}^n \mid Ax \leq b\} \ , \tag{4.29}$$

où b est un vecteur entier fini et A est une matrice dont chaque ligne est soit de la forme $\pm D_i$, avec D_i une ligne quelconque de D, soit de la forme $\pm e_j$, où e_j est le j-ème vecteur de la base canonique de $\mathbb{R}^n$, pour un indice quelconque $1 \leq j \leq n$.

On montre d'abord que A est totalement unimodulaire. Soit donc M une sous-matrice $k \times k$ extraite de A. Montrons par récurrence sur k que $\det M \in \{\pm 1, 0\}$. Si $k = 1$, cela résulte aussitôt de la totale unimodularité de D. Supposons maintenant le résultat prouvé pour toutes les sous-matrices carrées de A de dimension au plus $k-1$, et montrons le pour M. Si M contient une ligne égale à un vecteur $\pm e_j$, on développe $\det M$ par rapport à cette ligne, et par récurrence, le résultat est prouvé. Si M contient deux lignes égales au signe près, $\det M = 0$, et le résultat est encore prouvé. Sinon, M coïncide, au changement du signe de certaines lignes près, avec une sous-matrice de D, et comme D est totalement unimodulaire, $\det M \in \{\pm 1, 0\}$, ce qui achève la preuve de la totale unimodularité de A.

Comme Q est donné par (4.29), avec b entier et A totalement unimodulaire, il résulte du Lemme 4.3.8 et de la Proposition 4.3.9 que les points extrémaux de Q, s'ils existent, sont entiers.

Si l'on suppose en outre que Q est borné, Q est compact, et d'après le Théorème de Minkowski 4.3.2, $Q = \operatorname{co} \operatorname{extr} Q$. Comme $\operatorname{extr} Q$ est formé de vecteurs entiers, $Q_e = \operatorname{co}(Q \cap \mathbb{Z}^n) \supset \operatorname{co} \operatorname{extr} Q = Q$, et par ailleurs l'inclusion $Q_e \subset Q$ est triviale. L'égalité (4.28) est alors obtenue en appliquant le Corollaire 4.3.6. □

Il existe de nombreux résultats sur les matrices totalement unimodulaires. Nous nous bornons ici à donner une condition suffisante très utile.

Proposition 4.3.13 (Poincaré) *Si A est une matrice à coefficients ± 1 ou 0, avec au plus un coefficient 1 par colonne, et au plus un coefficient -1 par colonne, alors A est totalement unimodulaire.*

Démonstration. Comme la propriété que vérifie A passe aux sous-matrices, il suffit de vérifier que si A est carrée, alors $\det A \in \{\pm 1, 0\}$. Si A a une colonne nulle, $\det A = 0$. Si A a une colonne avec seulement un coefficient non-nul, on développe le déterminant par rapport à cette colonne, et l'on conclut par récurrence

que $\det A \in \{\pm 1, 0\}$. Il ne reste qu'a considérer le cas où chaque colonne de A a exactement un coefficient 1 et un coefficient -1 (tous les autres étant nuls) : alors, chaque colonne a la some de ces coefficients qui est nulle. Donc le vecteur $(1, \cdots, 1)$ appartient au noyau de la transposée de A et ainsi $\det A = 0$. □

Nous appliquerons la Proposition 4.3.13 aux problèmes de flots dans la sous-section suivante.

4.3.3 Problèmes de flots

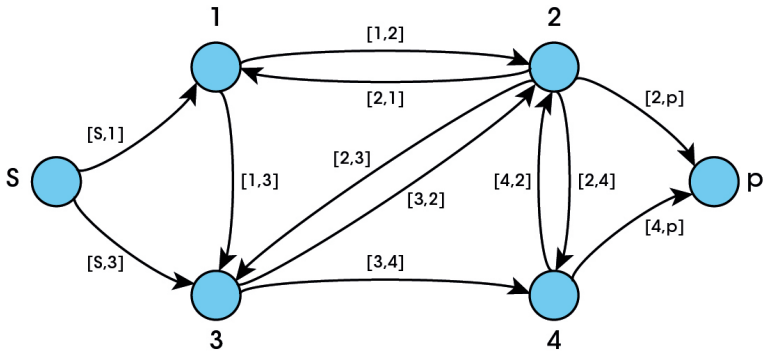

FIGURE 4.2 – Graphe orienté.

Afin de définir les problèmes de flots, considérons un graphe orienté $\mathcal{G} = (\mathcal{N}, \mathcal{A})$: $\mathcal{N}$ est l'ensemble des nœuds, et $\mathcal{A} \subset \mathcal{N} \times \mathcal{N}$ est l'ensemble des **arcs**. Un arc allant du nœud i au nœud j est ainsi noté (i, j). On munit chaque arc $(i, j) \in \mathcal{A}$ d'une **capacité** $u_{ij} \in \mathbb{R}_+ \cup \{+\infty\}$ et d'un **coût** $c_{ij} \in \mathbb{R}$. Le "u" dans u_{ij} est pour "upper bound". Attention, l'arc (i, j) n'est pas le même que l'arc (j, i) (voir la Figure 4.2). On se donne aussi en chaque nœud du graphe un flot entrant exogène $b_i \in \mathbb{R}$ (si $b_i < 0$, il s'agit d'un flot sortant, compté algébriquement). On appelle **flot** une fonction $x \in \mathbb{R}^{\mathcal{A}}$, $(i, j) \mapsto x_{ij}$, vérifiant la **loi des nœuds de Kirchoff**

$$b_i + \sum_{j \in \mathcal{N}, (j,i) \in \mathcal{A}} x_{ji} = \sum_{j \in \mathcal{N}, (i,j) \in \mathcal{A}} x_{ij}, \qquad \forall i \in \mathcal{N}, \tag{4.30}$$

ainsi que la contrainte de positivité

$$0 \leq x_{ij}, \qquad \forall (i, j) \in \mathcal{A} \ . \tag{4.31}$$

En sommant les lois des nœuds (4.30), on voit qu'une condition nécessaire pour l'existence d'un flot est que la somme des flots exogènes entrants soit nulle

$$\sum_{i \in \mathcal{N}} b_i = 0 \ . \tag{4.32}$$

Nous supposerons toujours que la condition (4.32) est vérifiée. Un flot est dit **admissible** s'il satisfait les contraintes de capacité

$$x_{ij} \leq u_{ij}, \qquad \forall (i, j) \in \mathcal{A} \ . \tag{4.33}$$

Définition 4.3.14 *On appelle* **problème de flot à coût minimum** *le programme linéaire*

$$\min_{x\in\mathbb{R}^{\mathcal{A}}} \sum_{(i,j)\in\mathcal{A}} c_{ij}x_{ij} \text{ sous les contraintes } (4.30), (4.31), (4.33). \tag{4.34}$$

Un cas particulier fondamental du problème de flot à coût minimum est le problème du flot maximal, ou **problème de flot** proprement dit, qui concerne seulement les capacités (et non les coûts). Il sera commode de supposer que $\mathcal{G}$ a deux nœuds distingués, s et p, appelés respectivement **source** et **puits**, tels que s n'a pas de prédécesseur ($\{i \in \mathcal{N} \mid (i,s) \in \mathcal{A}\} = \emptyset$), et p n'a pas de successeur ($\{i \in \mathcal{N} \mid (p,i) \in \mathcal{A}\} = \emptyset$). Soit $v \in \mathbb{R}_+$. On appelle **flot admissible de s à p de valeur** v une solution x de (4.30),(4.31),(4.33), avec

$$b_i = \begin{cases} v & \text{si } i = s, \\ -v & \text{si } i = p, \\ 0 & \text{sinon.} \end{cases}$$

Définition 4.3.15 *Le* **problème du flot maximal** *consiste à trouver un flot admissible de s à p de valeur v maximale.*

Exercice 4.3.1 Montrer que le problème du flot maximal est effectivement un cas particulier de problème de flot à coût minimal. (Indication : rajouter un arc reliant p à s au graphe intervenant dans la définition du problème du flot maximal, supprimer les deux flots externes b_s et b_p et fixer le coût $c_{ps} = -1$.)

Les problèmes de flot à coût minimum permettent de modéliser plusieurs exemples d'optimisation importants, comme le problème de transport de l'Exemple 1.2.1, ou le problème d'affectation de l'Exemple 1.2.2 (en ignorant pour l'instant la distinction entre solution réelle ou entière).

Exercice 4.3.2 Expliciter le problème de flot à coût minimum correspondant au problème de transport de l'Exemple 1.2.1 ou au problème d'affectation de l'Exemple 1.2.2. (On dessinera le graphe.)

En pratique, on cherche souvent des solutions entières d'un problème de flot : par exemple, pour le problème de transport de l'Exemple 1.2.1, les marchandises à livrer peuvent être des colis, et livrer un demi-colis peut ne pas avoir de sens. Il est donc naturel de se demander si un problème de flot à coût minimum a automatiquement des solutions optimales entières. Afin d'appliquer le Théorème 4.3.11, notons que la loi des nœuds de Kirchoff (4.30) peut s'écrire $Ax = b$, où la matrice $A \in \mathbb{R}^{\mathcal{N}\times\mathcal{A}}$, appelée **matrice d'incidence nœuds-arcs** de $\mathcal{G}$, est définie par

$$A_{i,(j,k)} = \begin{cases} -1 & \text{si } i = k, \\ 1 & \text{si } i = j, \\ 0 & \text{sinon.} \end{cases}$$

La matrice A est bien définie, sauf dans le cas dégénéré où le graphe a une boucle, c'est-à-dire un arc (j,k) tel que $j = k$. Un flot circulant sur une boucle a une

contribution qui se simplifie dans la loi des nœuds de Kirchoff (4.30), aussi n'y a-t-il aucune perte de généralité à supposer le graphe sans boucle, ce que nous ferons dans la suite de la section.

Nous pouvons maintenant écrire l'ensemble des flots admissibles

$$\{x \in \mathbb{R}^{\mathcal{A}} \mid Ax = b,\ 0 \leq x \leq u\} \ . \tag{4.35}$$

Proposition 4.3.16 *La matrice d'incidence nœuds-arcs A d'un graphe est totalement unimodulaire.*

Démonstration. Par construction la matrice A a exactement un seul 1 et un seul -1 dans chacune de ses colonnes car chaque arc possède un unique nœud de départ et un unique nœud d'arrivée. Le résultat s'obtient donc par application de la Proposition 4.3.13. □

Théorème 4.3.17 (Optimalité des flots entiers) *Si les flots entrants exogènes b_i sont entiers, et si les capacités u_{ij} sont entières ou infinies, alors, les points extrémaux de l'ensemble* (4.35) *des flots admissibles sont entiers. En particulier, si cet ensemble* (4.35) *est borné et non-vide, le problème* (4.34) *de flot à coût minimal admet une solution optimale entière $x \in \mathbb{N}^{\mathcal{A}}$.*

Démonstration. C'est une conséquence immédiate du Théorème 4.3.11 et de la Proposition 4.3.16. □

Une conséquence pratique importante du Théorème 4.3.17 est qu'on peut résoudre les problèmes de flots entiers par l'algorithme du simplexe. Non seulement la valeur optimale du coût est la même si on optimise sur les réels ou sur les entiers mais, en plus, comme l'algorithme du simplexe itère d'un sommet à un autre du polyèdre des solutions admissibles, les solutions trouvées par le simplexe sont entières car les sommets (ou points extrémaux) sont entiers.

Chapitre 5

CONTRÔLABILITÉ DES SYSTÈMES DIFFÉRENTIELS

5.1 Contrôlabilité des systèmes linéaires

Cette section est consacrée à la contrôlabilité des systèmes linéaires. Le principal résultat est le **critère de Kalman** qui fournit une condition nécessaire et suffisante pour la contrôlabilité d'un système linéaire autonome. De manière tout à fait remarquable, ce critère se formule de manière purement algébrique et la condition à vérifier est indépendante de la condition initiale et de l'horizon temporel. Dans un deuxième temps, nous considérons des systèmes de contrôle linéaires avec des bornes sur le contrôle. Cela nous conduit à introduire la notion importante d'**ensemble atteignable**.

5.1.1 Systèmes de contrôle linéaires

Soit $T > 0$ un horizon temporel fixé. On considère un système dynamique dont l'état $x(t) \in \mathbb{R}^d$ pour tout $t \in [0,T]$ est régi par le système différentiel

$$\dot{x}(t) = Ax(t) + Bu(t), \quad \forall t \in [0,T], \qquad x(0) = x_0 \in \mathbb{R}^d, \tag{5.1}$$

avec des matrices $A \in \mathbb{R}^{d\times d}$, $B \in \mathbb{R}^{d\times k}$, où $d \geq 1$ et $k \geq 1$. La fonction temporelle

$$u : [0,T] \to \mathbb{R}^k$$

nous permet d'agir sur le système afin d'en modifier l'état. On dit que u est le **contrôle**. Une fois le contrôle u fixé, (5.1) est un **problème de Cauchy**. Afin d'expliciter le fait que la trajectoire x, solution de (5.1), dépend du contrôle u, nous la noterons souvent x_u, et nous écrirons (5.1) sous la forme

$$\dot{x}_u(t) = Ax_u(t) + Bu(t), \quad \forall t \in [0,T], \qquad x_u(0) = x_0 \in \mathbb{R}^d. \tag{5.2}$$

Par la suite, nous supposerons que

$$u \in L^1([0,T];\mathbb{R}^k),$$

et nous serons parfois amenés à faire des hypothèses un peu plus fortes sur le contrôle, comme par exemple que u prend ses valeurs dans un sous-ensemble fermé non-vide U de $\mathbb{R}^k$, ce que nous noterons $u \in L^1([0,T];U)$; nous ferons parfois des hypothèses d'intégrabilité plus forte en temps, comme par exemple $L^2([0,T];U)$ ou $L^\infty([0,T];U)$. Rappelons à toutes fins utiles que l'espace $L^1([0,T];\mathbb{R}^k)$ est équipé de la norme

$$\|u\|_{L^1([0,T];\mathbb{R}^k)} = \int_0^T |u(s)|_{\mathbb{R}^k}\,\mathrm{d}s,$$

où $|\cdot|_{\mathbb{R}^k}$ désigne la norme euclidienne sur $\mathbb{R}^k$. (On peut remplacer la norme euclidienne par toute autre norme sur $\mathbb{R}^k$.) Rappelons que la notation * désigne la transposition des vecteurs ou des matrices ; on écrit donc x^*y pour le produit scalaire entre deux vecteurs et Z^* pour la transposée de la matrice Z.

Définition 5.1.1 (Systèmes de contrôle linéaires) *On dit que* (5.2) *est un* ***système de contrôle linéaire****. On dit que ce système est* ***autonome*** *(ou* ***stationnaire****) lorsque les matrices A et B ne dépendent pas du temps. Plus généralement, on dit que le système de contrôle linéaire est* ***instationnaire*** *lorsqu'il s'écrit sous la forme*

$$\dot{x}_u(t) = A(t)x_u(t) + B(t)u(t), \quad \forall t \in [0,T], \qquad x_u(0) = x_0,$$

avec $A \in L^1([0,T];\mathbb{R}^{d\times d})$ et $B \in L^1([0,T];\mathbb{R}^{d\times k})$. Enfin, on dit que le système de contrôle linéaire a un ***terme de dérive*** *lorsqu'il s'écrit sous la forme*

$$\dot{x}_u(t) = Ax_u(t) + Bu(t) + f(t), \quad \forall t \in [0,T], \qquad x_u(0) = x_0,$$

avec $f \in L^1([0,T];\mathbb{R}^d)$, les matrices A et B pouvant ou non dépendre du temps.

Ici, nous considérerons le système de contrôle linéaire autonome (5.1). La première question à se poser est si, pour tout contrôle $u \in L^1([0,T];\mathbb{R}^k)$ fixé, il existe une unique trajectoire $x : [0,T] \to \mathbb{R}^d$ associée à ce contrôle, solution du problème de Cauchy (5.1). Comme le contrôle u n'est *a priori* pas une fonction continue du temps, on ne peut pas chercher une trajectoire de classe $C^1([0,T];\mathbb{R}^d)$. Un bon cadre fonctionnel pour la trajectoire est celui des fonctions absolument continues sur $[0,T]$, dont on donne la définition.

Définition 5.1.2 (Fonction absolument continue) *On dit qu'une fonction $F : [0,T] \to \mathbb{R}^d$ est absolument continue sur $[0,T]$ et on écrit $F \in AC([0,T];\mathbb{R}^d)$ s'il existe $f \in L^1([0,T];\mathbb{R}^d)$ telle que*

$$F(t) - F(0) = \int_0^t f(s)\,\mathrm{d}s, \quad \forall t \in [0,T].$$

Si une fonction F est absolument continue sur $[0,T]$, alors elle est continue sur $[0,T]$ et elle est dérivable presque partout, de dérivée égale à f.

Proposition 5.1.3 (Formule de Duhamel) *Pour tout contrôle $u \in L^1([0,T];\mathbb{R}^k)$, il existe une unique trajectoire $x_u \in AC([0,T];\mathbb{R}^d)$ solution de (5.1) au sens où cette trajectoire vérifie la condition initiale $x_u(0) = 0$ et le système différentiel $\dot{x}_u(t) = Ax_u(t) + Bu(t)$ presque partout (p.p.) sur $[0,T]$. Cette trajectoire est donnée par la formule de Duhamel*

$$x_u(t) = e^{tA}x_0 + \int_0^t e^{(t-s)A}Bu(s)\,\mathrm{d}s, \quad \forall t \in [0,T]. \tag{5.3}$$

On notera que cette expression a bien un sens pour $u \in L^1([0,T];\mathbb{R}^k)$ car la fonction $s \mapsto e^{(t-s)A}$ est bornée sur $[0,T]$.

Remarque 5.1.4 On rappelle que, pour une matrice carrée A, l'exponentielle de matrice est définie par $e^A = \sum_{n\geq 0} \frac{1}{n!}A^n$. Par ailleurs, $\frac{d}{dt}e^{tA} = Ae^{tA} = e^{tA}A$, et si A_1, A_2 commutent ($A_1A_2 = A_2A_1$), alors $e^{A_1}e^{A_2} = e^{A_2}e^{A_1} = e^{A_1+A_2}$.

5.1.2 Cas sans contraintes : critère de Kalman

Définition 5.1.5 (Contrôlabilité) *On dit que le système (5.2) est contrôlable en temps T à partir de x_0 si*

$$\forall x_1 \in \mathbb{R}^d, \quad \exists u \in L^\infty([0,T];\mathbb{R}^k), \quad x_u(T) = x_1.$$

On cherche donc à atteindre la cible x_1 au temps T à partir de x_0.

Remarque 5.1.6 Dans la Définition 5.1.5 on pourrait demander à ce que le contrôle u ne soit pas forcément borné, par exemple $u \in L^1([0,T];\mathbb{R}^k)$.

En posant $x_2 = x_1 - e^{TA}x_0$, la contrôlabilité en T à partir de x_0 équivaut à

$$\forall x_2 \in \mathbb{R}^d, \quad \exists u \in L^\infty([0,T];\mathbb{R}^k), \quad x_2 = \int_0^T e^{(T-s)A}Bu(s)\,\mathrm{d}s,$$

i.e., à la **surjectivité** de l'application

$$\Phi : L^\infty([0,T];\mathbb{R}^k) \to \mathbb{R}^d, \quad \Phi(u) = \int_0^T e^{(T-s)A}Bu(s)\,\mathrm{d}s. \tag{5.4}$$

Un résultat remarquable, dû à Kalman, permet de caractériser la surjectivité de cette application à partir d'une condition **purement algébrique** ne faisant intervenir que les matrices A et B.

Théorème 5.1.7 (Critère de Kalman) *Le système linéaire autonome $\dot{x}_u(t) = Ax_u(t)+Bu(t)$ est contrôlable pour tout $T > 0$ et pour tout $x_0 \in \mathbb{R}^d$* ***si et seulement si*** *la matrice de Kalman $C \in \mathbb{R}^{d\times dk}$, définie par*

$$C = \left(B, AB, \cdots, A^{d-1}B\right),$$

est de rang maximal, ce qui signifie que

$$\mathrm{rang}(C) = d. \tag{5.5}$$

Remarque 5.1.8 La condition de Kalman (5.5) est **indépendante** de l'horizon temporel $T > 0$ et de la donnée initiale $x_0 \in \mathbb{R}^d$. La contrôlabilité d'un système linéaire autonome est donc indépendante de ces deux paramètres. Cela signifie en particulier que lorsqu'un système de contrôle linéaire autonome est contrôlable, on peut atteindre à partir d'une donnée initiale toute cible, même très lointaine, en un horizon temporel même très court. Ce n'est pas très surprenant dans la mesure où on ne s'est pas imposé de bornes sur la valeur du contrôle ; celui-ci peut donc prendre des valeurs très grandes si nécessaire.

Remarque 5.1.9 On vérifie facilement que la condition de Kalman est invariante par changement de base. En effet, soit $P \in \mathbb{R}^{d\times d}$ une matrice inversible de changement de base. On considère le système linéaire autonome $\dot{x}(t) = Ax(t) + Bu(t)$. Dans la nouvelle base, ce système s'écrit

$$\dot{y}(t) = \widetilde{A}y(t) + \widetilde{B}u(t),$$

avec $y(t) = P^{-1}x(t)$, $\widetilde{A} = P^{-1}AP$, $\widetilde{B} = P^{-1}B$, si bien que

$$\widetilde{C} = \left(\widetilde{B}, \widetilde{A}\widetilde{B}, \cdots, \widetilde{A}^{d-1}\widetilde{B}\right) = P^{-1}C.$$

Par conséquent, $\mathrm{rang}(C) = \mathrm{rang}(\widetilde{C})$.

Démonstration. (1) Supposons d'abord que $\mathrm{rang}(C) < d$. Par conséquent les lignes de C sont liées et il existe un vecteur $\Psi \in \mathbb{R}^d$, $\Psi \neq 0$, tel que

$$\Psi^*B = \Psi^*AB = \cdots = \Psi^*A^{d-1}B = 0\,(\in \mathbb{R}^k),$$

où Ψ^* désigne le transposé de Ψ (Ψ^* est un vecteur ligne). D'après le théorème de Cayley-Hamilton, il existe des réels $s_0, \cdots, s_{d-1}$ tels que

$$A^d = s_0 I_d + \cdots + s_{d-1}A^{d-1},$$

où I_d est la matrice identité dans $\mathbb{R}^{d\times d}$. On en déduit par récurrence que $\Psi^*A^kB = 0$ pour tout $k \in \mathbb{N}$, puis que $\Psi^*e^{tA}B = 0$ pour tout $t \in [0,T]$. Par conséquent, $\Psi^*\Phi(u) = 0$ pour tout contrôle u, i.e., l'application Φ ne peut être surjective.
(2) Réciproquement, si l'application Φ n'est pas surjective, il existe un vecteur $\Psi \in \mathbb{R}^d$, $\Psi \neq 0$, tel que

$$\Psi^* \int_0^T e^{(T-s)A}Bu(s)\,\mathrm{d}s = 0, \quad \forall u \in L^\infty([0,T];\mathbb{R}^k).$$

En choisissant le contrôle $u(s) = B^*e^{(T-s)A^*}\Psi$, qui est bien dans $L^\infty([0,T];\mathbb{R}^k)$, on en déduit que

$$\Psi^*e^{tA}B = 0\ (\in \mathbb{R}^k), \quad \forall t \in [0,T].$$

En $t = 0$, il vient $\Psi^*B = 0$, puis en dérivant par rapport à t, il vient $\Psi^*AB = 0$ et ainsi de suite ; d'où

$$\Psi^*B = \Psi^*AB = \cdots = \Psi^*A^{d-1}B = 0\,(\in \mathbb{R}^k).$$

La matrice C ne peut donc être de rang maximal. □

Exemple 5.1.10 [Contrôle d'un tram] On reprend l'Exemple 1.3.1 où l'état du tram (supposé de masse unité) est décrit par sa position $X(t)$ et sa vitesse $V(t)$ le long d'un axe unidirectionnel et on contrôle l'accélération du tram sous la forme

$$\ddot{X}(t) = u(t), \quad \forall t \in [0,T].$$

Cette équation différentielle du second ordre en temps se récrit comme un système d'ordre un en temps (avec $d = 2$, $k = 1$) :

$$\dot{x}(t) = \underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{=:A} x(t) + \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{=:B} u(t), \qquad x(t) = \begin{pmatrix} X(t) \\ V(t) \end{pmatrix}.$$

La matrice de Kalman $C \in \mathbb{R}^{2\times 2}$ est

$$C = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \text{rang}(C) = 2.$$

Le tram est donc contrôlable en tout temps T à partir de tout $x_0 = (X_0, V_0)^*$ (position et vitesse initiales) : cela signifie que quel que soit $x_1 = (X_1, V_1)^*$ (position et vitesse cibles en T), il existe un contrôle $u \in L^\infty([0,T];\mathbb{R})$ amenant le tram de x_0 en x_1 au temps T.

Exemple 5.1.11 [Circuit RLC] Considérons maintenant un exemple issu de l'électronique : le circuit RLC. Ici, x (l'état) représente la charge du circuit et u (le contrôle) la tension appliquée

$$u(t) = L\ddot{x}(t) + R\dot{x}(t) + C^{-1}x(t),$$

ou encore $\ddot{x}(t) = -\frac{R}{L}\dot{x}(t) - \frac{1}{LC}x(t) + \frac{1}{L}u(t)$ On obtient le système de contrôle linéaire (avec $d = 2$, $k = 1$) sous la forme

$$\dot{X}(t) = \begin{pmatrix} 0 & 1 \\ \frac{-1}{LC} & \frac{-R}{L} \end{pmatrix} X(t) + \begin{pmatrix} 0 \\ \frac{1}{L} \end{pmatrix} u(t), \qquad X(t) = \begin{pmatrix} x(t) \\ \dot{x}(t) \end{pmatrix}.$$

La matrice de Kalman $C \in \mathbb{R}^{2\times 2}$ est

$$C = \begin{pmatrix} 0 & \frac{1}{L} \\ \frac{1}{L} & \frac{-R}{L^2} \end{pmatrix}, \qquad \text{rang}(C) = 2,$$

ce qui montre que le circuit RLC est contrôlable.

Il est intéressant de considérer une reformulation du critère de Kalman. On introduit la matrice $G_T \in \mathbb{R}^{d\times d}$ telle que

$$G_T = \int_0^T e^{(T-s)A} BB^* e^{(T-s)A^*} \, \mathrm{d}s. \tag{5.6}$$

Il est clair que la matrice G_T est symétrique, et on vérifie facilement qu'elle est semi-définie positive car $y^* G_T y = \int_0^T |B^* e^{(T-s)A^*} y|^2_{\mathbb{R}^k} \, \mathrm{d}s \geq 0$ pour tout vecteur $y \in \mathbb{R}^d$.

Lemme 5.1.12 (Reformulation du critère de Kalman) *Le système linéaire autonome $\dot{x}(t) = Ax(t) + Bu(t)$ est contrôlable pour tout $T > 0$ et pour tout $x_0 \in \mathbb{R}^d$* ***si et seulement si*** *la matrice G_T, définie par (5.6), est inversible.*

Démonstration. (1) Soit $x_1 \in \mathbb{R}^d$. Supposons la matrice G_T inversible et posons

$$\overline{u}(t) = B^* e^{(T-s)A^*} y \quad \text{où} \quad y = G_T^{-1}(x_1 - e^{TA}x_0).$$

Par la formule de Duhamel, on voit que

$$x_{\overline{u}}(T) = e^{TA}x_0 + \int_0^T e^{(T-s)A} B\overline{u}(s)\,\mathrm{d}s = e^{TA}x_0 + G_T y = x_1.$$

Ceci montre que le système est contrôlable.
(2) Supposons qu'il existe $\Psi \in \mathbb{R}^d$, $\Psi \neq 0$, dans $\mathrm{Ker}(G_T)$. Il vient

$$0 = \Psi^* G_T \Psi = \int_0^T |B^* e^{(T-s)A^*}\Psi|^2_{\mathbb{R}^k}\,\mathrm{d}s,$$

si bien que $\Psi^* e^{(T-s)A}B = 0$ pour tout $s \in [0,T]$. Par la formule de Duhamel, on obtient $\Psi^*(x_u(T) - e^{TA}x_0) = 0$, ce qui montre que $x_u(T)$ est dans un hyperplan affine. Par conséquent, le système n'est pas contrôlable. □

Remarque 5.1.13 Le critère de Kalman $\mathrm{rang}(C) = d$ étant indépendant de T, on en déduit que l'inversibilité de la matrice G_T est donc, elle aussi, indépendante de T. Dans le cas des systèmes de contrôle linéaires autonomes, le critère de Kalman est plus simple à vérifier que l'inversibilité de G_T. Toutefois, la matrice G_T nous sera utile dans le Chapitre 6 lorsque nous étudierons la synthèse d'un contrôle optimal pour la minimisation d'un critère quadratique.

Concluons cette section par une extension du critère de Kalman au cas de la contrôlabilité des systèmes linéaires instationnaires, i.e., de la forme

$$\dot{x}_u(t) = A(t)x_u(t) + B(t)u(t), \quad \forall t \in [0,T], \qquad x_u(0) = x_0, \tag{5.7}$$

avec $A \in L^1([0,T];\mathbb{R}^{d\times d})$ et $B \in L^1([0,T];\mathbb{R}^{d\times k})$. Pour de tels systèmes, la formule de Duhamel n'est plus valable. On utilise la notion de **résolvante** $R : [0,T] \to \mathbb{R}^{d\times d}$ définie comme l'unique solution de

$$\dot{R}(t) = A(t)R(t), \quad R(0) = I,$$

où I est la matrice identité de $\mathbb{R}^{d\times d}$. On notera que

$$\begin{array}{ll} A \in L^1([0,T];\mathbb{R}^{d\times d}) & \Longrightarrow R \in AC([0,T];\mathbb{R}^{d\times d}), \\ A \in C^0([0,T];\mathbb{R}^{d\times d}) & \Longrightarrow R \in C^1([0,T];\mathbb{R}^{d\times d}). \end{array}$$

Comme $\frac{d}{dt}\det(R(t)) = \mathrm{tr}(A(t))\det(R(t))$ et $\det(R(0)) = 1$, la matrice $R(t)$ est inversible à tout temps (la quantité $\det(R(t))$ s'appelle le Wronskien au temps t). On notera également que, dans le cas autonome où $A(t) = A$, on a $R(t) = e^{tA}$. On vérifie sans peine que la solution du système différentiel instationnaire (5.7) est

$$x_u(t) = R(t)x_0 + R(t)\int_0^t R(s)^{-1}B(s)u(s)\,\mathrm{d}s, \quad \forall t \in [0,T].$$

Lemme 5.1.14 (Critère de contrôlabilité, cas instationnaire) *Le système instationnaire* (5.7) *est contrôlable en temps T à partir de x_0* ***si et seulement si*** *la matrice de contrôlabilité*

$$K_T := \int_0^T R(s)^{-1}B(s)B(s)^*(R(s)^{-1})^* \,\mathrm{d}s \in \mathbb{R}^{d\times d} \tag{5.8}$$

est inversible.

Démonstration. Identique au cas autonome. □

Remarque 5.1.15 La condition (5.8) dépend de T, mais pas de x_0. Ainsi, la contrôlabilité en temps T à partir de x_0 implique la contrôlabilité en temps T à partir de tout point ; en revanche, on ne peut s'affranchir de la dépendance en T. On notera également que dans le cas autonome, on a $R(s) = e^{sA}$ et $B(s) = B$, si bien que

$$K_T = e^{-TA}\left(\int_0^T e^{(T-s)A}BB^*e^{(T-s)A^*}\,\mathrm{d}s\right)e^{-TA^*} = e^{-TA}G_Te^{-TA^*}$$

On retrouve donc le critère du Lemme 5.1.12 sur la matrice G_T.

Contre-exemple 5.1.16 [Non-contrôlabilité] On considère le système de contrôle linéaire instationnaire ($d = 2$ et $k = 1$)

$$\dot{x}_u(t) = \begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}x_u(t) + \begin{pmatrix}\cos(t)\\ \sin(t)\end{pmatrix}u(t). \tag{5.9}$$

On vérifie facilement que $R(s) = e^{sA} = \begin{pmatrix}\cos(s) & -\sin(s)\\ \sin(s) & \cos(s)\end{pmatrix}$, d'où

$$R(s)^{-1}B(s) = \begin{pmatrix}1\\ 0\end{pmatrix} \implies K_T = \begin{pmatrix}T & 0\\ 0 & 0\end{pmatrix}.$$

La matrice K_T n'est donc pas inversible, si bien que le système (5.9) n'est pas contrôlable. Le problème vient du fait que la matrice $R(s)^{-1}B(s)$ est indépendante de s. En revanche, si le vecteur B était constant (et non-nul), le système serait contrôlable car B et AB seraient alors des vecteurs orthogonaux non-nuls, si bien que la matrice de Kalman $C = (B, AB)$ serait de rang plein.

5.1.3 Cas avec contraintes : ensemble atteignable

On considère toujours le système de contrôle linéaire autonome (5.2) mais désormais on suppose que le contrôle u est à valeurs dans un sous-ensemble **compact non-vide**

$$U \subset \mathbb{R}^k. \tag{5.10}$$

En particulier, le contrôle $u(t)$ est borné pour tout $t \in [0,T]$. On a donc $u \in L^\infty([0,T];U)$. (On notera que $L^1([0,T];U) = L^\infty([0,T];U)$ lorsque l'ensemble U est borné.) Les résultats de cette section s'étendent au cas instationnaire avec terme de dérive, mais pour simplifier, nous ne traiterons pas ce cas plus général.

Définition 5.1.17 (Ensemble atteignable) *Pour tout $t \in [0,T]$ et tout $x_0 \in \mathbb{R}^d$, l'**ensemble atteignable** en temps t à partir de x_0 est défini comme suit :*

$$\mathcal{A}(t, x_0) = \{x_1 \in \mathbb{R}^d \mid \exists u \in L^\infty([0,t]; U) \text{ tel que } x_u(t) = x_1\}. \tag{5.11}$$

Théorème 5.1.18 (Propriétés de l'ensemble atteignable) *Pour tout $t \in [0,T]$, l'ensemble atteignable $\mathcal{A}(t, x_0)$ est **compact**, **convexe**, et varie **continûment** en t. La continuité en temps est uniforme, i.e., pour tout $\epsilon > 0$, il existe $\delta > 0$ tel que*

$$\forall t_1, t_2 \in [0,T], \quad |t_1 - t_2| \leq \delta \implies d(\mathcal{A}(t_1, x_0), \mathcal{A}(t_2, x_0)) \leq \epsilon, \tag{5.12}$$

où la distance de Hausdorff entre deux sous-ensembles $\mathcal{A}_1$ et $\mathcal{A}_2$ de $\mathbb{R}^d$ est définie comme suit (cf. la figure 5.1) :

$$\begin{aligned} d(\mathcal{A}_1, \mathcal{A}_2) &:= \max\left(\sup_{x_1 \in \mathcal{A}_1} d(x_1, \mathcal{A}_2), \sup_{x_2 \in \mathcal{A}_2} d(x_2, \mathcal{A}_1) \right) \\ &= \max\left(\sup_{x_1 \in \mathcal{A}_1} \inf_{y_2 \in \mathcal{A}_2} |x_1 - y_2|_{\mathbb{R}^d}, \sup_{x_2 \in \mathcal{A}_2} \inf_{y_1 \in \mathcal{A}_1} |x_2 - y_1|_{\mathbb{R}^d} \right). \end{aligned} \tag{5.13}$$

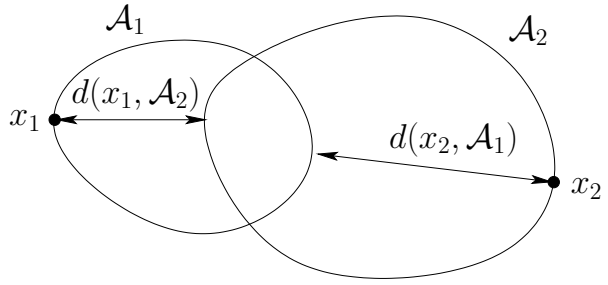

FIGURE 5.1 – Distance de Hausdorff entre deux sous-ensembles $\mathcal{A}_1$ et $\mathcal{A}_2$ de $\mathbb{R}^d$.

Démonstration. Nous verrons les preuves de variation continue en temps et de compacité dans la Section 5.2 dans le cas plus général des systèmes de contrôle non-linéaires. Nous nous contentons ici de prouver la convexité de l'ensemble atteignable $\mathcal{A}(t, x_0)$, propriété qui est, quant à elle, spécifique au cas linéaire.
(1) Cas où le sous-ensemble U est convexe. Dans ce cas, la preuve de convexité de l'ensemble atteignable $\mathcal{A}(t, x_0)$ est élémentaire. Soit $x_1, x_2 \in \mathcal{A}(t, x_0)$, soit $\theta \in [0, 1]$ et montrons que $\theta x_1 + (1 - \theta)x_2 \in \mathcal{A}(t, x_0)$. Par définition, il existe des contrôles $u_i \in L^\infty([0,t]; U)$, $i \in \{1, 2\}$, tels que

$$x_i = e^{tA} x_0 + \int_0^t e^{(t-s)A} B u_i(s)\, \mathrm{d}s,$$

où x_i est la trajectoire associée au contrôle u_i, $i \in \{1, 2\}$. Posons $u(s) = \theta u_1(s) + (1 - \theta)u_2(s)$, pour tout $s \in [0,t]$. La fonction u est mesurable et cette fonction est à valeurs dans U grâce à la convexité du sous-ensemble U. De plus, par linéarité, la

trajectoire x_u associée au contrôle u vérifie

$$\begin{aligned} x_u(t) &= e^{tA}x_0 + \int_0^t e^{(t-s)A}Bu(s)\,\mathrm{d}s \\ &= e^{tA}x_0 + \theta\int_0^t e^{(t-s)A}Bu_1(s)\,\mathrm{d}s + (1-\theta)\int_0^t e^{(t-s)A}Bu_2(s)\,\mathrm{d}s \\ &= \theta x_1 + (1-\theta)x_2, \end{aligned}$$

ce qui montre que $\theta x_1 + (1-\theta)x_2 \in \mathcal{A}(t,x_0)$.
(2) Cas général pour U. Dans ce cas, on invoque le Lemme de Lyapunov 5.1.19 rappelé ci-dessous (pour la preuve, voir par exemple la référence [17]). Soit $x_1, x_2 \in \mathcal{A}(t,x_0)$, soit $\theta \in [0,1]$ et montrons à nouveau que $\theta x_1 + (1-\theta)x_2 = x(t) \in \mathcal{A}(t,x_0)$. Par définition, il existe des contrôles $u_i \in L^\infty([0,t];U)$, $i \in \{1,2\}$, tels que $x_i = e^{tA}x_0 + \int_0^t e^{(t-s)A}Bu_i(s)\,\mathrm{d}s$. Posons $y_i = x_i - e^{tA}x_0$ et considérons la fonction $f \in L^1([0,t];\mathbb{R}^{2d})$ telle que

$$f(s) = \begin{pmatrix} e^{(t-s)A}Bu_1(s) \\ e^{(t-s)A}Bu_2(s) \end{pmatrix} \in \mathbb{R}^{2d}.$$

On a $\int_{\{0\}} f(s)\,\mathrm{d}s = (0,0)^*$ et $\int_{[0,t]} f(s)\,\mathrm{d}s = (y_1,y_2)^*$. En invoquant le lemme de Lyapunov, on en déduit qu'il existe un sous-ensemble mesurable $E \subset [0,t]$ tel que

$$\int_E f(s)\,\mathrm{d}s = \begin{pmatrix} \theta y_1 \\ \theta y_2 \end{pmatrix}.$$

En notant E^c le complémentaire de E dans $[0,t]$, on a

$$\int_{E^c} f(s)\,\mathrm{d}s = \int_{[0,t]} f(s)\,\mathrm{d}s - \int_E f(s)\,\mathrm{d}s = \begin{pmatrix} (1-\theta)y_1 \\ (1-\theta)y_2 \end{pmatrix}.$$

Finalement, on pose

$$u(s) = \begin{cases} u_1(s) & \text{si } s \in E, \\ u_2(s) & \text{si } s \in E^c. \end{cases}$$

Le contrôle ainsi défini est bien une fonction mesurable de $[0,t]$ dans U car les ensembles E et E^c sont mesurables. De plus, la trajectoire x_u associée à ce contrôle satisfait

$$\begin{aligned} x_u(t) - e^{tA}x_0 &= \int_{[0,t]} e^{(t-s)A}Bu(s)\,\mathrm{d}s \\ &= \int_E e^{(t-s)A}Bu_1(s)\,\mathrm{d}s + \int_{E^c} e^{(t-s)A}Bu_2(s)\,\mathrm{d}s = \theta y_1 + (1-\theta)y_2, \end{aligned}$$

ce qui montre que $\theta x_1 + (1-\theta)x_2 = x_u(t) \in \mathcal{A}(t,x_0)$. □

Lemme 5.1.19 (Lyapunov) *Soit $t > 0$. Soit une fonction $f \in L^1([0,t];\mathbb{R}^n)$. Alors, le sous-ensemble*

$$\left\{ \int_E f(s)\,\mathrm{d}s \mid E \subset [0,t] \text{ mesurable} \right\} \tag{5.14}$$

est un sous-ensemble ***convexe*** *de $\mathbb{R}^n$.*

Remarque 5.1.20 On peut montrer que l'ensemble atteignable pour des contrôles à valeurs dans U est le même que pour des contrôles à valeurs dans $\mathrm{conv}(U)$ (l'enveloppe convexe de U).

Exemple 5.1.21 [Mouvement d'un point matériel] On considère un point matériel en mouvement rectiligne. On contrôle la vitesse de ce point par un contrôle à valeurs dans l'intervalle borné $U := [-1, 1]$:

$$\dot{x}(t) = u(t),\ \forall t \in [0,T], \quad x(0) = 0, \quad u(t) \in U = [-1, 1],$$

où on a fixé l'origine à la position initiale du point matériel. L'ensemble atteignable est $\mathcal{A}(t, 0) = [-t, t]$ (qui est bien compact, convexe et varie continûment en t). On constate qu'on obtient le même ensemble atteignable en se restreignant à des contrôles à valeurs dans $\partial U = \{-1, 1\}$. De tels contrôles sont appelés des **contrôles bang-bang** car ils ne prennent que des valeurs extrémales dans ∂U. Une illustration est présentée à la Figure 5.2.

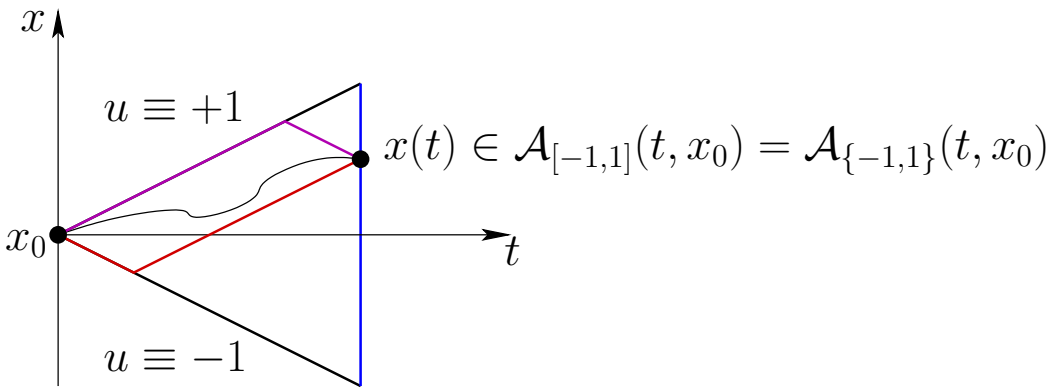

FIGURE 5.2 – Ensemble atteignable par un point matériel dont on contrôle la vitesse dans $U = [-1, 1]$.

5.2 Contrôlabilité des systèmes non-linéaires

Cette section porte sur la contrôlabilité des systèmes de contrôle non-linéaires. Comme à la section précédente, la notion d'**ensemble atteignable** joue un rôle important. Le résultat principal de cette section est un critère de **contrôlabilité locale** au voisinage d'une cible située dans l'ensemble atteignable, ce critère se formulant à l'aide de la contrôlabilité du système linéarisé. Afin d'établir ce résultat, nous montrerons que, sous certaines hypothèses, la différentielle de l'application entrée-sortie (qui à un contrôle associe l'état du système au temps final) est différentiable et que sa différentielle est l'application entrée-sortie du système linéarisé. Dans cette section on utilisera certains résultats sur les équations différentielles ordinaires qui sont rappelés dans l'Annexe 8.

5.2.1 Ensemble atteignable

On fixe un horizon temporel $T > 0$ et une condition initiale $x_0 \in \mathbb{R}^d$. On considère le système de contrôle non-linéaire

$$\dot{x}_u(t) = f(t, x_u(t), u(t)), \quad \forall t \in [0,T], \qquad x_u(0) = x_0. \tag{5.15}$$

Soit $U \subset \mathbb{R}^k$ un sous-ensemble compact non-vide de $\mathbb{R}^k$. La définition de l'ensemble atteignable (en temps $t \in [0,T]$ à partir de x_0) est identique à celle que nous avons introduite dans le cas linéaire (cf. la Définition 5.1.17).

Définition 5.2.1 (Ensemble atteignable) *Pour tout $t \in [0,T]$, l'**ensemble atteignable** en temps t à partir de x_0 est défini comme suit :*

$$\mathcal{A}(t, x_0) = \{x_1 \in \mathbb{R}^d \mid \exists u \in L^\infty([0,t]; U) \text{ tel que } x_u(t) = x_1\}.$$

Nous allons établir deux propriétés importantes et utiles de l'ensemble atteignable : sa variation continue en temps et sa compacité.

Lemme 5.2.2 (Variation continue en temps) *On suppose que*
(i) *f est continue sur $\mathbb{R} \times \mathbb{R}^d \times U$;*
(ii) *U est un sous-ensemble compact non-vide de $\mathbb{R}^k$;*
(iii) *les trajectoires sont uniformément bornées, i.e.,*

$$\exists M > 0, \quad \forall u \in L^\infty([0,T]; U), \quad \sup_{t \in [0,T]} |x_u(t)|_{\mathbb{R}^d} \leq M.$$

Alors, l'ensemble $\mathcal{A}(t, x_0)$ varie continûment en temps, et ce de manière uniforme, i.e., pour tout $\epsilon > 0$, il existe $\delta > 0$ tel que

$$\forall t_1, t_2 \in [0,T], \quad |t_1 - t_2| \leq \delta \implies d(\mathcal{A}(t_1, x_0), \mathcal{A}(t_2, x_0)) \leq \epsilon,$$

où d est la distance de Hausdorff d entre deux sous-ensembles, définie en (5.13) *(cf. la Figure 5.1).*

Remarque 5.2.3 Les hypothèses du Lemme 5.2.2 sont bien vérifiées dans le cas linéaire. L'ensemble atteignable varie donc continûment en temps dans ce cas.

Démonstration. Soit $\epsilon > 0$. On va montrer qu'il existe $\delta > 0$ tel que

$$\forall t_1, t_2 \in [0,T], \quad |t_1 - t_2| \leq \delta \implies d(\mathcal{A}_1, \mathcal{A}_2) \leq \epsilon,$$

où $\mathcal{A}_1 = \mathcal{A}(t_1, x_0)$ et $\mathcal{A}_2 = \mathcal{A}(t_2, x_0)$. Supposons pour fixer les idées que $t_2 > t_1$. Soit $x_2 \in \mathcal{A}_2$. Il existe donc un contrôle $u \in L^\infty([0,t_2]; U)$ tel que

$$x_2 = x_0 + \int_0^{t_2} f(s, x(s), u(s))\, \mathrm{d}s.$$

Avec ce même contrôle, on pose

$$x_1 = x_0 + \int_0^{t_1} f(s, x(s), u(s))\, \mathrm{d}s \in \mathcal{A}(t_1, x_0).$$

D'après les hypothèses sur f, x et u, on a

$$|x_2 - x_1|_{\mathbb{R}^d} \leq \int_{t_1}^{t_2} |f(s, x(s), u(s))|_{\mathbb{R}^d} \, \mathrm{d}s \leq C|t_2 - t_1|.$$

Ceci montre que $d(x_2, \mathcal{A}_1) \leq |x_2 - x_1|_{\mathbb{R}^d} \leq C|t_2 - t_1|$. On raisonne de même pour $x_1 \in \mathcal{A}_1$, ce qui conclut la preuve. □

Lemme 5.2.4 (Compacité) *On suppose que*
(i) *f est continue sur $\mathbb{R} \times \mathbb{R}^d \times U$ et de classe C^1 en x ;*
(ii) *U est un sous-ensemble compact non-vide de $\mathbb{R}^k$;*
(iii) *les trajectoires sont uniformément bornées, i.e.,*

$$\exists M > 0, \quad \forall u \in L^\infty([0,T]; U), \quad \sup_{t \in [0,T]} |x_u(t)|_{\mathbb{R}^d} \leq M;$$

(iv) *pour tout $(t, x) \in [0,T] \times \mathbb{R}^d$, l'ensemble des vecteurs vitesse $K(t, x) := \{f(t, x, u) \mid u \in U\}$ est un sous-ensemble **convexe** de $\mathbb{R}^d$.*
Alors, pour tout $t \in [0,T]$, l'ensemble atteignable $\mathcal{A}(t, x_0)$ est un sous-ensemble compact de $\mathbb{R}^d$.

Remarque 5.2.5 Les hypothèses du Lemme 5.2.4 sont bien vérifiées dans le cas linéaire avec U convexe. L'ensemble atteignable est donc compact dans ce cas.

Démonstration. On se place dans l'espace de Hilbert $V = L^2([0,T]; \mathbb{R}^d)$ et on va montrer la compacité de l'ensemble atteignable $\mathcal{A}(T, x_0)$. La preuve utilise des notions de topologie faible dans les espaces de Hilbert (voir la Sous-section 2.3.3 pour cette notion).
(1) Soit $(y_n)_{n \in \mathbb{N}}$ une suite d'éléments de $\mathcal{A}(T, x_0) \subset \mathbb{R}^d$. Soit $(u_n)_{n \in \mathbb{N}}$ une suite de contrôles dans $L^\infty([0,T]; U)$ et $(x_n)_{n \in \mathbb{N}}$ la suite de trajectoires correspondantes dans $AC([0,T]; \mathbb{R}^d)$ menant de x_0 à y_n. Posons $g_n(s) = f(s, x_n(s), u_n(s)) \, \mathrm{d}s$ pour tout $n \in \mathbb{N}$ et $s \in [0,T]$. On a

$$x_n(t) = x_0 + \int_0^t g_n(s) \, \mathrm{d}s, \ \forall t \in [0,T] \quad \text{et} \quad y_n = x_n(T).$$

D'après les hypothèses, la suite $(g_n)_{n \in \mathbb{N}}$ est bornée dans V. En invoquant le Lemme 2.3.13 sur la compacité faible dans les espaces de Hilbert, on en déduit qu'à une sous-suite près, la suite $(g_n)_{n \in \mathbb{N}}$ converge vers une fonction $g \in V$ pour la topologie faible. On définit la trajectoire $x \in AC([0,T]; \mathbb{R}^d)$ en posant

$$x(t) = x_0 + \int_0^t g(s) \, \mathrm{d}s, \ \forall t \in [0,T].$$

Par convergence faible, on a $\int_0^t g_n(s) \, \mathrm{d}s = (g_n, 1_{[0,t]})_V \to (g, 1_{[0,t]})_V = \int_0^t g(s) \, \mathrm{d}s$, i.e.,

$$\lim_{n \to +\infty} x_n(t) = x(t), \quad \forall t \in [0,T].$$

En particulier, on a donc

$$\lim_{n\to+\infty} y_n = x(T).$$

Il reste à montrer que la trajectoire $x(t)$ peut bien être engendrée par un contrôle $u \in L^\infty([0,T];U)$.
(2) Posons $\theta_n(s) = f(s, x(s), u_n(s))$ et introduisons l'ensemble

$$\Theta = \{\theta \in V \mid \theta(s) \in K(s, x(s)),\ \forall s \in [0,T]\},$$

de sorte que $(\theta_n)_{n\in\mathbb{N}}$ est une suite de Θ. Par hypothèse, $K(s, x(s))$ est un sous-ensemble convexe de $\mathbb{R}^d$ pour tout $s \in [0,T]$. On en déduit que Θ est un sous-ensemble convexe de V. De plus, Θ est fermé dans V car la convergence dans V implique la convergence p.p. d'une sous-suite, et $K(s, x(s))$ est fermé dans $\mathbb{R}^d$. Grâce au Lemme 2.3.15 sur la fermeture faible des convexes dans les espaces de Hilbert, on en déduit que Θ est faiblement fermé dans V. De plus, comme la suite $(\theta_n)_{n\in\mathbb{N}}$ est bornée dans V, on déduit du Lemme 2.3.13 qu'elle converge faiblement, à une sous-suite près, vers une fonction $\theta \in \Theta$. Il existe donc une fonction $u : [0,T] \to U$ telle que $\theta(s) = f(s, x(s), u(s))$ p.p. dans $[0,T]$, et la fonction u peut être choisie mesurable (cf. la Section 8.2 pour plus de précisions sur ce point). Pour tout $\varphi \in V$, on a

$$\int_0^T g_n(s)\varphi(s)\,\mathrm{d}s = \int_0^T \theta_n(s)\varphi(s)\,\mathrm{d}s + \int_0^T (f(s, x_n(s), u_n(s)) - f(s, x(s), u_n(s))\varphi(s)\,\mathrm{d}s. \tag{5.16}$$

Comme $|f(s, x_n(s), u_n(s)) - f(s, x(s), u_n(s))|_{\mathbb{R}^d} \le C|x_n(s) - x(s)|_{\mathbb{R}^d}$ et $|x_n(s) - x(s)|_{\mathbb{R}^d}$ tend vers zéro p.p. dans $[0,T]$, le deuxième terme au membre de droite de (5.16) tend vers zéro (invoquer le théorème de convergence dominée de Lebesgue). En outre, par convergence faible, on a $\int_0^T g(s)\varphi(s)\,\mathrm{d}s = \int_0^T \theta(s)\varphi(s)\,\mathrm{d}s$, i.e., $g(s) = \theta(s)$ p.p. dans $[0,T]$. En conclusion, on a bien $g(s) = f(s, x(s), u(s))$ p.p. sur $[0,T]$. □

5.2.2 Contrôlabilité locale des systèmes non-linéaires

On considère toujours le système de contrôle non-linéaire (5.15). On suppose désormais que la fonction $f(t, x, u)$ est de classe C^1 en (x, u).

Définition 5.2.6 (Application entrée-sortie) *L'**application entrée-sortie** en temps T à partir de x_0 est l'application*

$$E_{T,x_0} : \mathcal{U}_{T,x_0} \to \mathcal{A}(T, x_0), \quad E_{T,x_0}(u) = x_u(T), \tag{5.17}$$

où $\mathcal{U}_{T,x_0} \subset L^\infty([0,T];U)$, U étant un sous-ensemble fermé non-vide de $\mathbb{R}^k$, est le domaine de E_{T,x_0}, i.e., l'ensemble des contrôles tels que la trajectoire associée x_u est bien définie sur $[0,T]$. L'ensemble atteignable $\mathcal{A}(T, x_0)$ est l'image de l'application entrée-sortie E_{T,x_0}.

Soit $y \in \mathcal{A}(T, x_0)$. Par définition, il existe un contrôle $u_y \in \mathcal{U}_{T,x_0}$ amenant l'état de x_0 à y en temps T. Le problème de la contrôlabilité locale consiste à savoir si cette propriété reste satisfaite dans un voisinage du point $y \in \mathcal{A}(T, x_0)$.

Définition 5.2.7 (Contrôlabilité locale) *On dit que le système de contrôle non-linéaire* (5.15) *est contrôlable localement en un point $y \in \mathcal{A}(T, x_0)$ s'il existe un voisinage V_y de y dans $\mathbb{R}^d$ tel que $V_y \subset \mathcal{A}(T, x_0)$, i.e., pour tout $y' \in V_y$, il existe un contrôle $u_{y'} \in \mathcal{U}_{T,x_0}$ amenant l'état de x_0 à y' en temps T.*

Afin d'étudier la contrôlabilité locale du système de contrôle non-linéaire (5.15), nous allons considérer la différentielle (de Fréchet) de l'application entrée-sortie E_{T,x_0}. Pour simplifier, on se place pour le reste de cette section dans le cas **sans contrainte**, i.e., on suppose que $U = \mathbb{R}^k$ si bien que l'on a $\mathcal{U}_{T,x_0} \subset L^\infty([0,T];\mathbb{R}^k)$. Par des arguments de dépendance de la solution d'un système différentiel en des paramètres, on vérifie facilement que $\mathcal{U}_{T,x_0}$ est un sous-ensemble ouvert de $L^\infty([0,T];\mathbb{R}^k)$. On est donc dans la situation où

$$E_{T,x_0} : \mathcal{U}_{T,x_0} \subset L^\infty([0,T];\mathbb{R}^k) \to \mathcal{A}(T, x_0) \subset \mathbb{R}^d.$$

Soit $u \in \mathcal{U}_{T,x_0}$ et $x_u \in AC([0,T];\mathbb{R}^d)$ la trajectoire associée. Soit

$$\delta u \in L^\infty([0,T];\mathbb{R}^k),$$

une perturbation du contrôle ; on suppose cette perturbation suffisamment petite pour que $u+\delta u \in \mathcal{U}_{T,x_0}$ (ceci est possible puisque $\mathcal{U}_{T,x_0}$ est un sous-ensemble ouvert de $L^\infty([0,T];\mathbb{R}^k)$). On considère le système différentiel linéarisé le long de la trajectoire x_u, i.e.,

$$\dot{\delta x}(t) = A_u(t)\delta x(t) + B_u(t)\delta u(t), \quad \forall t \in [0,T], \qquad \delta x(0) = 0, \tag{5.18}$$

où pour tout $t \in [0,T]$,

$$A_u(t) = \frac{\partial f}{\partial x}(t, x_u(t), u(t)) \in \mathbb{R}^{d\times d}, \qquad B_u(t) = \frac{\partial f}{\partial u}(t, x_u(t), u(t)) \in \mathbb{R}^{d\times k}.$$

Lemme 5.2.8 *L'application entrée-sortie E_{T,x_0} est **différentiable** (au sens de Fréchet) en tout $u \in \mathcal{U}_{T,x_0}$ et sa différentielle $E'_{T,x_0}(u) : L^\infty([0,T];\mathbb{R}^k) \to \mathbb{R}^d$ est l'application entrée-sortie du système linéarisé le long de la trajectoire x_u ; plus explicitement, pour tout $\delta u \in L^\infty([0,T];\mathbb{R}^k)$, on a*

$$\langle E'_{T,x_0}(u), \delta u\rangle = \delta x(T),$$

où δx est solution du système différentiel linéarisé (5.18).

Remarque 5.2.9 La différentielle $E'_{T,x_0}(u)$ est bien une forme linéaire continue en δu car on a

$$\langle E'_{T,x_0}(u), \delta u\rangle = R(T)\int_0^T R(s)^{-1}B_u(s)\delta u(s)\,\mathrm{d}s,$$

où $R(t)$ est la résolvante du système linéarisé, i.e., la solution matricielle dans $\mathbb{R}^{d\times d}$ de $\dot{R}(t) = A_u(t)R(t)$, pour tout $t \in [0,T]$, et $R(0) = I_d$. On a donc bien $|\langle E'_{T,x_0}(u), \delta u\rangle| \le C\|\delta u\|_{L^\infty([0,T];\mathbb{R}^k)}$. En outre, $E'_{T,x_0}(u)$ dépend continûment de u.

Démonstration. Nous nous contentons d'esquisser la preuve. Soit $\delta u \in V = L^\infty([0,T];\mathbb{R}^k)$ tel que $u+\delta u \in \mathcal{U}_{T,x_0}$ (qui est ouvert dans V). On note $x_{u+\delta u}$ la trajectoire associée à $u+\delta u$ issue de x_0. En effectuant des développements de Taylor sur f, il vient

$$\begin{aligned}\dot{x}_{u+\delta u}(t) - \dot{x}_u(t) &= f(t, x_{u+\delta u}(t), u(t)+\delta u(t)) - f(t, x_u(t), u(t)) \\ &= \tfrac{\partial f}{\partial x}(t, x_u(t), u(t))(x_{u+\delta u}(t) - x_u(t)) + \tfrac{\partial f}{\partial u}(t, x_u(t), u(t))\delta u(t) + o(\delta u) \\ &= A_u(t)(x_{u+\delta u}(t) - x_u(t)) + B_u(t)\delta u(t) + o(\delta u),\end{aligned}$$

car $x_{u+\delta u} - x_u = O(\delta u)$ (dépendance continue en un paramètre de la solution d'un système différentiel). En posant $\epsilon(t) = x_{u+\delta u}(t) - x_u(t) - \delta x(t)$, on en déduit que $\epsilon(0) = 0$ et que

$$\begin{aligned}\dot{\epsilon}(t) &= \dot{x}_{u+\delta u}(t) - \dot{x}_u(t) - \dot{\delta x}(t) \\ &= A_u(t)(x_{u+\delta u}(t) - x_u(t) - \delta x(t)) + o(\delta u) = A_u(t)\epsilon(t) + o(\delta u).\end{aligned}$$

Par des arguments de stabilité, on montre que $\epsilon = o(\delta u)$. En conclusion, on obtient

$$\begin{aligned}E_{T,x_0}(u+\delta u) - E_{T,x_0}(u) &= x_{u+\delta u}(T) - x_u(T) \\ &= \delta x(T) + \epsilon(T) = \delta x(T) + o(\delta u),\end{aligned}$$

et on a vu que $\delta u \mapsto \delta x(T)$ définit une forme linéaire continue sur δu pour la topologie de V. Ceci conclut la preuve. □

Théorème 5.2.10 (Contrôlabilité locale) *Si le système différentiel linéarisé le long de la trajectoire x_u est **contrôlable** (en temps T), alors le système différentiel non-linéaire est **localement contrôlable** (en temps T à partir de x_0).*

Démonstration. Si le système différentiel linéarisé est contrôlable, alors la différentielle de l'application entrée-sortie E'_{T,x_0} est surjective. On conclut par le théorème de la submersion rappelé ci-dessous (qui est une variante du théorème des fonctions implicites, voir par exemple la référence [19]). □

Théorème 5.2.11 (Submersion) *Soit V et W deux espaces de Banach, et $F : V \to W$ une application continûment différentiable. Soit $v \in V$. Si l'application différentielle $F'(v) : V \to W$ est surjective, alors F est localement surjective au voisinage de $F(v) \in W$.*

Remarque 5.2.12 On considère le cas particulier d'un **point d'équilibre** d'un système différentiel autonome, i.e., un couple (x_0, u_0) tel que $f(x_0, u_0) = 0$. Noter que $x_0 \in \mathcal{A}(t, x_0)$ en utilisant le contrôle constant égal à u_0. Le critère de contrôlabilité locale en x_0 consiste à vérifier que les matrices $A = \frac{\partial f}{\partial x}(x_0, u_0)$ et $B = \frac{\partial f}{\partial u}(x_0, u_0)$ vérifient la condition de Kalman. En effet, comme $f(x_0, u_0) = 0$, la trajectoire de référence est réduite à un point, si bien que le système linéarisé est également autonome, et on peut appliquer la condition de Kalman pour en vérifier la contrôlabilité.

Remarque 5.2.13 [Inversion du temps] En cas de contrôlabilité locale et lorsque la dynamique est autonome et de la forme $f(x,u) = ug(x)$ (en supposant pour simplifier u à valeurs scalaires), on déduit par inversion du temps que pour tout $y \in \mathcal{A}(T,x_0)$ tel que $V_y \subset \mathcal{A}(T,x_0)$, on peut ramener tout point $y' \in V_y$ à x_0. En effet, en notant u' le contrôle amenant x_0 en y' en temps T, on pose $\tilde{u}'(t) = -u'(T-t)$ et on vérifie que $\tilde{x}(t) = x_{u'}(T-t)$ vérifie bien $\tilde{x}(0) = y'$, $\tilde{x}(T) = x_0$ et $\frac{d}{dt}\tilde{x}(t) = -\frac{d}{dt}x_{u'}(T-t) = -u'(T-t)g(x_{u'}(T-t)) = \tilde{u}'(t)g(\tilde{x}(t))$, ce qui montre que $\tilde{x}$ est bien la trajectoire associée au contrôle $\tilde{u}'$.

Exemple 5.2.14 [Pendule inversé] On considère l'exemple du pendule inversé (masse vers le haut, tige vers le bas) avec pour simplifier une masse et une longueur unités ($m = 1$, $l = 1$). On suppose que le pendule a un mouvement dans un plan et on repère l'extrémité supérieure du pendule par son angle θ avec la verticale (dans le sens horaire). On contrôle l'accélération horizontale du point inférieur de la tige. La dynamique s'écrit sous la forme

$$\ddot{\theta}(t) = \sin(\theta(t)) - u(t)\cos(\theta(t)).$$

En posant $x = (x_1, x_2) = (\theta, \dot{\theta}) \in \mathbb{R}^2$, on se ramène à un système d'ordre un :

$$\dot{x}(t) = f(x(t), u(t)), \qquad f(x,u) = \begin{pmatrix} x_2 \\ \sin(x_1) - u\cos(x_1) \end{pmatrix}.$$

On calcule

$$\frac{\partial f}{\partial x}(x,u) = \begin{pmatrix} 0 & 1 \\ \cos(x_1) + u\sin(x_1) & 0 \end{pmatrix}, \qquad \frac{\partial f}{\partial u}(x,u) = \begin{pmatrix} 0 \\ -\cos(x_1) \end{pmatrix}.$$

On considère le point d'équilibre instable $(x_0, u_0) = ((0,0)^*, 0)$. Le système linéarisé autour de ce point s'écrit sous la forme $\dot{\delta x}(t) = A\delta x(t) + B\delta u(t)$ avec

$$A = \frac{\partial f}{\partial x}(x_0,u_0) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad B = \frac{\partial f}{\partial u}(x_0,u_0) = \begin{pmatrix} 0 \\ -1 \end{pmatrix}.$$

La condition de Kalman est bien satisfaite car

$$C = (B, AB) = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}.$$

On a donc montré que le pendule inversé est **localement contrôlable** autour de son point d'équilibre instable $(x_0, u_0) = ((0,0)^*, 0)$. Enfin, en adaptant le raisonnement présenté à la Remarque 5.2.13, on montre qu'on peut ramener tout point au voisinage du point d'équilibre instable vers ce point.

Chapitre 6

LE SYSTÈME LINÉAIRE-QUADRATIQUE

Ce chapitre est consacré à l'étude du système linéaire-quadratique (LQ). Il s'agit d'un problème de contrôle optimal régi par une dynamique linéaire et où le critère à minimiser est quadratique en le contrôle et en la trajectoire associée. Ce problème étant relativement simple, il nous sera possible d'en mener une analyse mathématique complète. D'une part, nous montrerons l'existence et l'unicité du contrôle optimal. D'autre part, cette analyse nous permettra de dégager plusieurs notions importantes pour la suite : l'**état adjoint** pour le calcul de la différentielle du critère, le **Hamiltonien** pour la formulation du contrôle optimal à tout temps comme un minimiseur fonction des valeurs instantanées de l'état adjoint et enfin, celle de **feedback** (ou rétroaction) grâce à l'équation de Riccati afin de formuler le contrôle optimal en **boucle fermée**, c'est-à-dire comme une fonction instantanée de l'état du système.

6.1 Présentation du système LQ

On se donne un intervalle de temps $[0,T]$, avec $T > 0$, une matrice $A \in \mathbb{R}^{d\times d}$ et une matrice $B \in \mathbb{R}^{d\times k}$. On se donne également une condition initiale $x_0 \in \mathbb{R}^d$ et (pour un peu plus de généralité) un terme de dérive $f \in L^1([0,T];\mathbb{R}^d)$. Le système de contrôle linéaire s'écrit sous la forme

$$\dot{x}_u(t) = Ax_u(t) + Bu(t) + f(t), \quad \forall t \in [0,T], \qquad x_u(0) = x_0. \tag{6.1}$$

L'ensemble des contrôles admissibles est ici l'espace

$$V = L^2([0,T];\mathbb{R}^k). \tag{6.2}$$

Pour chaque contrôle $u \in L^2([0,T];\mathbb{R}^k)$, il existe une unique trajectoire $x_u \in AC([0,T];\mathbb{R}^d)$ associée à ce contrôle.

L'objectif de ce chapitre est de chercher un **contrôle optimal** (en fait **le** contrôle optimal, car nous verrons qu'il est unique) qui minimise dans $L^2([0,T];\mathbb{R}^k)$

le critère

$$J(u) = \frac{1}{2}\int_0^T u(t)^* R u(t)\,\mathrm{d}t + \frac{1}{2}\int_0^T e_{x_u}(t)^* Q e_{x_u}(t)\,\mathrm{d}t + \frac{1}{2} e_{x_u}(T)^* D e_{x_u}(T), \quad (6.3)$$

où $e_{x_u} = x_u - \xi$ avec $\xi \in C^0([0,T];\mathbb{R}^d)$ une **trajectoire cible** donnée. On s'intéresse donc au problème suivant :

$$\text{Chercher } \overline{u} \in V \text{ tel que } J(\overline{u}) = \inf_{u\in V} J(u). \quad (6.4)$$

Dans la définition du critère J, les matrices $Q, D \in \mathbb{R}^{d\times d}$ sont symétriques **semi-définies positives**, tandis que la matrice $R \in \mathbb{R}^{k\times k}$ est symétrique **définie positive**. La définie positivité de la matrice R jouera un rôle clé pour assurer l'existence et l'unicité du contrôle optimal minimisant J sur $L^2([0,T];\mathbb{R}^k)$. On notera que le critère J résulte d'une pondération au sens des moindres carrés entre l'atteinte de la trajectoire cible décrite par la fonction ξ et le fait que le contrôle ne soit pas "trop grand" dans $L^2([0,T];\mathbb{R}^k)$. En revanche, on ne s'impose pas ici d'atteindre exactement la cible au temps final T (ni à aucun temps intermédiaire). Une illustration générale du problème de contrôle optimal LQ est présentée à la Figure 6.1.

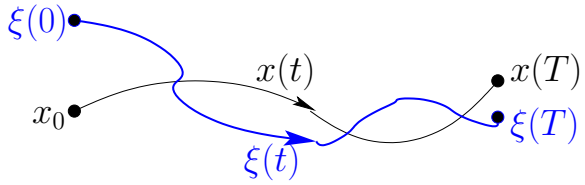

FIGURE 6.1 – Illustration du problème de contrôle optimal LQ : trajectoire cible et trajectoire optimale.

Remarque 6.1.1 On peut prendre $Q = D = 0$ dans le critère (6.3). La solution du problème (6.4) est alors triviale : $u \equiv 0$ sur $[0,T]$.

Afin d'étudier les propriétés de la fonctionnelle J, il sera utile de poser

$$J(u) = J_R(u) + J_{QD}(u), \quad \forall u \in V,$$

avec

$$\begin{aligned} J_R(u) &= \tfrac{1}{2}\int_0^T u(t)^* R u(t)\,\mathrm{d}t, \\ J_{QD}(u) &= \tfrac{1}{2}\int_0^T e_{x_u}(t)^* Q e_{x_u}(t)\,\mathrm{d}t + \tfrac{1}{2} e_{x_u}(T)^* D e_{x_u}(T). \end{aligned}$$

Lemme 6.1.2 *La fonctionnelle J définie en* (6.3) *est fortement convexe et continue sur l'espace de Hilbert $V = L^2([0,T];\mathbb{R}^k)$.*

Démonstration. Comme la matrice R est symétrique définie positive, la fonctionnelle J_R est fortement convexe sur V de paramètre $\alpha = \lambda_{\min}(R)$ (la plus petite valeur

propre de la matrice R). En effet, pour deux vecteurs $v_1, v_2 \in \mathbb{R}^k$, on a

$$\left(\frac{v_1+v_2}{2}\right)^* R\left(\frac{v_1+v_2}{2}\right) = \frac{v_1^* R v_1 + v_2^* R v_2}{2} - \frac{1}{4}(v_1-v_2)^* R (v_1-v_2)$$
$$\leq \frac{v_1^* R v_1 + v_2^* R v_2}{2} - \frac{1}{4}\lambda_{\min}(R)|v_1-v_2|^2_{\mathbb{R}^k}.$$

On en déduit que pour deux contrôles $u_1, u_2 \in V$, on a

$$J_R\left(\frac{u_1+u_2}{2}\right) \leq \frac{J_R(u_1)+J_R(u_2)}{2} - \frac{1}{8}\lambda_{\min}(R)\|u_1-u_2\|_V^2,$$

ce qui prouve la forte convexité de la fonctionnelle J_R sur V avec paramètre $\alpha = \lambda_{\min}(R)$. De plus, la fonctionnelle J_R est clairement continue en u. Par ailleurs, la fonctionnelle J_{QD} est convexe sur V comme composée d'une application convexe par une application affine. En effet,

— comme $x_u(t) = e^{tA}x_0 + \int_0^t e^{(t-s)A}(Bu(s)+f(s))\,\mathrm{d}s$, l'application qui à $u \in L^2([0,T];\mathbb{R}^k)$ associe $e_{x_u} = x_u - \xi \in C^0([0,T];\mathbb{R}^d)$ est affine ;

— comme les matrices Q et D sont symétriques semi-définies positives, on montre facilement que l'application qui à $y \in C^0([0,T];\mathbb{R}^d)$ associe $\frac{1}{2}\int_0^T y(t)^* Q y(t)\,\mathrm{d}t + \frac{1}{2}y(T)^* D y(T) \in \mathbb{R}$ est convexe (même raisonnement que ci-dessus).

La fonctionnelle J_{QD} est en outre continue comme composée de deux applications continues. En conclusion, la fonctionnelle J est fortement convexe sur V comme somme d'une application fortement convexe (J_R) et d'une application convexe (J_{QD}), et J est également continue comme somme de deux applications continues. □

Corollaire 6.1.3 *Il existe un unique contrôle optimal $\overline{u} \in V$ solution de* (6.4).

Démonstration. Il suffit de combiner le Théorème 2.3.8 (avec $K = V$) avec le Lemme 6.1.2. □

6.2 Différentielle du critère : état adjoint

L'objectif de cette section est d'établir une condition nécessaire et suffisante d'optimalité formulée à l'aide de la différentielle de la fonctionnelle J, en utilisant les résultats de la Section 2.5.

Lemme 6.2.1 *La fonctionnelle J est différentiable sur V et on a, pour tout $u \in V$,*

$$\nabla J(u) = Ru + B^* p \ (\in V),$$

*où l'**état adjoint** $p \in C^1([0,T];\mathbb{R}^d)$ est l'unique solution de l'équation différentielle rétrograde en temps*

$$\dot{p}(t) = -A^* p(t) - Q e_{x_u}(t), \quad \forall t \in [0,T], \qquad p(T) = D e_{x_u}(T). \tag{6.5}$$

Démonstration. Comme $J = J_R + J_{QD}$, nous allons considérer séparément la différentiabilité des fonctionnelles J_R et J_{QD}.
(1) La différentiabilité de J_R est immédiate puisque, en utilisant la symétrie de la matrice R, il vient, pour toute perturbation du contrôle $\delta u \in V$,

$$\begin{aligned} J_R(u+\delta u) &= \frac{1}{2}\int_0^T (u(t)+\delta u(t))^* R(u(t)+\delta u(t))\,\mathrm{d}t \\ &= J_R(u) + \int_0^T \delta u(t)^* R u(t)\,\mathrm{d}t + J_R(\delta u) \\ &= J_R(u) + (Ru, \delta u)_V + J_R(\delta u). \end{aligned}$$

Comme $\frac{J_R(\delta u)}{\|\delta u\|_V} \leq \frac{1}{2}\lambda_{\max}(R)\|\delta u\|_V$, on conclut que $\nabla J_R(u) = Ru \in V$, ce qui signifie que p.p. sur $[0,T]$, on a $(\nabla J_R(u))(t) = Ru(t)$.
(2) Pour différentier J_{QD}, on considère la trajectoire perturbée $x_{u+\delta u}$, associée au contrôle perturbé $u+\delta u$. Par linéarité, on a $x_{u+\delta u} = x_u + \delta x$ avec

$$\frac{d}{dt}\delta x(t) = A\delta x(t) + B\delta u(t), \quad \forall t \in [0,T], \qquad \delta x(0) = 0.$$

La perturbation de la trajectoire δx est donc linéaire en δu et on a $\|\delta x\|_{C^0([0,T];\mathbb{R}^d)} \leq C\|\delta u\|_V$ car $\delta x(t) = \int_0^t e^{(t-s)A}B\delta u(s)\,\mathrm{d}s$, où C est une constante dépendant de A, B et T mais qui est uniforme en δu. Comme les matrices Q et D sont symétriques, et en raisonnant comme ci-dessus, on obtient

$$\begin{aligned} J_{QD}(u+\delta u) = J_{QD}(u) &+ \int_0^T \delta x(t)^* Q e_{x_u}(t)\,\mathrm{d}t + \delta x(T)^* D e_{x_u}(T) \\ &+ \frac{1}{2}\int_0^T \delta x(t)^* Q \delta x(t)\,\mathrm{d}t + \frac{1}{2}\delta x(T)^* D \delta x(T), \end{aligned}$$

ce qui montre que

$$(\nabla J_{QD}(u), \delta u)_V = \int_0^T \delta x(t)^* Q e_{x_u}(t)\,\mathrm{d}t + \delta x(T)^* D e_{x_u}(T).$$

Au membre de droite, la perturbation du contrôle δu n'apparaît pas explicitement, mais uniquement de manière implicite par le fait que la perturbation de la trajectoire δx dépend (linéairement) de la perturbation du contrôle δu. Afin de faire apparaître explicitement δu au membre de droite, on utilise l'état adjoint $p \in C^1([0,T];\mathbb{R}^d)$ solution de (6.5). En effet, en intégrant par parties en temps, on constate que

$$\begin{aligned} (\nabla J_{QD}(u), \delta u)_V &= \int_0^T \delta x(t)^* Q e_x(t)\,\mathrm{d}t + \delta x(T)^* D e_x(T) \\ &= -\int_0^T \delta x(t)^*(\dot{p}(t) + A^* p(t))\,\mathrm{d}t + \delta x(T)^* p(T) \\ &= \int_0^T \big(\dot{\delta x}(t)^* p(t) - \delta x(t)^* A^* p(t)\big)\,\mathrm{d}t \\ &= \int_0^T (B\delta u(t))^* p(t)\,\mathrm{d}t = \int_0^T \delta u(t)^* B^* p(t)\,\mathrm{d}t = (B^*p, \delta u)_V. \end{aligned}$$

En conclusion, on a montré que $\nabla J_{QD}(u) = B^*p$, ce qui conclut la preuve. □

Théorème 6.2.2 (CNS d'optimalité) *Le contrôle $\overline{u} \in V$ est optimal pour le problème LQ* ***si et seulement si*** *on a*

$$\overline{u}(t) = -R^{-1}B^*\overline{p}(t) \quad \forall t \in [0,T], \tag{6.6}$$

où l'état adjoint $\overline{p} : [0,T] \to \mathbb{R}^d$ est tel que

$$\frac{d\overline{p}}{dt}(t) = -A^*\overline{p}(t) - Qe_{\overline{x}}(t), \quad \forall t \in [0,T], \qquad \overline{p}(T) = De_{\overline{x}}(T), \tag{6.7}$$

où $e_{\overline{x}} = \overline{x} - \xi$ et où $\overline{x} = x_{\overline{u}}$ est la trajectoire associée au contrôle optimal $\overline{u}$, i.e.,

$$\frac{d\overline{x}}{dt}(t) = A\overline{x}(t) + B\overline{u}(t) + f(t), \quad \forall t \in [0,T], \qquad \overline{x}(0) = x_0. \tag{6.8}$$

Le triplet $(\overline{x}, \overline{p}, \overline{u})$ satisfaisant les conditions ci-dessus est appelé une ***extrémale****.*

Démonstration. Il suffit de combiner le Théorème 2.5.1 avec le Lemme 6.2.1, le caractère suffisant de la condition (6.6) résultant de la convexité de la fonctionnelle J. □

Remarque 6.2.3 [État adjoint] Attention, il n'y a pas de condition initiale sur $\overline{p}$, mais une condition finale en T. Par ailleurs, dans la littérature, la convention est parfois de définir l'état adjoint comme un vecteur ligne $\widehat{p} := \overline{p}^*$. Dans ce cas, le système différentiel rétrograde s'écrit $\frac{d}{dt}\widehat{p}(t) = -\widehat{p}(t)A - e_{\overline{x}}(t)^*Q$, pour tout $t \in [0,T]$, et $\widehat{p}(T) = e_{\overline{x}}(T)^*D$. Enfin, le contrôle optimal est $u(t) = -R^{-1}B^*\widehat{p}(t)^*$.

Remarque 6.2.4 [Régularité] On notera que si $(\overline{x}, \overline{p}, \overline{u})$ est une extrémale, on a $\overline{p} \in C^1([0,T];\mathbb{R}^d)$ et par conséquent $\overline{u} \in C^1([0,T];\mathbb{R}^k)$. Il n'y a pas ici de phénomène de commutation pour le contrôle optimal.

Remarque 6.2.5 [Unicité de l'extrémale] Même si on sait déjà qu'on a unicité du contrôle optimal $\overline{u}$, donc de la trajectoire optimale $\overline{x}$ et de la trajectoire adjointe $\overline{p}$, il est instructif de montrer directement l'unicité de l'extrémale. Par linéarité (considérer la différence entre deux extrémales), il suffit de montrer que dans le cas sans dérive et avec cible nulle, une extrémale est nécessairement nulle. Considérons donc une extrémale $(\overline{x}, \overline{p}, \overline{u})$ telle que

$$\begin{aligned}
&\frac{d\overline{x}}{dt}(t) = A\overline{x}(t) + B\overline{u}(t), \qquad \overline{x}(0) = 0,\\
&\frac{d\overline{p}}{dt}(t) = -A^*\overline{p}(t) - Q\overline{x}(t), \quad \overline{p}(T) = D\overline{x}(T),\\
&\overline{u}(t) = -R^{-1}B^*\overline{p}(t).
\end{aligned}$$

L'observation cruciale est que

$$\begin{aligned}
\frac{d}{dt}\big(\overline{p}(t)^*\overline{x}(t)\big) &= \left(\frac{d\overline{p}}{dt}(t)\right)^*\overline{x}(t) + \overline{p}(t)^*\frac{d\overline{x}}{dt}(t)\\
&= -\overline{x}(t)^*Q\overline{x}(t) - (B^*\overline{p}(t))^*R^{-1}B^*\overline{p}(t) \le 0.
\end{aligned}$$

Comme $\overline{x}(0) = 0$, en intégrant de 0 à T, il vient

$$\begin{aligned} 0 &= \overline{p}(T)^* \overline{x}(T) - \int_0^T \frac{d}{dt}\big(\overline{p}(t)^* \overline{x}(t)\big)\,\mathrm{d}t \\ &= \overline{x}(T)^* D\overline{x}(T) + \int_0^T \Big\{ \overline{x}(t)^* Q \overline{x}(t) + (B^*\overline{p}(t))^* R^{-1} B^* \overline{p}(t) \Big\}\,\mathrm{d}t. \end{aligned}$$

Comme les matrices D et Q sont positives et que la matrice R est définie positive, on en déduit que $B^*\overline{p}(t) = 0$ sur $[0,T]$. Donc, $\overline{u}(t) = 0$, ce qui implique que $\overline{x}(t) = 0$, et ce qui implique enfin que $\overline{p}(t) = 0$.

Exemple 6.2.6 [Mouvement d'un point matériel] On considère un point matériel qui peut se déplacer sur une droite et dont on contrôle la vitesse (cf. l'Exemple 5.1.21). Le système de contrôle linéaire s'écrit, avec $d = k = 1$,

$$\dot{x}_u(t) = u(t), \quad \forall t \in [0,T], \qquad x_u(0) = x_0.$$

Le critère à minimiser dans $V = L^2([0,T];\mathbb{R})$ est

$$J(u) = \frac{1}{2}\int_0^T x_u(t)^2\,\mathrm{d}t + \frac{1}{2}\int_0^T u(t)^2\,\mathrm{d}t,$$

qui réalise une pondération au sens des moindres carrés entre l'atteinte de la cible nulle sur $[0,T]$ et le fait que le contrôle ne soit pas trop grand dans $L^2([0,T];\mathbb{R})$. Ce problème rentre dans le cadre du système LQ introduit à la section 6.1 en posant

$$A = 0, \quad B = 1, \quad R = 1, \quad Q = 1, \quad D = 0, \quad \xi \equiv 0.$$

En appliquant le Théorème 6.2.2, on déduit que le contrôle optimal est

$$\overline{u}(t) = -\overline{p}(t),$$

où l'état adjoint est solution de

$$\frac{d\overline{p}}{dt}(t) = -\overline{x}(t), \quad \forall t \in [0,T], \qquad \overline{p}(T) = 0.$$

On a donc

$$\frac{d}{dt}\begin{pmatrix}\overline{x}(t)\\ \overline{p}(t)\end{pmatrix} = \underbrace{\begin{pmatrix}0 & -1\\ -1 & 0\end{pmatrix}}_{=Z}\begin{pmatrix}\overline{x}(t)\\ \overline{p}(t)\end{pmatrix}, \qquad e^{tZ} = \begin{pmatrix}\cosh(t) & -\sinh(t)\\ -\sinh(t) & \cosh(t)\end{pmatrix},$$

si bien que

$$\begin{aligned} \overline{x}(t) &= x_0\cosh(t) - \overline{p}(0)\sinh(t), \\ \overline{p}(t) &= -x_0\sinh(t) + \overline{p}(0)\cosh(t). \end{aligned}$$

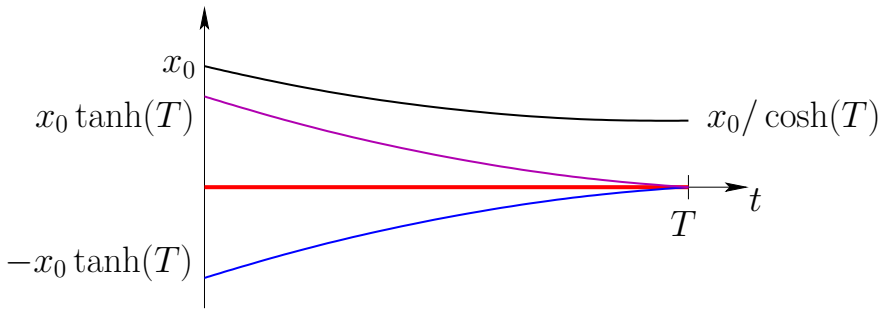

FIGURE 6.2 – Illustration de l'extrémale obtenue à l'Exemple 6.2.6 (mouvement d'un point matériel) : trajectoire $\overline{x}(t)$, état adjoint $\overline{p}(t)$, contrôle optimal $\overline{u}(t)$; la cible $\xi(t)$ est identiquement nulle.

On notera que l'état adjoint initial est, à ce stade, encore inconnu. Afin de le déterminer, on utilise la condition en $t = T$ sur l'état adjoint, à savoir $\overline{p}(T) = 0$. On obtient facilement que $\overline{p}(0) = x_0 \tanh(T)$. En conclusion, l'extrémale s'écrit

$$\begin{aligned}
\overline{x}(t) &= x_0 \tfrac{1}{\cosh(T)} \cosh(T-t), \\
\overline{p}(t) &= x_0 \tfrac{1}{\cosh(T)} \sinh(T-t), \\
\overline{u}(t) &= -\overline{p}(t) = -x_0 \tfrac{1}{\cosh(T)} \sinh(T-t).
\end{aligned}$$

Cette extrémale est illustrée à la Figure 6.2.

6.3 Principe du minimum : Hamiltonien

L'objectif de cette section est de reformuler le Théorème 6.2.2 à l'aide de la notion de Hamiltonien. Ce point de vue nous sera très utile au chapitre suivant lorsque nous aborderons les systèmes de contrôle non-linéaires et formulerons le principe du minimum de Pontryaguine.

Définition 6.3.1 *Le* ***Hamiltonien*** *associé au système de contrôle linéaire* (6.1) *et à la fonctionnelle J définie en* (6.3) *est l'application $H : [0,T] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$ telle que*

$$H(t,x,p,u) = p^*(Ax + Bu + f(t)) + \frac{1}{2}u^*Ru + \frac{1}{2}(x - \xi(t))^*Q(x - \xi(t)).$$

On notera bien que dans cette écriture, (x,p,u) désigne un vecteur générique de $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^k$ et pas les solutions d'équations différentielles.

Un calcul élémentaire sur les dérivées partielles du Hamiltonien (qui sont ici identifiées à des vecteurs colonnes) montre que

$$\begin{aligned}
\nabla_x H(t,x,p,u) &= A^*p + Q(x - \xi(t)), \\
\nabla_p H(t,x,p,u) &= Ax + Bu + f(t), \\
\nabla_u H(t,x,p,u) &= B^*p + Ru.
\end{aligned}$$

On considère maintenant l'extrémale $(\overline{x}, \overline{p}, \overline{u})$ obtenue au Théorème 6.2.2. Pour tout $t \in [0,T]$, on évalue H et ses dérivées partielles en $(t, \overline{x}(t), \overline{p}(t), \overline{u}(t))$. On constate d'une part que

$$\frac{d\overline{x}}{dt}(t) = A\overline{x}(t) + B\overline{u}(t) + f(t) = \nabla_p H(t, \overline{x}(t), \overline{p}(t), \overline{u}(t)), \tag{6.9a}$$

$$\frac{d\overline{p}}{dt}(t) = -A^*\overline{p}(t) - Q(\overline{x}(t) - \xi(t)) = -\nabla_x H(t, \overline{x}(t), \overline{p}(t), \overline{u}(t)), \tag{6.9b}$$

et d'autre part que

$$\nabla_u H(t, \overline{x}(t), \overline{p}(t), \overline{u}(t)) = 0. \tag{6.10}$$

Comme la fonction $v \mapsto H(t, x, p, v)$ est fortement convexe en $v \in \mathbb{R}^k$ pour tout triplet (t, x, p) fixé dans $[0,T] \times \mathbb{R}^d \times \mathbb{R}^d$, l'équation (6.10) ne signifie rien d'autre que

$$\overline{u}(t) = \arg\min_{v \in \mathbb{R}^k} H(t, \overline{x}(t), \overline{p}(t), v), \quad \forall t \in [0,T].$$

Il s'agit du **principe du minimum de Pontryaguine (PMP)** dans le cas particulier du système LQ. Résumons ce résultat sous la forme d'une proposition.

Proposition 6.3.2 (PMP pour le système LQ) *Le contrôle $\overline{u} \in V$ est optimal pour le problème LQ* ***si et seulement si*** *on a*

$$\overline{u}(t) = \arg\min_{v \in \mathbb{R}^k} H(t, \overline{x}(t), \overline{p}(t), v), \quad \forall t \in [0,T],$$

avec

$$\frac{d\overline{x}}{dt}(t) = \nabla_p H(t, \overline{x}(t), \overline{p}(t), \overline{u}(t)) = A\overline{x}(t) + B\overline{u}(t), \qquad \overline{x}(0) = x_0, \tag{6.11a}$$

$$\frac{d\overline{p}}{dt}(t) = -\nabla_x H(t, \overline{x}(t), \overline{p}(t), \overline{u}(t)) = -A^*\overline{p}(t) - Qe_{\overline{x}}(t), \quad \overline{p}(T) = De_{\overline{x}}(T), \tag{6.11b}$$

où $e_{\overline{x}}(t) = \overline{x}(t) - \xi(t)$.

Remarque 6.3.3 [Convention de signe] On aurait pu définir $\widehat{H} := -H$ et aboutir à un principe du maximum pour $\widehat{H}$.

Dans le cas particulier avec dérive et cible nulles, i.e., lorsque $f \equiv 0$ et $\xi \equiv 0$ sur $[0,T]$, le Hamiltonien H ne dépend pas du temps, i.e., on a

$$\frac{\partial H}{\partial t}(t, x, p, u) = 0.$$

Dans ce cas onn dit que le Hamiltonien est **autonome**.

Proposition 6.3.4 (Conservation du Hamiltonien le long de l'extrémale) *On suppose que dérive et cible sont nulles, i.e., que le Hamiltonien est autonome. Alors, la valeur du Hamiltonien se conserve le long de l'extrémale $(\overline{x}, \overline{p}, \overline{u})$.*

Démonstration. On considère l'application $\mathcal{H} : [0,T] \to \mathbb{R}$ telle que

$$\mathcal{H}(t) = H(\overline{x}(t), \overline{p}(t), \overline{u}(t)), \quad \forall t \in [0,T].$$

En dérivant cette fonction par rapport au temps, il vient

$$\begin{aligned}\frac{d\mathcal{H}}{dt}(t) &= \big(\nabla_x H\big)^* \frac{d\overline{x}}{dt}(t) + \big(\nabla_p H\big)^* \frac{d\overline{p}}{dt}(t) + \big(\nabla_u H\big)^* \frac{d\overline{u}}{dt}(t) \\ &= -\frac{d\overline{p}}{dt}(t)^* \frac{d\overline{x}}{dt}(t) + \frac{d\overline{x}}{dt}(t)^* \frac{d\overline{p}}{dt}(t) + 0 = 0,\end{aligned}$$

ce qui conclut la preuve. □

Exemple 6.3.5 On reprend l'Exemple 6.2.6 du mouvement d'un point matériel le long d'une droite et dont on contrôle la vitesse, i.e., $\dot{x}_u(t) = u(t)$, pour tout $t \in [0,T]$, et $x_u(0) = x_0$. Le critère à minimiser est à nouveau $J(u) = \frac{1}{2}\int_0^T x_u(t)^2\,\mathrm{d}t + \frac{1}{2}\int_0^T u(t)^2\,\mathrm{d}t$. Ce problème rentre dans le cadre du système LQ avec $d = k = 1$ et

$$A = 0, \quad B = 1, \quad R = 1, \quad Q = 1, \quad D = 0, \quad \xi \equiv 0.$$

Le Hamiltonien est l'application de $\mathbb{R} \times \mathbb{R} \times \mathbb{R}$ dans $\mathbb{R}$ telle que

$$H(x,p,u) = pu + \frac{1}{2}u^2 + \frac{1}{2}x^2.$$

À (x,p) fixés, l'application $u \mapsto H(x,p,u)$ est quadratique. Le principe du minimum de Pontryaguine (cf. la Proposition 6.3.2) implique que le contrôle optimal $\overline{u}(t)$ est, pour tout $t \in [0,T]$, le minimiseur de $u \mapsto H(\overline{x}(t), \overline{p}(t), u)$ sur $\mathbb{R}$. En utilisant l'expression de H, on obtient facilement

$$\overline{u}(t) = -\overline{p}(t).$$

On retrouve ainsi le même résultat que celui obtenu en considérant la différentielle de J. De plus, si on évalue le Hamiltonien le long de l'extrémale, il vient

$$\mathcal{H}(t) = H(\overline{x}(t), \overline{p}(t), \overline{u}(t)) = \frac{1}{2}(\overline{x}(t)^2 - \overline{p}(t)^2) = \frac{1}{2}\left(\frac{x_0}{\cosh(T)}\right)^2,$$

car

$$\overline{x}(t) = \frac{x_0}{\cosh(T)}\cosh(T-t), \qquad \overline{p}(t) = \frac{x_0}{\cosh(T)}\sinh(T-t).$$

Ce calcul confirme que le Hamiltonien est bien constant le long de l'extrémale, comme annoncé à la Proposition 6.3.4.

6.4 Équation de Riccati : feedback

L'objectif de cette section est de montrer qu'il est possible, en résolvant l'équation de Riccati, de formuler à tout temps $t \in [0,T]$ le contrôle optimal $\overline{u}(t)$ comme un feedback (ou rétroaction) sur l'état $\overline{x}(t)$. Pour simplifier, on suppose que dérive et cible sont nulles.

Théorème 6.4.1 *On suppose que dérive et cible sont nulles. Il existe une unique matrice $P \in C^1([0,T];\mathbb{R}^{d\times d})$ solution de l'**équation de Riccati***

$$\dot{P}(t) = -A^*P(t) - P(t)A + P(t)BR^{-1}B^*P(t) - Q,\ \forall t \in [0,T], \quad P(T) = D,$$

et on a

$$\overline{p}(t) = P(t)\overline{x}(t), \quad \forall t \in [0,T],$$

*si bien que le contrôle optimal s'écrit sous forme de **boucle fermée** :*

$$\overline{u}(t) = K(t)\overline{x}(t), \quad K(t) = -R^{-1}B^*P(t), \quad \forall t \in [0,T].$$

De plus, la matrice $P(t)$ est symétrique semi-définie positive, et définie positive si la matrice D est définie positive. Enfin, la valeur optimale du critère est

$$J(\overline{u}) = \frac{1}{2}x_0^*P(0)x_0.$$

Démonstration. (1) Dépendance linéaire. Le problème LQ étant bien posé, on sait qu'il existe un unique couple $(\overline{x},\overline{p}) \in C^1([0,T];\mathbb{R}^d \times \mathbb{R}^d)$ tel que

$$\begin{aligned}
\frac{d\overline{x}}{dt}(t) &= A\overline{x}(t) - BR^{-1}B\overline{p}(t), \quad \overline{x}(0) = x_0,\\
\frac{d\overline{p}}{dt}(t) &= -A^*\overline{p}(t) - Q\overline{x}(t), \qquad \overline{p}(T) = D\overline{x}(T).
\end{aligned}$$

Par linéarité, le couple $(\overline{x},\overline{p})$ dépend linéairement de la condition initiale $x_0 \in \mathbb{R}^d$. Il existe donc des matrices $\mathcal{X},\mathcal{P}$ dans $C^1([0,T];\mathbb{R}^{d\times d})$ telles que

$$\overline{x}(t) = \mathcal{X}(t)x_0, \quad \overline{p}(t) = \mathcal{P}(t)x_0, \quad \forall t \in [0,T],$$

et on a $\mathcal{X}(0) = I_d$.
(2) Inversibilité de $\mathcal{X}(t)$. Nous allons montrer que la matrice $\mathcal{X}(t)$ est inversible pour tout $t \in [0,T]$. Pour ce faire, on raisonne par l'absurde. Soit $s \in [0,T]$ et $0 \neq x_0 \in \mathbb{R}^d$ tels que $\overline{x}(s) = \mathcal{X}(s)x_0 = 0$. On a nécessairement $s > 0$ car $\mathcal{X}(0) = I_d$. De plus, on a vu que

$$\frac{d}{dt}\big(\overline{p}(t)^*\overline{x}(t)\big) = -\overline{x}(t)^*Q\overline{x}(t) - (B^*\overline{p}(t))^*R^{-1}B^*\overline{p}(t).$$

En intégrant de s à T, et comme $\overline{x}(s) = 0$, il vient

$$0 = (D\overline{x}(T))^*\overline{x}(T) + \int_s^T \Big(\overline{x}(t)^*Q\overline{x}(t) + (B^*\overline{p}(t))^*R^{-1}B^*\overline{p}(t)\Big)\,\mathrm{d}t \geq 0.$$

Les matrices D, Q, R étant symétriques (semi-)définies positives, on en déduit que

$$\overline{u}(t) = -R^{-1}B^*\overline{p}(t) = 0, \quad \forall t \in [s,T].$$

On a donc $\frac{d\overline{x}}{dt}(t) = A\overline{x}(t)$ et $\overline{x}(s) = 0$; d'où $\overline{x}(t) = 0$ sur $[s,T]$. De même, comme on a $\frac{d\overline{p}}{dt}(t) = -A^*\overline{p}(t)$ et $\overline{p}(T) = D\overline{x}(T) = 0$, il vient $\overline{p}(t) = 0$ sur $[s,T]$. On en déduit que $(\overline{x},\overline{p})$ vérifie un système différentiel linéaire avec conditions finales $\overline{x}(T) = \overline{p}(T) = 0$.

Ceci implique que $\overline{x}(t) = \overline{p}(t) = 0$ sur $[0,T]$; en particulier, on obtient $x_0 = 0$, d'où la contradiction.
(3) Équation de Riccati. On pose

$$P(t) = \mathcal{P}(t)\mathcal{X}(t)^{-1}, \quad \forall t \in [0,T].$$

Par construction, on a $P \in C^1([0,T];\mathbb{R}^{d\times d})$. De plus, on constate que

$$\begin{aligned}\frac{d\overline{p}}{dt}(t) &= \frac{dP}{dt}(t)\overline{x}(t) + P(t)\frac{d\overline{x}}{dt}(t)\\ &= \left(\frac{dP}{dt}(t) + P(t)A - P(t)BR^{-1}B^*P(t)\right)\overline{x}(t),\end{aligned}$$

et par ailleurs, on a également $\frac{d\overline{p}}{dt}(t) = -A^*\overline{p}(t) - Q\overline{x}(t)$. On en déduit que

$$\left(\frac{dP}{dt}(t) + P(t)A + A^*P(t) - P(t)BR^{-1}B^*P(t) + Q\right)\overline{x}(t) = 0,$$

pour tout $t \in [0,T]$ et pour tout $x_0 \in \mathbb{R}^d$. Pour tout $t \in [0,T]$ fixé, le vecteur $\overline{x}(t)$ décrit $\mathbb{R}^d$ lorsque x_0 décrit $\mathbb{R}^d$ (car $\mathcal{X}(t)$ est inversible). Par conséquent, la fonction $t \mapsto P(t)$ est bien solution de l'équation de Riccati pour tout $t \in [0,T]$. En raisonnant de manière analogue, on constate que $\overline{p}(T) = D\overline{x}(T) = P(T)\overline{x}(T)$. Comme $\overline{x}(T)$ décrit $\mathbb{R}^d$ lorsque x_0 décrit $\mathbb{R}^d$, on conclut que $P(T) = D$.
(4) Propriétés de $P(t)$. La fonction $t \mapsto P(t)$ est solution d'un système différentiel quadratique. La non-linéarité satisfait donc une condition de Lipschitz locale, ce qui assure l'unicité de la solution. L'unicité prouve que $P(t)$ est symétrique pour tout $t \in [0,T]$ car la fonction $t \mapsto P(t)^*$ satisfait la même équation. Afin d'établir la positivité de $P(t)$ pour tout $t \in [0,T]$, on raisonne comme suit. Soit $x \in \mathbb{R}^d$. Posons $x_0 = \mathcal{X}(t)^{-1}x$ de sorte que $x = \overline{x}(t)$ où $\overline{x}$ est la trajectoire optimale issue de x_0. Comme la fonction $t \mapsto \overline{p}(t)^*\overline{x}(t)$ est décroissante, il vient

$$x^*P(t)x = \overline{x}(t)^*P(t)\overline{x}(t) \geq \overline{x}(T)^*D\overline{x}(T) \geq 0,$$

ce qui montre que $P(t)$ est semi-définie positive. Enfin, si la matrice D est définie positive, cela entraîne $\overline{x}(T) = 0$, d'où $x = \mathcal{X}(t)\mathcal{X}(T)^{-1}\overline{x}(T) = 0$, i.e., la matrice $P(t)$ est alors définie positive.
(5) Valeur optimale du critère. Il vient

$$\begin{aligned}J(\overline{u}) &= \frac{1}{2}\int_0^T \left(\overline{x}(t)^*Q\overline{x}(t) + \overline{u}(t)^*R\overline{u}(t)\right)\mathrm{d}t + \frac{1}{2}\overline{x}(T)^*D\overline{x}(T)\\ &= \frac{1}{2}\int_0^T \left(\overline{x}(t)^*Q\overline{x}(t) - \overline{p}(t)^*B\overline{u}(t)\right)\mathrm{d}t + \frac{1}{2}\overline{x}(T)^*D\overline{x}(T)\\ &= \frac{1}{2}\int_0^T \left(\overline{x}(t)^*Q\overline{x}(t) - \overline{p}(t)^*B\overline{u}(t)\right)\mathrm{d}t + \frac{1}{2}\overline{p}(T)^*\overline{x}(T)\\ &= \frac{1}{2}\int_0^T -\frac{d}{dt}\big(\overline{p}(t)^*\overline{x}(t)\big)\,\mathrm{d}t + \frac{1}{2}\overline{p}(T)^*\overline{x}(T)\\ &= \frac{1}{2}\overline{p}(0)^*\overline{x}(0) = \frac{1}{2}\overline{x}(0)^*P(0)\overline{x}(0) = \frac{1}{2}x_0^*P(0)x_0,\end{aligned}$$

ce qui conclut la preuve. □

Remarque 6.4.2 [Représentation linéaire de l'équation de Riccati] Au lieu de résoudre un système différentiel quadratique de taille $\frac{d(d+1)}{2}$ (P est symétrique), on peut considérer le système différentiel **linéaire** suivant qui est de taille $2d$:

$$\frac{d}{dt}\begin{pmatrix} x(t) \\ p(t) \end{pmatrix} = \underbrace{\begin{pmatrix} A & -BR^{-1}B^* \\ -Q & -A^* \end{pmatrix}}_{=\mathbb{A}\in\mathbb{R}^{(2d)\times(2d)}} \begin{pmatrix} x(t) \\ p(t) \end{pmatrix}$$

On note $R(t) = e^{(T-t)\mathbb{A}}$ la résolvante associée à ce système différentiel (telle que $R(T) = I_{2d}$). On pose

$$R(t) = \begin{pmatrix} R_1(t) & R_2(t) \\ R_3(t) & R_4(t) \end{pmatrix} \in \mathbb{R}^{(2d)\times(2d)},$$

où les quatre blocs sont à valeurs dans $\mathbb{R}^{d\times d}$. On a $x(t) = R_1(t)x(T) + R_2(t)p(T)$ et $p(t) = R_3(t)x(T) + R_4(t)p(T)$. Or $p(T) = Dx(T)$, si bien qu'en posant $\mathcal{X}_T(t) = R_1(t) + R_2(t)D$ et $\mathcal{P}_T(t) = R_3(t) + R_4(t)D$, il vient $x(t) = \mathcal{X}_T(t)x(T)$ et $p(t) = \mathcal{P}_T(t)x(T)$. En conclusion, la matrice $P(t)$ solution de l'équation de Riccati s'obtient également à partir de la résolvante du système linéaire de taille $2d$ ci-dessus en posant

$$P(t) = (R_3(t) + R_4(t)D)(R_1(t) + R_2(t)D)^{-1} \in \mathbb{R}^{d\times d}.$$

Cette expression est intéressante en pratique car elle évite de devoir résoudre un système différentiel non-linéaire.

Exemple 6.4.3 On reprend l'Exemple 6.2.6 du mouvement d'un point matériel le long d'une droite, dont on contrôle la vitesse, i.e., $\dot{x}_u(t) = u(t)$, pour tout $t \in [0,T]$, et $x_u(0) = x_0$. Le critère à minimiser est à nouveau $J(u) = \frac{1}{2}\int_0^T x_u(t)^2\,\mathrm{d}t + \frac{1}{2}\int_0^T u(t)^2\,\mathrm{d}t$. Ce problème rentre dans le cadre du système LQ avec $d = k = 1$ et

$$A = 0, \quad B = 1, \quad R = 1, \quad Q = 1, \quad D = 0, \quad \xi \equiv 0.$$

L'équation de Riccati pour la fonction $P(t)$, ici à valeurs scalaires, s'écrit

$$\dot{P}(t) = P(t)^2 - 1, \quad \forall t \in [0,T], \qquad P(T) = 0.$$

On obtient $P(t) = \tanh(T-t)$. Le contrôle optimal se met alors sous forme de boucle fermée

$$\overline{u}(t) = K(t)\overline{x}(t), \quad K(t) = -P(t) = -\tanh(T-t).$$

Pour mémoire, on avait trouvé que

$$\begin{aligned} \overline{x}(t) &= \tfrac{x_0}{\cosh(T)}\cosh(T-t), \\ \overline{u}(t) &= -\overline{p}(t) = -\tfrac{x_0}{\cosh(T)}\sinh(T-t), \end{aligned}$$

ce qui permet de retrouver l'expression ci-dessus liant $\overline{u}(t)$ à $\overline{x}(t)$. Enfin, la valeur optimale du critère est $J(\overline{u}) = \frac{1}{2}x_0^2 P(0) = \frac{1}{2}x_0^2\tanh(T)$.

Chapitre 7

PRINCIPE DU MINIMUM DE PONTRYAGUINE

Ce chapitre est consacré au problème de contrôle optimal pour des systèmes non-linéaires. Le résultat phare est le **principe du minimum de Pontryaguine** (PMP) dont nous nous contenterons d'esquisser la preuve. Nous verrons que le PMP ne fournit que des **conditions nécessaires d'optimalité** dont la formulation fait intervenir, comme pour le système LQ du chapitre précédent, les notions d'**état adjoint** et de **Hamiltonien**. En revanche, le PMP ne dit rien sur l'existence d'un contrôle optimal ni sur le caractère suffisant de ces conditions. L'intérêt pratique du PMP est de nous permettre de faire un premier tri des contrôles candidats à l'optimalité ; en espérant que les contrôles vérifiant les conditions nécessaires d'optimalité du PMP ne sont pas trop nombreux, on pourra ensuite les examiner individuellement pour en déterminer le caractère optimal ou non. Afin de nous familiariser avec l'emploi du PMP, nous présentons dans ce chapitre deux exemples d'application : le système LQ avec des contraintes sur le contrôle d'une part et un modèle non-linéaire de dynamique de populations d'autre part.

7.1 Systèmes de contrôle non-linéaires

On se donne un intervalle de temps $[0,T]$ avec $T > 0$, on considère un état à valeurs dans $\mathbb{R}^d$, $d \geq 1$, et un contrôle à valeurs dans un sous-ensemble fermé non-vide $U \subset \mathbb{R}^k$. On s'intéresse au système de contrôle non-linéaire

$$\dot{x}_u(t) = f(t, x_u(t), u(t)), \quad \forall t \in [0,T], \qquad x_u(0) = x_0, \tag{7.1}$$

avec une dynamique décrite par la fonction $f : [0,T] \times \mathbb{R}^d \times U \to \mathbb{R}^d$. L'ensemble des contrôles admissibles est ici le sous-ensemble

$$\mathcal{U} = L^1([0,T]; U) \subset L^1([0,T]; \mathbb{R}^k). \tag{7.2}$$

L'objectif est de trouver un contrôle optimal $\overline{u} \in \mathcal{U}$ qui minimise le critère

$$J(u) = \int_0^T g(t, x_u(t), u(t))\,\mathrm{d}t + h(x_u(T)), \tag{7.3}$$

où les fonctions $g : [0,T] \times \mathbb{R}^d \times U \to \mathbb{R}$ et $h : \mathbb{R}^d \to \mathbb{R}$ sont données. Le problème de contrôle optimal est donc le suivant :

$$\text{Chercher } \overline{u} \in \mathcal{U} \text{ tel que } J(\overline{u}) = \inf_{u \in \mathcal{U}} J(u). \tag{7.4}$$

Nous allons formuler quelques hypothèses (en général, raisonnables) sur les différents ingrédients intervenant dans la formulation du problème de contrôle optimal (7.4), à savoir la fonction f pour la dynamique et les fonctions g et h pour le critère. Commençons par les hypothèses sur la dynamique. On suppose que

(a) $f \in C^0([0,T] \times \mathbb{R}^d \times U; \mathbb{R}^d)$ et f est de classe C^1 par rapport à x ;

(b) $\exists C$, $|f(t,y,v)|_{\mathbb{R}^d} \leq C(1 + |y|_{\mathbb{R}^d} + |v|_{\mathbb{R}^k})$, $\forall t \in [0,T]$, $\forall y \in \mathbb{R}^d$, $\forall v \in U$;

(c) Pour tout $R > 0$, $\exists C_R$, $|\frac{\partial f}{\partial x}(t,y,v)|_{\mathbb{R}^{d\times d}} \leq C_R(1 + |v|_{\mathbb{R}^d})$, $\forall t \in [0,T]$, $\forall y \in \overline{B}(0,R)$, $\forall v \in U$.

Dans ces hypothèses, C et C_R désignent des constantes génériques indépendantes de (t,y,v), C_R dépendant du rayon R de la boule fermée $\overline{B}(0,R)$; par la suite, nous utiliserons les symboles C et C_R avec la convention que les valeurs de C et de C_R peuvent changer à chaque utilisation tant qu'elles restent indépendantes du temps, de l'état du système et de la valeur du contrôle. L'objectif des trois hypothèses ci-dessus est d'assurer, pour tout contrôle $u \in \mathcal{U}$, l'existence et l'unicité de la trajectoire associée $x_u \in AC([0,T]; \mathbb{R}^d)$.

Lemme 7.1.1 (Existence et unicité des trajectoires) *Dans le cadre des hypothèses* (a), (b), (c) *ci-dessus, pour tout contrôle $u \in \mathcal{U}$, il existe une unique trajectoire associée $x_u \in AC([0,T]; \mathbb{R}^d)$ solution de* (7.1).

Démonstration. Il s'agit d'une conséquence de la version locale du théorème de Cauchy–Lipschitz avec une dynamique mesurable en temps uniquement (cf. le Théorème 8.3.6). On considère le système dynamique $\dot{x}(t) = F(t,x(t))$ avec la fonction $F : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ telle que $F(t,x) = f(t,x,u(t))$. La fonction F est mesurable en t, et elle est continue en x. De plus, F est localement lipschitzienne par rapport à x puisque l'on a, pour tout $t \in [0,T]$ et tout $x_1, x_2 \in \overline{B}(0,R)$,

$$|F(t,x_1) - F(t,x_2)|_{\mathbb{R}^d} \leq C_0(t)|x_1 - x_2|_{\mathbb{R}^d}, \quad C_0(t) = \sup_{y \in \overline{B}(0,R)} \left|\tfrac{\partial f}{\partial x}(t,y,u(t))\right|_{\mathbb{R}^{d\times d}}.$$

Comme $C_0(t) \leq C_R(1+|u(t)|_{\mathbb{R}^k})$ grâce à l'hypothèse (c), on a bien $C_0 \in L^1([0,T]; \mathbb{R}_+)$. En outre, la fonction F est localement intégrable grâce à l'hypothèse (b) puisque l'on a, pour tout $x \in \mathbb{R}^d$ et tout $t \in [0,T]$,

$$|F(t,x)|_{\mathbb{R}^d} \leq C(1 + |x|_{\mathbb{R}^d} + |u(t)|_{\mathbb{R}^k}) \in L^1([0,T]; \mathbb{R}_+).$$

Il reste enfin à s'assurer que la trajectoire maximale est bien définie sur tout l'intervalle $[0,T]$ (i.e., qu'il n'y a pas eu d'explosion en un temps $t_* < T$). Pour cela, on utilise le lemme de Gronwall rappelé ci-dessous. Comme on a $x(t) = x_0 + \int_0^t f(s,x(s),u(s))\,\mathrm{d}s$, on peut appliquer ce lemme avec $z(t) = |x(t)|_{\mathbb{R}^d}$ et $\psi(t) \equiv C$. L'estimation (7.5) est satisfaite avec $\alpha = |x_0|_{\mathbb{R}^d} + C(T + \|u\|_{L^1([0,T];\mathbb{R}^k)})$ grâce à l'hypothèse (b). On en déduit que la trajectoire reste bien bornée sur $[0,T]$, i.e., il n'y a pas d'explosion. □

Lemme 7.1.2 (Gronwall) *Soit $\psi, z : [0,T] \to \mathbb{R}_+$ deux fonctions continues telles que*

$$\exists \alpha \geq 0, \quad \forall t \in [0,T], \quad z(t) \leq \alpha + \int_0^t \psi(s) z(s)\, \mathrm{d}s. \tag{7.5}$$

Alors, on a $z(t) \leq \alpha e^{\int_0^t \psi(s)\,\mathrm{d}s}$ pour tout $t \in [0,T]$.

Démonstration. Posons $\Psi(t) = \int_0^t \psi(s)\,\mathrm{d}s$ et $v(t) = e^{-\Psi(t)} \int_0^t \psi(s) z(s)\,\mathrm{d}s$. En utilisant (7.5), on constate que

$$\begin{aligned}\frac{dv}{dt}(t) &= -\psi(t) e^{-\Psi(t)} \int_0^t \psi(s) z(s)\,\mathrm{d}s + e^{-\Psi(t)} \psi(t) z(t) \\ &= \psi(t) e^{-\Psi(t)} \left(z(t) - \int_0^t \psi(s) z(s)\,\mathrm{d}s \right) \leq \alpha \psi(t) e^{-\Psi(t)}.\end{aligned}$$

Comme $v(0) = 0$ et $\Psi(0) = 0$, en intégrant cette majoration de 0 à t, il vient

$$e^{-\Psi(t)} \int_0^t \psi(s) z(s)\,\mathrm{d}s = v(t) \leq \alpha \int_0^t \psi(s) e^{-\Psi(s)}\,\mathrm{d}s = \alpha(1 - e^{-\Psi(t)}),$$

et en ré-arrangeant les termes, on obtient

$$\alpha + \int_0^t \psi(s) z(s)\,\mathrm{d}s \leq \alpha e^{\Psi(t)}.$$

On conclut en utilisant à nouveau la borne (7.5) sur $z(t)$. □

Venons en maintenant aux hypothèses sur le critère. On suppose que

(d) $g \in C^0([0,T] \times \mathbb{R}^d \times U; \mathbb{R})$ et g est de classe C^1 par rapport à x ; de plus, $h \in C^1(\mathbb{R}^d; \mathbb{R})$;

(e) Pour tout $R > 0$, $\exists C_R$, $|g(t,y,v)| \leq C_R(1 + |v|_{\mathbb{R}^k})$, $\forall t \in [0,T]$, $\forall y \in \overline{B}(0,R)$, $\forall v \in U$;

(f) Pour tout $R > 0$, $\exists C_R$, $|\frac{\partial g}{\partial x}(t,y,v)|_{\mathbb{R}^d} \leq C_R(1 + |v|_{\mathbb{R}^k})$, $\forall t \in [0,T]$, $\forall y \in \overline{B}(0,R)$, $\forall v \in U$;

(g) Les fonctions g et h sont minorées respectivement sur $[0,T] \times \mathbb{R}^d \times U$ et sur $\mathbb{R}^d$.

Ces hypothèses nous permettent d'affirmer que, pour tout $u \in \mathcal{U}$, le critère $J(u)$ est bien défini car la trajectoire associée x_u est bien définie et $x_u(t) \in \overline{B}(0, R(u))$, pour tout $t \in [0,T]$, si bien que grâce à l'hypothèse (e), la fonction $t \mapsto g(t, x(t), u(t))$ est bien intégrable. En outre, l'infimum de J sur $\mathcal{U}$ est bien fini grâce à l'hypothèse (g). Il est donc raisonnable de considérer le problème de minimisation (7.4). Les hypothèses (d) et (f) nous seront utiles à la section suivante pour définir l'état adjoint.

7.2 PMP : énoncé et commentaires

L'objectif de cette section est d'énoncer le principe du minimum de Pontryaguine (PMP) pour le système de contrôle non-linéaire (7.1) et la fonctionnelle J définie en (7.3). Nous nous contentons d'énoncer le PMP et d'en voir quelques premiers

exemples d'application. La preuve du PMP sera esquissée dans la section suivante. Comme dans le cas plus simple du système linéaire-quadratique (cf. la Section 6.3), le PMP repose sur la notion de Hamiltonien.

Définition 7.2.1 *Le **Hamiltonien** associé au système de contrôle non-linéaire* (7.1) *et à la fonctionnelle J définie en* (7.3) *est l'application $H : [0,T] \times \mathbb{R}^d \times \mathbb{R}^d \times U \to \mathbb{R}$ définie par*

$$H(t,x,p,u) = p^* f(t,x,u) + g(t,x,u). \tag{7.6}$$

*On notera bien que dans cette écriture, (x,p,u) désigne un vecteur générique de $\mathbb{R}^d \times \mathbb{R}^d \times U$. Lorsque l'application H ne dépend pas explicitement du temps, on dit que le Hamiltonien est **autonome**.*

Théorème 7.2.2 (PMP) *Si $\overline{u} \in \mathcal{U}$ est un contrôle optimal, i.e., si $\overline{u}$ est une solution de* (7.4)*, alors en notant $\overline{x} = x_{\overline{u}} \in AC([0,T];\mathbb{R}^d)$ la trajectoire associée au contrôle $\overline{u}$ et en définissant l'**état adjoint** $\overline{p} \in AC([0,T];\mathbb{R}^d)$ solution de*

$$\frac{d\overline{p}}{dt}(t) = -\overline{A}(t)^* \overline{p}(t) - \overline{b}(t), \ \forall t \in [0,T], \quad \overline{p}(T) = \frac{\partial h}{\partial x}(\overline{x}(T)) \in \mathbb{R}^d, \tag{7.7}$$

où pour tout $t \in [0,T]$,

$$\overline{A}(t) = \frac{\partial f}{\partial x}(t,\overline{x}(t),\overline{u}(t)) \in \mathbb{R}^{d\times d}, \qquad \overline{b}(t) = \frac{\partial g}{\partial x}(t,\overline{x}(t),\overline{u}(t)) \in \mathbb{R}^d, \tag{7.8}$$

on a, p.p. $t \in [0,T]$,

$$\overline{u}(t) \in \arg\min_{v\in U} H(t,\overline{x}(t),\overline{p}(t),v), \tag{7.9}$$

où le Hamiltonien $H : [0,T] \times \mathbb{R}^d \times \mathbb{R}^d \times U \to \mathbb{R}$ est défini en (7.6)*. Un triplet $(\overline{x},\overline{p},\overline{u})$ satisfaisant les conditions ci-dessus est appelé une **extrémale**. (On notera qu'avec les conventions adoptées, $\frac{\partial g}{\partial x}$ et $\frac{\partial h}{\partial x}$ sont des vecteurs colonne.)*

Remarque 7.2.3 L'état adjoint $\overline{p}$ est solution d'un système linéaire (à $(\overline{x},\overline{u})$ fixés) instationnaire et rétrograde en temps. Ce système, ainsi que la condition finale sur $\overline{p}(T)$, sont bien définis grâce aux hypothèses (a) et (d) ci-dessus. De plus, ce système admet une unique solution car la fonction $\overline{b}$ est bien intégrable en temps grâce à l'hypothèse (f) et la fonction $\overline{A}$ est dans $L^1([0,T];\mathbb{R}^{d\times d})$ grâce à l'hypothèse (c).

Remarque 7.2.4 Dans le cas du système de contrôle non-linéaire (7.1) avec la fonctionnelle J définie en (7.3), le PMP ne fournit qu'une **condition nécessaire d'optimalité**. En revanche, le PMP ne dit rien sur l'existence d'un contrôle optimal, et il ne fournit pas *en général* de condition suffisante (cf. toutefois la Proposition 7.2.8 ci-dessous). L'intérêt pratique du PMP est de restreindre le champ des possibles en vue de l'obtention d'un contrôle optimal : on commence par considérer les extrémales et, en espérant qu'elles ne sont pas trop nombreuses, on en fait ensuite le tri.

Exemple 7.2.5 [Système LQ] Appliquons le Théorème 7.2.2 au système LQ étudié au chapitre précédent. Pour simplifier, on omet le terme de dérive. On a

$$f(t,x,u) = Ax+Bu, \quad g(t,x,u) = \tfrac{1}{2}u^*Ru+\tfrac{1}{2}e_x(t)^*Qe_x(t), \quad h(x) = \tfrac{1}{2}e_x(T)^*De_x(T),$$

où $e_x(t) = x - \xi(t)$; on rappelle que les matrices $Q, D \in \mathbb{R}^{d\times d}$ sont symétriques semi-définies positives, que la matrice $R \in \mathbb{R}^{k\times k}$ est symétrique définie positive et que $\xi \in C^0([0,T];\mathbb{R}^d)$ est la trajectoire cible. Pour le système LQ, il n'y a pas de contraintes sur le contrôle, on a donc $U = \mathbb{R}^k$. Le Hamiltonien s'écrit

$$H(t,x,p,u) = p^*(Ax + Bu) + \tfrac{1}{2}u^*Ru + \tfrac{1}{2}e_x(t)^*Qe_x(t).$$

On a donc (noter l'unicité du minimiseur)

$$\overline{u}(t) = \arg\min_{v\in\mathbb{R}^k}\left(\overline{p}^*Bv + \frac{1}{2}v^*Rv\right),$$

ce qui équivaut à

$$\overline{u}(t) = -R^{-1}B^*\overline{p}(t).$$

Comme $\frac{\partial f}{\partial x} = A$, $\frac{\partial g}{\partial x} = Qe_x$, $\frac{\partial h}{\partial x} = De_x$, l'équation (7.7) sur l'état adjoint devient

$$\frac{d\overline{p}}{dt}(t) = -A^*\overline{p}(t) - Qe_{\overline{x}}(t), \quad \forall t \in [0,T], \qquad \overline{p}(T) = De_{\overline{x}}(T),$$

qui est bien l'équation différentielle rétrograde et la condition finale qui avaient été obtenues au chapitre précédent pour l'état adjoint (cf. le Théorème 6.2.2).

Contre-exemple 7.2.6 Donnons un exemple relativement simple de non-existence de contrôle optimal. On considère le système de contrôle linéaire $\dot{x}_u(t) = u(t)$ avec $x_u(0) = x_0 = 0$ et $T = 1$. Le critère à minimiser est

$$J(u) = \int_0^1 x_u(t)^2\,\mathrm{d}t + \int_0^1 (u(t)^2 - 1)^2\,\mathrm{d}t, \quad U = [-1,1].$$

Alors, on a $\inf_{u\in\mathcal{U}} J(u) = 0$ et il n'existe pas de contrôle optimal. Pour le montrer, on considère pour tout $n \in \mathbb{N}_*$ la suite minimisante de contrôles

$$u_n(t) = (-1)^k, \quad t \in [\tfrac{k}{2n}, \tfrac{k+1}{2n}[, \quad k \in \{0,\dots,2n-1\},$$

dont la trajectoire associée, x_n, est en dents de scie et vérifie $\|x_n\|_{L^\infty(0,1)} \le \frac{1}{2n}$ (cf. la Figure 7.1). On en déduit que $J(u_n) \le \frac{1}{4n^2}$. S'il existait $\overline{u} \in \mathcal{U}$ tel que $J(\overline{u}) = 0$, alors on aurait $\overline{x}(t) \equiv 0$ et $\overline{u}(t) \in \{-1,1\}$, mais $\overline{u}(t) = \frac{d\overline{x}}{dt}(t) = 0$. La difficulté rencontrée dans cet exemple provient de la non-convexité du critère. Evidemment, le lecteur attentif aura reconnu l'Exemple 2.3.2, adapté au contexte du contrôle, et qui donnait aussi un contre-exemple à l'existence d'une solution pour un problème d'optimisation.

Exemple 7.2.7 [Absence de condition suffisante] Donnons maintenant un exemple où le PMP ne fournit pas de condition suffisante d'optimalité. On considère à nouveau le système de contrôle linéaire $\dot{x}(t) = u(t)$ avec $x_0 = 0$ et $T = 1$. Le critère à minimiser est cette fois

$$J(u) = \int_0^1 (x_u(t)^2 - 1)^2\,\mathrm{d}t, \quad U = [-1,1].$$

FIGURE 7.1 – Illustration du Contre-exemple 7.2.6 : contrôle issu d'une suite minimisante et trajectoire associée.

On cherche donc à minimiser la distance de $x(t)$ à l'ensemble $\{-1,1\}$; les contraintes sur u font que $x(t) \in [-1,1]$, $\forall t \in [0,T]$. Il y a donc deux contrôles optimaux, qui sont $\overline{u}_\pm(t) \equiv \pm 1$, pour tout $t \in [0,T]$, et on a $\inf_{u\in\mathcal{U}} J(u) = \int_0^1 (t^2-1)^2 \,\mathrm{d}t = \frac{8}{15}$. Or, si on considère le contrôle $\overline{u}(t) \equiv 0$, celui-ci vérifie les conditions du PMP mais ce n'est pas un contrôle optimal car $J(0) = 1 > \frac{8}{15}$. En effet, on a $f(t,x,u) = u$, $g(t,x,u) = (x^2-1)^2$, $h = 0$, la trajectoire associée est $\overline{x}(t) \equiv 0$ et l'état adjoint est $\overline{p}(t) \equiv 0$. Le Hamiltonien à minimiser est $H(t,\overline{x}(t),\overline{p}(t),v) = (\overline{x}(t)^2-1)^2$ dont un minimiseur est bien $v = 0$. La difficulté rencontrée dans cet exemple provient à nouveau de la non-convexité du critère.

Concluons cette section par un résultat positif quant au caractère suffisant de la condition d'optimalité du PMP.

Proposition 7.2.8 (Condition suffisante) *Le PMP fournit une* ***condition suffisante*** *d'optimalité sous les hypothèses suivantes :*

- $f(t,x,u) = A(t)x + B(t)u$ *avec* $A \in C^0([0,T];\mathbb{R}^{d\times d})$ *et* $B \in C^0([0,T];\mathbb{R}^{d\times k})$ *;*
- $\mathcal{U} = L^2([0,T];U)$ *où* U *est un ensemble* ***convexe, compact*** *non-vide ;*
- *la fonction* g *est* ***convexe*** *et différentiable en* $(x,u) \in \mathbb{R}^d \times U$ *;*
- *la fonction* h *est* ***convexe*** *et différentiable en* $x \in \mathbb{R}^d$.

Démonstration. Nous nous contenterons d'esquisser la preuve. La fonctionnelle J est convexe en u sur l'ensemble convexe $K = L^2([0,T];U)$ (on travaille dans L^2 afin de se placer dans le cadre des espaces de Hilbert). De par le Théorème 2.5.1, $\overline{u}$ est un contrôle optimal dans K si et seulement si

$$(\nabla J(\overline{u}), v - \overline{u})_{L^2([0,T];\mathbb{R}^k)} \geq 0, \quad \forall v \in K.$$

Grâce à l'introduction de l'état adjoint $\overline{p}$ solution de (7.7), ceci se récrit

$$\int_0^T \left(\overline{p}(t)^* B(v(t) - \overline{u}(t)) + \frac{\partial g}{\partial u}(t,\overline{x}(t),\overline{u}(t))^*(v(t) - \overline{u}(t)) \right) \mathrm{d}t \geq 0, \quad \forall v \in K.$$

Cette inégalité, toujours grâce au Théorème 2.5.1, équivaut au fait que $\overline{u}$ soit minimiseur sur K de la fonctionnelle

$$\tilde{J}(u) = \int_0^T \left(\overline{p}(t)^* B u(t) + g(t,\overline{x}(t),u(t)) \right) \mathrm{d}t.$$

En posant $\Phi(t,v) := \overline{p}(t)^* Bv + g(t,\overline{x}(t),v)$, nous avons donc établi que

$$\int_0^T \Phi(t,\overline{u}(t))\,\mathrm{d}t \leq \int_0^T \Phi(t,u(t))\,\mathrm{d}t, \qquad \forall u \in K.$$

Supposons par l'absurde que $\overline{u}(t) > \min_{v\in U} \Phi(t,v)$ sur un sous-ensemble de $[0,T]$ de mesure strictement positive. Comme $\Phi(t,\overline{u}(t)) \geq \min_{v\in U} \Phi(t,v)$ puisque $\overline{u}(t) \in U$ par hypothèse, ceci implique que

$$\int_0^T \min_{v\in U} \Phi(t,v)\,\mathrm{d}t < \int_0^T \Phi(t,\overline{u}(t))\,\mathrm{d}t.$$

En posant $\hat{u}(t) := \arg\min_{v\in U}\Phi(t,v)$ sur $[0,T]$, on peut montrer (voir le Théorème 8.2.9 de sélection mesurable dont la démonstration sort du cadre de ce cours) que la fonction $\hat{u}$ ainsi définie est bien mesurable. Elle est de plus à valeurs dans U par construction, et comme U est borné par hypothèse, $\hat{u}$ est bien de carré sommable. En conclusion, $\hat{u} \in K$, ce qui fournit la contradiction attendue puisqu'il vient

$$\int_0^T \Phi(t,\hat{u}(t))\,\mathrm{d}t = \int_0^T \min_{v\in U} \Phi(t,v)\,\mathrm{d}t < \int_0^T \Phi(t,\overline{u}(t))\,\mathrm{d}t \leq \int_0^T \Phi(t,\hat{u}(t))\,\mathrm{d}t.$$

Ainsi $\overline{u}(t)$ est minimiseur instantané de $v \mapsto \overline{p}(t)^* Bv + g(t,\overline{x}(t),v)$, ce qui n'est rien d'autre que minimiser le Hamiltonien par rapport à v. □

7.3 Application au système LQ avec contraintes

L'objectif de cette section est d'illustrer le PMP dans le cas du système LQ (dynamique linéaire et critère quadratique), mais contrairement au Chapitre 6, nous supposons ici qu'il y a des contraintes sur le contrôle. Malgré la présence de ces contraintes, ce nouveau problème de contrôle optimal reste relativement simple, et il nous sera en fait possible de prouver le PMP (et d'en établir le caractère suffisant) en nous appuyant sur l'inéquation d'Euler caractérisant le minimiseur d'une fonctionnelle convexe sur un sous-ensemble convexe, fermé, non-vide d'un espace de Hilbert (cf. le Théorème 2.5.1).

Soit $T > 0$, une matrice $A \in \mathbb{R}^{d\times d}$, une matrice $B \in \mathbb{R}^{d\times k}$ et une condition initiale $x_0 \in \mathbb{R}^d$. Le système de contrôle linéaire s'écrit sous la forme

$$\dot{x}_u(t) = Ax_u(t) + Bu(t), \quad \forall t \in [0,T], \qquad x_u(0) = x_0. \tag{7.10}$$

Soit U un sous-ensemble **convexe, fermé, non-vide** de $\mathbb{R}^k$. L'ensemble des contrôles admissibles est ici le sous-ensemble

$$K = L^2([0,T];U). \tag{7.11}$$

On s'intéresse au problème de minimisation sous contraintes

$$\text{Chercher } \overline{u} \in K \text{ tel que } J(\overline{u}) = \inf_{u\in K} J(u), \tag{7.12}$$

avec le critère quadratique

$$J(u) = \frac{1}{2}\int_0^T u(t)^* R u(t)\,\mathrm{d}t + \frac{1}{2}\int_0^T e_{x_u}(t)^* Q e_{x_u}(t)\,\mathrm{d}t + \frac{1}{2} e_{x_u}(T)^* D e_{x_u}(T), \quad (7.13)$$

où $e_{x_u} = x_u - \xi$ et $\xi \in C^0([0,T];\mathbb{R}^d)$ est la trajectoire cible. Comme dans le Chapitre 6, les matrices $Q, D \in \mathbb{R}^{d\times d}$ sont symétriques semi-définies positives, tandis que la matrice $R \in \mathbb{R}^{k\times k}$ est symétrique définie positive.

Lemme 7.3.1 *Il existe une unique solution au problème* (7.12)*, i.e., la fonctionnelle J définie par* (7.13) *admet un unique minimiseur sur le sous-ensemble K défini par* (7.11)*.*

Démonstration. Nous allons appliquer le Théorème 2.3.8. D'une part, K est un sous-ensemble convexe, fermé, non-vide de l'espace de Hilbert $V = L^2([0,T];\mathbb{R}^k)$. En effet,

- K est non-vide car le sous-ensemble U est non-vide (considérer un contrôle constant en temps égal à un élément de U) ;
- K est convexe car le sous-ensemble U est convexe (pour tout $u_1, u_2 \in K$ et $\theta \in [0,1]$, on a $\theta u_1(t) + (1-\theta)u_2(t) \in U$ p.p. $t \in [0,T]$ car U est convexe, si bien que $\theta u_1 + (1-\theta)u_2 \in K$) ;
- enfin, K est fermé dans V car si $(u_n)_{n\in\mathbb{N}}$ est une suite de K convergeant vers u dans V, comme la convergence dans $L^2([0,T];\mathbb{R}^k)$ implique la convergence p.p. (à une sous-suite près) et que le sous-ensemble U est fermé, on en déduit que $u(t) \in U$ p.p. $t \in [0,T]$, i.e., $u \in K$.

D'autre part, la fonctionnelle J est fortement convexe et continue (elle est même différentiable) sur V (cf. les Lemmes 6.1.2 et 6.2.1). □

Dans la suite de cette section, on notera $\overline{u} \in K = L^2([0,T];U)$ l'unique contrôle optimal solution de (7.12) et $\overline{x} = x_{\overline{u}}$ la trajectoire associée. Le système LQ avec contraintes rentre dans le champ d'application du PMP. En procédant comme à l'Exemple 7.2.5 (qui traitait le cas sans contraintes), on introduit l'état adjoint $\overline{p} \in C^1([0,T];\mathbb{R}^d)$ tel que

$$\frac{d\overline{p}}{dt}(t) = -A^*\overline{p}(t) - Q e_{\overline{x}}(t), \quad \forall t \in [0,T], \qquad \overline{p}(T) = D e_{\overline{x}}(T), \quad (7.14)$$

où $e_{\overline{x}}(t) = \overline{x}(t) - \xi(t)$ p.p. $t \in [0,T]$, et le Hamiltonien $H : [0,T]\times\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^k \to \mathbb{R}$ tel que

$$H(t,x,p,u) = p^*(Ax+Bu) + \frac{1}{2}u^*Ru + \frac{1}{2}(x-\xi(t))^*Q(x-\xi(t)). \quad (7.15)$$

En appliquant le PMP (cf. le Théorème 7.2.2), on en déduit qu'une condition nécessaire d'optimalité est que, p.p. $t \in [0,T]$, $\overline{u}(t)$ est un minimiseur de $H(t,\overline{x}(t),\overline{p}(t),v)$ sur U, i.e.,

$$\overline{u}(t) \in \arg\min_{v\in U} H(t,\overline{x}(t),\overline{p}(t),v). \quad (7.16)$$

En inspectant l'expression de H, on voit que de manière équivalente, on a

$$\overline{u}(t) \in \arg\min_{v\in U}\left(v^*B^*\overline{p}(t) + \frac{1}{2}v^*Rv\right). \tag{7.17}$$

Or, la fonctionnelle en v au membre de droite est quadratique et fortement convexe. On en déduit qu'elle admet un unique minimiseur sur le sous-ensemble convexe, fermé, non-vide U de $\mathbb{R}^k$. De manière plus précise, on a donc

$$\overline{u}(t) = \arg\min_{v\in U}\left(v^*B^*\overline{p}(t) + \frac{1}{2}v^*Rv\right). \tag{7.18}$$

Lorsque $U = \mathbb{R}^k$, on retrouve bien le résultat du chapitre 6, à savoir $\overline{u}(t) = -R^{-1}B^*\overline{p}(t)$. Dans le cas général pour le sous-ensemble U, on n'a pas forcément d'expression explicite de $\overline{u}(t)$ en fonction de $\overline{p}(t)$ car celle-ci dépend de la forme du sous-ensemble U.

Proposition 7.3.2 *La condition* (7.18) *est une* ***condition nécessaire et suffisante*** *d'optimalité pour le problème* (7.12)*. En outre, cette condition définit un unique contrôle optimal $\overline{u} \in K$ et celui-ci est une fonction lipschitzienne du temps.*

Remarque 7.3.3 Le fait que le contrôle optimal $\overline{u} \in K$ soit une fonction lipschitzienne du temps montre que pour le système LQ avec contraintes, il n'y a pas de phénomènes de type bang-bang pour le contrôle optimal.

Démonstration. (1) La fonctionnelle J étant convexe et différentiable sur V, une condition nécessaire et suffisante d'optimalité pour le problème (7.12) est l'inéquation d'Euler (cf. le Théorème 2.5.1)

$$(\nabla J(\overline{u}), v - \overline{u})_V \geq 0, \quad \forall v \in K.$$

En utilisant l'expression de la différentielle de J obtenue au Lemme 6.2.1, on en déduit que

$$(R\overline{u} + B^*\overline{p}, v - \overline{u})_V \geq 0, \quad \forall v \in K,$$

ou encore, en explicitant le produit scalaire dans $V = L^2([0,T];\mathbb{R}^k)$,

$$\int_0^T (v(t) - \overline{u}(t))^*(R\overline{u}(t) + B^*\overline{p}(t))\,\mathrm{d}t \geq 0, \quad \forall v \in K = L^2([0,T];U).$$

En utilisant à nouveau l'inéquation d'Euler, ceci ne signifie rien d'autre que

$$\overline{u} = \arg\min_{v\in K}\mathcal{J}_{\overline{p}}(v),$$

où la fonctionnelle

$$\mathcal{J}_{\overline{p}} : V \to \mathbb{R}, \qquad \mathcal{J}_{\overline{p}}(v) = \int_0^T \left(v(t)^*B^*\overline{p}(t) + \frac{1}{2}v(t)^*Rv(t)\right)\mathrm{d}t$$

est quadratique, différentiable et fortement convexe sur V. On pose pour tout $t \in [0,T]$,

$$u_\sharp(t) = \arg\min_{v\in U}\left(v^* B^* \overline{p}(t) + \frac{1}{2} v^* R v\right).$$

De l'inéquation d'Euler dans $U \subset \mathbb{R}^k$, on déduit que pour tout $t \in [0,T]$,

$$(v - u_\sharp(t))^*(R u_\sharp(t) + B^*\overline{p}(t)) \geq 0, \quad \forall v \in U.$$

(2) Montrons que la fonction $u_\sharp(t)$ ainsi définie est lipschitzienne en t sur $[0,T]$. Soit $t_1, t_2 \in [0,T]$. On a

$$\begin{aligned}(u_\sharp(t_2) - u_\sharp(t_1))^*(R u_\sharp(t_1) + B^*\overline{p}(t_1)) \geq 0,\\ (u_\sharp(t_1) - u_\sharp(t_2))^*(R u_\sharp(t_2) + B^*\overline{p}(t_2)) \geq 0.\end{aligned}$$

En posant $\delta u_\sharp = u_\sharp(t_2) - u_\sharp(t_1)$, il vient

$$(\delta u_\sharp)^* R \delta u_\sharp \leq (\delta u_\sharp)^* B^*(\overline{p}(t_1) - \overline{p}(t_2)).$$

Comme la matrice R est par hypothèse définie positive, on en déduit que

$$|u_\sharp(t_2) - u_\sharp(t_1)|_{\mathbb{R}^k} = |\delta u_\sharp|_{\mathbb{R}^k} \leq \lambda_{\min}(R)^{-1} \|B^*\|_{\mathbb{R}^{k\times d}} |\overline{p}(t_2) - \overline{p}(t_1)|_{\mathbb{R}^d},$$

où $\lambda_{\min}(R) > 0$ désigne la plus petite valeur propre de la matrice R. Comme la fonction $t \mapsto \overline{p}(t)$ est de classe C^1 en t, cela montre que la fonction $t \mapsto u_\sharp(t)$ est lipschitzienne en t.
(3) En conclusion, la fonction $u_\sharp : [0,T] \to \mathbb{R}^k$ est mesurable (car lipschitzienne), de carré sommable et à valeurs dans U. On a donc $u_\sharp \in K$. De plus, comme $\overline{u}(t) \in U$ p.p. $t \in [0,T]$, l'inégalité suivante est satisfaite p.p. $t \in [0,T]$:

$$\overline{u}(t)^* B^* \overline{p}(t) + \frac{1}{2}\overline{u}(t)^* R \overline{u}(t) \geq u_\sharp(t)^* B^* \overline{p}(t) + \frac{1}{2} u_\sharp(t)^* R u_\sharp(t).$$

En intégrant cette inégalité de 0 à T, il vient

$$\mathcal{J}_{\overline{p}}(\overline{u}) \geq \mathcal{J}_{\overline{p}}(u_\sharp).$$

Par unicité du minimiseur de $\mathcal{J}_{\overline{p}}$ sur K, on conclut que $\overline{u} = u_\sharp$. □

7.4 Exemple non-linéaire : ruche d'abeilles

On considère un modèle relativement simple de dynamique de populations. Pour fixer les idées, nous allons le décliner dans le contexte de la modélisation d'une ruche d'abeilles. On suppose que dans la ruche, la population d'abeilles $a(t)$ et celle des reines $r(t)$ évolue selon la dynamique

$$\dot{x}(t) = \begin{pmatrix}\dot{a}(t)\\ \dot{r}(t)\end{pmatrix} = \begin{pmatrix}\varphi(u(t))a(t)\\ \gamma u(t) a(t)\end{pmatrix}, \quad \forall t \in [0,T], \tag{7.19}$$

où le contrôle $u \in L^\infty([0,T];U)$ avec $U = [0,1]$ représente l'effort des abeilles pour fournir des reines et où nous avons introduit la fonction

$$\varphi : [0,1] \to \mathbb{R}, \quad \varphi(v) = \alpha(1-v) - \beta. \tag{7.20}$$

Les paramètres du modèle α, β, γ sont des réels strictement positifs et on suppose que $\alpha > \beta$. On suppose également que $a(0) > 0$; comme $\dot{a}(t) = \varphi(u(t))a(t)$, on a $a(t) > 0$ pour tout $t \in [0,T]$. On notera également que

— si u est constant égal à 1, on a $\dot{a}(t) = -\beta a(t) < 0$: la population d'abeilles décroît (exponentiellement) ;

— si u est constant égal à 0, on a $\dot{a}(t) = (\alpha - \beta)a(t) > 0$: la population d'abeilles croît (exponentiellement).

Notre objectif ici est de chercher un contrôle optimal afin de maximiser la population de reines au temps T. En introduisant la fonctionnelle $J : \mathcal{U} = L^1([0,T];U) \to \mathbb{R}$ telle que

$$J(u) = -r(T), \tag{7.21}$$

le problème de contrôle optimal est donc le suivant :

$$\text{Chercher } \overline{u} \in \mathcal{U} \text{ tel que } J(\overline{u}) = \inf_{u \in \mathcal{U}} J(u). \tag{7.22}$$

On commence par chercher une condition nécessaire d'optimalité en appliquant le PMP. L'état de la ruche est décrit par le vecteur $x = (a,r)^* \in \mathbb{R}^2$. Le problème de contrôle optimal (7.22) rentre dans le cadre d'application du PMP en posant

$$f(x,u) = \begin{pmatrix} \varphi(u)a \\ \gamma u a \end{pmatrix}, \quad g(x,u) = 0, \quad h(x) = -r. \tag{7.23}$$

Soit $\overline{u} \in \mathcal{U}$ un contrôle optimal, de trajectoire associée $(\overline{a}, \overline{r})^*$. Comme $\frac{\partial f}{\partial x}(x,u) = \begin{pmatrix} \varphi(u) & 0 \\ \gamma u & 0 \end{pmatrix}$ et $\frac{\partial g}{\partial x}(x,u) = 0$, l'état adjoint $\overline{p} = (\overline{p}_a, \overline{p}_r)^* : [0,T] \to \mathbb{R}^2$ est tel que

$$\begin{cases} \dfrac{d\overline{p}_a}{dt}(t) = -\varphi(\overline{u}(t))\overline{p}_a(t) - \gamma \overline{u}(t)\overline{p}_r(t), \\ \dfrac{d\overline{p}_r}{dt}(t) = 0, \end{cases} \quad \forall t \in [0,T], \tag{7.24}$$

et la condition finale sur l'état adjoint est

$$\overline{p}(T) = (\overline{p}_a(T), \overline{p}_r(T)) = (0,-1)^*. \tag{7.25}$$

On a donc

$$\frac{d\overline{p}_a}{dt}(t) = -\varphi(\overline{u}(t))\overline{p}_a(t) + \gamma \overline{u}(t), \quad \overline{p}_r(t) \equiv -1, \quad \forall t \in [0,T]. \tag{7.26}$$

Par ailleurs, le Hamiltonien est autonome (cf. la Définition 7.2.1) et s'écrit sous la forme

$$H(x,p,u) = p_a \varphi(u) a + \gamma p_r u a. \tag{7.27}$$

La condition de minimisation (7.9) s'écrit, en utilisant le fait que $\overline{a}(t) \neq 0$ pour tout $t \in [0,T]$,

$$\overline{u}(t) \in \arg\min_{v\in[0,1]} \psi(t)v, \tag{7.28}$$

où la **fonction de commutation** est donnée par

$$\psi(t) = -\overline{p}_a(t)\alpha - \gamma. \tag{7.29}$$

La solution du problème de minimisation (7.28) est élémentaire ; on obtient, pour tout $t \in [0,T]$,

— si $\psi(t) > 0$, $\overline{u}(t) = 0$;
— si $\psi(t) = 0$, $\overline{u}(t) \in [0,1]$;
— si $\psi(t) < 0$, $\overline{u}(t) = 1$.

Le contrôle optimal est donc nécessairement bang-bang, sauf si $\overline{p}_a(t) = -\frac{\gamma}{\alpha}$ sur un sous-intervalle de temps de mesure strictement positive. Reprenons alors l'équation de l'état adjoint :

— si $\overline{p}_a(t) > -\frac{\gamma}{\alpha}$, $\overline{u}(t) = 1$, et on a $\frac{d}{dt}\overline{p}_a(t) = \beta\overline{p}_a(t) + \gamma \geq 0$, i.e., $t \mapsto \overline{p}_a(t)$ est croissante ;
— si $\overline{p}_a(t) < -\frac{\gamma}{\alpha}$, $\overline{u} = 0$, et on a $\frac{d}{dt}\overline{p}_a(t) = (\beta - \alpha)\overline{p}_a(t) \geq 0$, i.e., $t \mapsto \overline{p}_a(t)$ est encore croissante ;
— enfin, il ne peut exister d'intervalle de mesure strictement positive où $\overline{p}_a$ est constant et égal à $-\frac{\gamma}{\alpha}$; en effet, dans ces conditions, on aurait $\varphi(u(t))\frac{\gamma}{\alpha} + \gamma u(t) = 1 - \frac{\beta}{\alpha} \neq 0$, donc $\overline{p}_a$ ne pourrait pas être constant.

Nous pouvons maintenant terminer la résolution du problème. Au temps final, $\psi(T) = -\gamma < 0$, ce qui montre que $\overline{u}(T) = 1$, i.e., au temps final, le contrôle optimal consiste à fournir des reines (ce qui n'est pas très surprenant puisque l'objectif est d'en maximiser le nombre). Le point qui reste à préciser est s'il est optimal d'en fournir depuis l'instant initial ou s'il convient plutôt de laisser d'abord croître la population d'abeilles avant de commencer à en fournir. Comme la fonction de commutation est continue, il existe un temps $t_* < T$ tel que $\overline{u}(t) = 1$ sur $]t_*, T]$. Sur cet intervalle, on a $\frac{d\overline{p}_a}{dt}(t) = \beta\overline{p}_a(t) + \gamma$ et par ailleurs la condition finale sur $\overline{p}_a$ étant $\overline{p}_a(T) = 0$, on en déduit que

$$\overline{p}_a(t) = -\frac{\gamma}{\beta}\left(1 - e^{\beta(t-T)}\right), \quad \forall t \in [t_*, T]. \tag{7.30}$$

La fonction $\overline{p}_a$ est donnée par l'expression ci-dessus tant que le contrôle optimal $\overline{u}$ reste égal à 1. Pour que la valeur du contrôle change, la fonction de commutation (qui est continue) doit s'annuler, i.e., $\overline{p}_a(t_*) = -\frac{\gamma}{\alpha}$. En utilisant l'expression de $\overline{p}_a$, on obtient

$$t_* = \frac{1}{\beta}\ln(1 - \frac{\beta}{\alpha}) + T. \tag{7.31}$$

On notera que $t_* < T$. Deux cas peuvent alors se produire en fonction des paramètres du problème.

— **Cas 1.** $t_* < 0$ (ce qui correspond au cas d'un horizon temporel T petit) ; le contrôle optimal est alors $\overline{u} \equiv 1$ sur $[0,T]$, ce qui signifie que l'on fournit des reines en continu depuis $t = 0$ jusqu'à $t = T$;

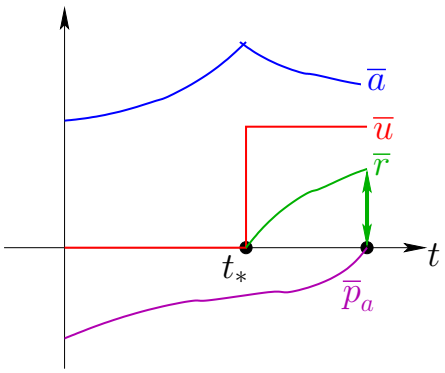

FIGURE 7.2 – Trajectoire, état adjoint et contrôle optimal pour le modèle de ruche.

— **Cas 2.** $t_* > 0$ (ce qui correspond au cas d'un horizon temporel T relativement grand) ; le contrôle optimal est $\overline{u} \equiv 0$ sur $[0, t_*[$ et $\overline{u} \equiv 1$ sur $]t_*, T]$. En effet, le contrôle $\overline{u}$ vérifie bien le PMP car $\frac{d\overline{p}_a}{dt}(t) = (\beta - \alpha)\overline{p}_a(t)$, $\overline{p}_a(t_*) = -\frac{\gamma}{\alpha}$, d'où $\overline{p}_a(t) = -\frac{\gamma}{\alpha}e^{(\beta-\alpha)(t-t_*)} < -\frac{\gamma}{\alpha}$ sur $[0, t_*]$, si bien que la fonction de commutation est positive, ce qui correspond bien à $\overline{u}(t) = 0$. L'ensemble $\{t \in [0,T] \mid \psi(t) = 0\}$ est réduit au singleton $\{t_*\}$ et est donc de mesure nulle.

Une illustration de la trajectoire, de l'état adjoint et du contrôle optimal est présentée à la Figure 7.2 dans le cas où il y a une commutation.

7.5 PMP : esquisse de preuve

Cette section est consacrée à une esquisse de preuve du Théorème 7.2.2 qui établissait le principe du minimum de Pontryaguine (PMP). On reprend le système de contrôle non-linéaire considéré à la Section 7.1. On rappelle que la dynamique s'écrit sous la forme

$$\dot{x}_u(t) = f(t, x_u(t), u(t)), \quad \forall t \in [0,T], \qquad x_u(0) = x_0, \tag{7.32}$$

avec $T > 0$, $f : [0,T] \times \mathbb{R}^d \times U \to \mathbb{R}^d$ et $x_0 \in \mathbb{R}^d$. L'ensemble des contrôles admissibles est

$$\mathcal{U} = L^1([0,T]; U),$$

où U est un sous-ensemble fermé non-vide de $\mathbb{R}^k$. L'objectif est de trouver un contrôle optimal $\overline{u} \in \mathcal{U}$ qui minimise le critère

$$J(u) = \int_0^T g(t, x_u(t), u(t))\,\mathrm{d}t + h(x_u(T)), \tag{7.33}$$

où les fonctions $g : [0,T] \times \mathbb{R}^d \times U \to \mathbb{R}$ et $h : \mathbb{R}^d \to \mathbb{R}$ sont données. Le problème de contrôle optimal est donc le suivant :

$$\text{Chercher } \overline{u} \in \mathcal{U} \text{ tel que } J(\overline{u}) = \inf_{u \in \mathcal{U}} J(u). \tag{7.34}$$

On rappelle les hypothèses qui avaient été introduites afin de garantir l'existence et l'unicité d'une trajectoire x_u pour un contrôle donné $u \in \mathcal{U}$ (cf. en particulier le Lemme 7.1.1) et le fait que la fonctionnelle $J(u)$ est bien définie :

1. $f \in C^0([0,T] \times \mathbb{R}^d \times U; \mathbb{R}^d)$ et f est de classe C^1 par rapport à x ;
2. $\exists C$, $|f(t,y,v)|_{\mathbb{R}^d} \leq C(1 + |y|_{\mathbb{R}^d} + |v|_{\mathbb{R}^k})$, $\forall t \in [0,T]$, $\forall y \in \mathbb{R}^d$, $\forall v \in U$;
3. Pour tout $R > 0$, $\exists C_R$, $|\frac{\partial f}{\partial x}(t,y,v)|_{\mathbb{R}^{d\times d}} \leq C_R(1 + |v|_{\mathbb{R}^d})$, $\forall t \in [0,T]$, $\forall y \in \overline{B}(0,R)$, $\forall v \in U$;
4. $g \in C^0([0,T] \times \mathbb{R}^d \times U; \mathbb{R})$ et g est de classe C^1 par rapport à x ; de plus, $h \in C^1(\mathbb{R}^d; \mathbb{R})$;
5. Pour tout $R > 0$, $\exists C_R$, $|g(t,y,v)| \leq C_R(1 + |v|_{\mathbb{R}^k})$, $\forall t \in [0,T]$, $\forall y \in \overline{B}(0,R)$, $\forall v \in U$;
6. Pour tout $R > 0$, $\exists C_R$, $|\frac{\partial g}{\partial x}(t,y,v)|_{\mathbb{R}^d} \leq C_R(1 + |v|_{\mathbb{R}^k})$, $\forall t \in [0,T]$, $\forall y \in \overline{B}(0,R)$, $\forall v \in U$;
7. Les fonctions g et h sont minorées respectivement sur $[0,T] \times \mathbb{R}^d \times U$ et sur $\mathbb{R}^d$.

Dans ces hypothèses, C et C_R désignent des constantes génériques indépendantes de (t,y,v), C_R dépendant du rayon R de la boule fermée $\overline{B}(0,R)$; comme précédemment, nous continuons à utiliser les symboles C et C_R avec la convention que les valeurs de C et de C_R peuvent changer à chaque utilisation tant qu'ils restent indépendants du temps, de l'état du système et de la valeur du contrôle.

Rappelons enfin l'énoncé du PMP (cf. le Théorème 7.2.2).

Théorème 7.5.1 (PMP) *Si $\overline{u} \in \mathcal{U}$ est un contrôle optimal, i.e., si $\overline{u}$ est une solution de (7.34), alors, en notant $\overline{x} = x_{\overline{u}} \in AC([0,T]; \mathbb{R}^d)$ la trajectoire associée à $\overline{u}$, et en définissant l'**état adjoint** $\overline{p} \in AC([0,T]; \mathbb{R}^d)$ solution de*

$$\frac{d\overline{p}}{dt}(t) = -\overline{A}(t)^* \overline{p}(t) - \overline{b}(t), \quad \forall t \in [0,T], \qquad \overline{p}(T) = \frac{\partial h}{\partial x}(\overline{x}(T)), \tag{7.35}$$

avec $\overline{A}(t) = \frac{\partial f}{\partial x}(t, \overline{x}(t), \overline{u}(t)) \in \mathbb{R}^{d\times d}$ et $\overline{b}(t) = \frac{\partial g}{\partial x}(t, \overline{x}(t), \overline{u}(t)) \in \mathbb{R}^d$ pour tout $t \in [0,T]$, on a, p.p. $t \in [0,T]$,

$$\overline{u}(t) \in \arg\min_{v \in U} H(t, \overline{x}(t), \overline{p}(t), v), \tag{7.36}$$

*où le **Hamiltonien** $H : [0,T] \times \mathbb{R}^d \times \mathbb{R}^d \times U \to \mathbb{R}$ est défini par*

$$H(t,x,p,u) = p^* f(t,x,u) + g(t,x,u). \tag{7.37}$$

On rappelle enfin qu'un triplet $(\overline{x}, \overline{p}, \overline{u})$ satisfaisant les conditions ci-dessus est appelé une **extrémale** et que le PMP ne fournit qu'une **condition nécessaire d'optimalité** ; en revanche, il ne dit rien sur l'existence d'un contrôle optimal et il ne fournit pas *a priori* de condition suffisante.

Démonstration. Nous allons nous contenter de donner une esquisse de la preuve, en insistant sur les idées principales sans nécessairement fournir tous les détails techniques pour certains résultats intermédiaires. Ce qui compte ici est donc davantage

l'esprit de la démonstration que sa lettre.
(1) L'idée fondamentale est de tester l'optimalité de $J(\overline{u})$ en faisant des **variations aiguille** : il s'agit de perturbations de $\overline{u}$ d'ordre un (!) mais sur un intervalle de temps de longueur très petite $\delta \ll 1$. Soit $t \in [0,T[$ et $\delta \in]0,T-t[$, avec $\delta \ll 1$. La perturbation reste donc petite dans $L^1([0,T];\mathbb{R}^k)$. Soit $v \in U$ arbitraire. On pose $I_\delta = [t,t+\delta]$ et on considère le contrôle perturbé

$$u_\delta(t) = \begin{cases} \overline{u}(t), & \forall t \in [0,T] \setminus I_\delta, \\ v, & \forall t \in I_\delta. \end{cases}$$

On note x_δ la trajectoire associée au contrôle perturbé. On admet par la suite que p.p. $t \in [0,T[$ (de tels points sont appelés points de Lebesgue), pour $\psi = f$ et $\psi = g$,

$$\lim_{\delta \to 0^+} \frac{1}{\delta} \int_{I_\delta} \psi(s, \overline{x}(s), \overline{u}(s))\, \mathrm{d}s = \psi(t, \overline{x}(t), \overline{u}(t)).$$

On suppose dans la suite de la preuve que t est un point de Lebesgue ; le résultat ci-dessus justifie donc que l'on considère bien tous les instants $t \in [0,T]$ à un sous-ensemble de mesure nulle près.
(2) Comparaison des trajectoires. Comme $x_\delta(t) = \overline{x}(t)$ et $x_\delta(s) = \overline{x}(s) + O(\delta)$ pour tout $s \in I_\delta$, on peut invoquer la continuité de f en (t,x) et la propriété des points de Lebesgue afin d'obtenir les estimations suivantes :

$$\begin{aligned} x_\delta(t+\delta) &= \overline{x}(t) + \int_{I_\delta} f(s, x_\delta(s), v)\, \mathrm{d}s = \overline{x}(t) + \delta f(t, \overline{x}(t), v) + o(\delta), \\ \overline{x}(t+\delta) &= \overline{x}(t) + \int_{I_\delta} f(s, \overline{x}(s), \overline{u}(s))\, \mathrm{d}s = \overline{x}(t) + \delta f(t, \overline{x}(t), \overline{u}(t)) + o(\delta), \end{aligned}$$

si bien que

$$x_\delta(t+\delta) - \overline{x}(t+\delta) = \delta(f(t, \overline{x}(t), v) - f(t, \overline{x}(t), \overline{u}(t))) + o(\delta).$$

Une illustration est présentée à la Figure 7.3. On va maintenant comparer $x_\delta(s)$ et $\overline{x}(s)$ pour tout $s \in I_\delta^+ = [t+\delta,T]$. Il est clair que $x_\delta(s) - \overline{x}(s) = O(\delta)$ pour tout $s \in I_\delta^+$, et on cherche à préciser la différence à l'ordre un en δ. On introduit la solution $y_\delta \in AC(I_\delta^+;\mathbb{R}^d)$ de l'équation différentielle

$$\dot{y}_\delta(s) = \overline{A}(s) y_\delta(s), \quad \forall s \in I_\delta^+, \qquad y_\delta(t+\delta) = f(t, \overline{x}(t), v) - f(t, \overline{x}(t), \overline{u}(t)),$$

où on rappelle que $\overline{A}(s) = \frac{\partial f}{\partial x}(s, \overline{x}(s), \overline{u}(s))$. On en déduit que

$$x_\delta(s) - \overline{x}(s) = \delta y_\delta(s) + \Phi_\delta(s), \quad \forall s \in I_\delta^+, \qquad \Phi_\delta = o(\delta) \text{ unif. sur } I_\delta^+.$$

En effet, on a vu que $\Phi_\delta(t+\delta) = o(\delta)$ et $\dot{\Phi}_\delta(s) = \Psi_\delta(s) + \overline{A}(s)\Phi_\delta(s)$, pour tout $s \in I_\delta^+$, où $\Psi_\delta(s) = o(s)$ uniformément sur I_δ^+, car

$$\Psi_\delta(s) = f(s, x_\delta(s), \overline{u}(s)) - f(s, \overline{x}(s), \overline{u}(s)) - \overline{A}(s)(x_\delta(s) - \overline{x}(s)).$$

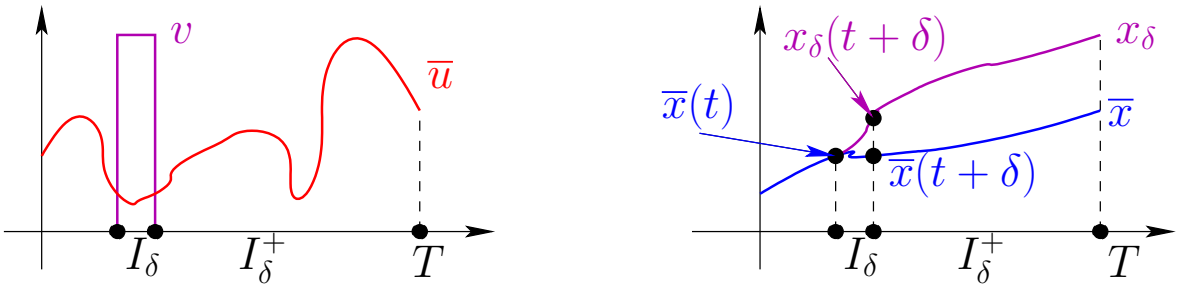

FIGURE 7.3 – Principe de la variation aiguille pour le contrôle optimal $\overline{u}$ (à gauche), trajectoire optimale et trajectoire perturbée (à droite).

En conclusion de cette première étape de la preuve, on a donc

$$x_\delta(s) - \overline{x}(s) = \delta y_\delta(s) + o(\delta) \text{ unif. sur } I_\delta^+.$$

(3) Comparaison des critères. Grâce à la comparaison des trajectoires, à la continuité de g en (t,x) et à la propriété des points de Lebesgue, il vient

$$\begin{aligned}
J(u_\delta) - J(\overline{u}) &= \int_t^T g(s,x_\delta(s),u_\delta(s)) - g(s,\overline{x}(s),\overline{u}(s))\,\mathrm{d}s + h(x_\delta(T)) - h(\overline{x}(T)) \\
&= \int_{I_\delta} g(s,x_\delta(s),v) - g(s,\overline{x}(s),\overline{u}(s))\,\mathrm{d}s + \int_{I_\delta^+} g(s,x_\delta(s),\overline{u}(s)) - g(s,\overline{x}(s),\overline{u}(s))\,\mathrm{d}s \\
&\quad + \delta \frac{\partial h}{\partial x}(\overline{x}(T))^* y_\delta(T) + o(\delta) \\
&= \delta(g(t,\overline{x}(t),v) - g(t,\overline{x}(t),\overline{u}(t))) + \delta \int_{t+\delta}^T \overline{b}(s)^* y_\delta(s)\,\mathrm{d}s \\
&\quad + \delta \frac{\partial h}{\partial x}(\overline{x}(T))^* y_\delta(T) + o(\delta),
\end{aligned}$$

où on rappelle que $\overline{b}(s) = \frac{\partial g}{\partial x}(s,\overline{x}(s),\overline{u}(s))$. L'optimalité de $\overline{u}$ implique donc que

$$0 \le g(t,\overline{x}(t),v) - g(t,\overline{x}(t),\overline{u}(t)) + \int_{t+\delta}^T \overline{b}(s)^* y_\delta(s)\,\mathrm{d}s + \frac{\partial h}{\partial x}(\overline{x}(T))^* y_\delta(T) + o(1).$$

(4) Introduction de l'état adjoint et conclusion. L'état adjoint $\overline{p}$, qui est par définition tel que $\frac{d\overline{p}}{dt}(s) = -\overline{A}(s)^*\overline{p}(s) - \overline{b}(s)$ sur $[0,T]$ et $\overline{p}(T) = \frac{\partial h}{\partial x}(\overline{x}(T))$, nous permet d'éliminer la fonction y_δ. En effet, il vient

$$\begin{aligned}
\int_{t+\delta}^T \overline{b}(s)^* y_\delta(s)\,\mathrm{d}s + \frac{\partial h}{\partial x}(\overline{x}(T))^* y_\delta(T) &= \int_{t+\delta}^T \left(-\frac{d\overline{p}}{dt}(s) - \overline{A}(s)^*\overline{p}(s) \right)^* y_\delta(s)\,\mathrm{d}s + \overline{p}(T)^* y_\delta(T) \\
&= \int_{t+\delta}^T -\frac{d\overline{p}}{dt}(s)^* y_\delta(s)\,\mathrm{d}s + \overline{p}(T)^* y_\delta(T) \\
&= \overline{p}(t+\delta)^* y_\delta(t+\delta) \\
&= \overline{p}(t+\delta)^*(f(t,\overline{x}(t),v) - f(t,\overline{x}(t),\overline{u}(t))).
\end{aligned}$$

En faisant tendre $\delta \downarrow 0$, il vient par continuité de $\overline{p}$,

$$0 \leq g(t, \overline{x}(t), v) - g(t, \overline{x}(t), \overline{u}(t)) + \overline{p}(t)^*(f(t, \overline{x}(t), v) - f(t, \overline{x}(t), \overline{u}(t))),$$

et en utilisant la définition du Hamiltonien, on obtient

$$0 \leq H(t, \overline{x}(t), \overline{p}(t), v) - H(t, \overline{x}(t), \overline{p}(t), \overline{u}(t)),$$

ce qui conclut la preuve car v est arbitraire dans U. □

Chapitre 8

ANNEXE : QUELQUES RAPPELS MATHÉMATIQUES

8.1 Rappels sur les espaces de Hilbert

Nous rappelons brièvement quelques propriétés des espaces de Hilbert (pour plus de détails, nous renvoyons au cours de mathématiques [16]). Pour simplifier la présentation, on ne considère que le cas d'espaces de Hilbert sur $\mathbb{R}$.

Définition 8.1.1 *Un espace de Hilbert réel est un espace vectoriel sur $\mathbb{R}$, muni d'un produit scalaire, noté $\langle x, y\rangle$, qui est complet pour la norme associée à ce produit scalaire, notée $\|x\| = \sqrt{\langle x, x\rangle}$. (On rappelle qu'un espace vectoriel normé est complet si toute suite de Cauchy est une suite convergente dont la limite appartient à cet espace.)*

Dans tout ce qui suit nous noterons V un espace de Hilbert réel, et $\langle x, y\rangle$ son produit scalaire associé.

Définition 8.1.2 *Un ensemble $K \subset V$ est dit convexe si, pour tout $x, y \in K$ et tout réel $\theta \in [0, 1]$, l'élément $(\theta x + (1 - \theta)y)$ appartient à K.*

Un résultat essentiel est le théorème de projection sur un ensemble convexe (voir le théorème 10.1 de [16]).

Théorème 8.1.3 (de projection sur un convexe) *Soit V un espace de Hilbert. Soit $K \subset V$ un convexe fermé non vide. Pour tout $x \in V$, il existe un unique $x_K \in K$ tel que*

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

De façon équivalente, x_K est caractérisé par la propriété

$$x_K \in K, \langle x_K - x, x_K - y\rangle \leq 0 \quad \forall y \in K. \tag{8.1}$$

On appelle x_K la projection orthogonale sur K de x.

Remarque 8.1.4 Le Théorème 8.1.3 permet de définir une application P_K, appelée opérateur de projection sur l'ensemble convexe K, en posant $P_K x = x_K$. On vérifie sans peine que P_K est continue et faiblement contractante, c'est-à-dire que

$$\|P_K x - P_K y\| \leq \|x - y\| \quad \forall\, x, y \in V\ . \tag{8.2}$$

•

Remarque 8.1.5 Un cas particulier de convexe fermé K est un sous-espace vectoriel fermé W. Dans ce cas, la caractérisation (8.1) de x_W devient

$$x_W \in W, \langle x_W - x, z\rangle = 0 \quad \forall z \in W.$$

En effet, dans (8.1) il suffit de prendre $y = x_K \pm z$ avec z quelconque dans W. •

Démonstration. Soit y^n une suite minimisante, c'est-à-dire que $y^n \in K$ vérifie

$$d_n = \|x - y^n\| \to d = \inf_{y\in K} \|x - y\| \text{ quand } n \to +\infty.$$

Montrons que y^n est une suite de Cauchy. En utilisant la symétrie du produit scalaire, il vient

$$\|x - \frac{1}{2}(y^n + y^p)\|^2 + \|\frac{1}{2}(y^n - y^p)\|^2 = \frac{1}{2}(d_n^2 + d_p^2).$$

Or, par convexité de K, $(y^n + y^p)/2 \in K$, et $\|x - \frac{1}{2}(y^n + y^p)\|^2 \geq d^2$. Par conséquent

$$\|y^n - y^p\|^2 \leq 2(d_n^2 + d_p^2) - 4d^2,$$

ce qui montre que y^n est une suite de Cauchy. Comme V est un espace de Hilbert, il est complet, donc la suite y^n est convergente vers une limite x_K. Par ailleurs, comme K est fermé, cette limite x_K appartient à K. Par conséquent, on a $d = \|x - x_K\|$. Comme toute la suite minimisante est convergente, la limite est forcément unique, et x_K est le seul point de minimum de $\min_{y\in K} \|x - y\|$.

Soit $x_K \in K$ ce point de minimum. Pour tout $y \in K$ et $\theta \in [0, 1]$, par convexité de K, $x_K + \theta(y - x_K)$ appartient à K et on a

$$\|x - x_K\|^2 \leq \|x - (x_K + \theta(y - x_K))\|^2.$$

En développant le terme de droite, il vient

$$\|x - x_K\|^2 \leq \|x - x_K\|^2 + \theta^2\|y - x_K\|^2 - 2\theta\langle x - x_K, y - x_K\rangle,$$

ce qui donne pour $\theta > 0$

$$0 \geq -2\langle x - x_K, y - x_K\rangle + \theta\|z\|^2.$$

En faisant tendre θ vers 0, on obtient la caractérisation (8.1). Réciproquement, soit x_K qui vérifie cette caractérisation. Pour tout $y \in K$ on a

$$\|x - y\|^2 = \|x - x_K\|^2 + \|x_K - y\|^2 + 2\langle x - x_K, x_K - y\rangle \geq \|x - x_K\|^2,$$

ce qui prouve que x_K est bien la projection orthogonale de x sur K. □

Définition 8.1.6 *Soit V un espace de Hilbert pour le produit scalaire $\langle , \rangle$. On appelle base hilbertienne (dénombrable) de V une famille dénombrable $(e_n)_{n\geq 1}$ d'éléments de V qui est orthonormale pour le produit scalaire et telle que l'espace vectoriel engendré par cette famille est dense dans V.*

Remarque 8.1.7 L'existence d'une base hilbertienne dénombrable n'est pas garantie pour tous les espaces de Hilbert. Néanmoins, on peut construire des bases hilbertiennes pour les espaces de Hilbert séparables (i.e. qui contiennent une famille dénombrable dense). •

Proposition 8.1.8 *Soit V un espace de Hilbert pour le produit scalaire $\langle , \rangle$. Soit $(e_n)_{n\geq 1}$ une base hilbertienne de V. Pour tout élément x de V, il existe une unique suite $(x_n)_{n\geq 1}$ de réels telle que la somme partielle $\sum_{n=1}^{p} x_n e_n$ converge vers x quand p tend vers l'infini, et cette suite est définie par $x_n = \langle x, e_n\rangle$. De plus, on a*

$$\|x\|^2 = \langle x, x\rangle = \sum_{n\geq 1} |\langle x, e_n\rangle|^2. \tag{8.3}$$

On écrit alors

$$x = \sum_{n\geq 1} \langle x, e_n\rangle e_n.$$

Démonstration. S'il existe une suite $(x_n)_{n\geq 1}$ de réels telle que $\lim_{p\to+\infty} \sum_{n=1}^{p} x_n e_n = x$, alors par projection sur e_n (et comme cette suite est par définition indépendante de p) on a $x_n = \langle x, e_n\rangle$, ce qui prouve l'unicité de la suite $(x_n)_{n\geq 1}$. Montrons maintenant son existence. Par définition d'une base hilbertienne, pour tout $x \in V$ et pour tout $\epsilon > 0$, il existe y, combinaison linéaire finie des $(e_n)_{n\geq 1}$, tel que $\|x - y\| < \epsilon$. Grâce au Théorème 8.1.3 on peut définir une application linéaire S_p qui, à tout point $z \in V$, fait correspondre $S_p z = z_W$, où z_W est la projection orthogonale sur le sous-espace vectoriel W engendré par les p premiers vecteurs $(e_n)_{1\leq n\leq p}$. En vertu de (8.1), $(z - S_p z)$ est orthogonal à tout élément de W, donc en particulier à $S_p z$. On en déduit que

$$\|z\|^2 = \|z - S_p z\|^2 + \|S_p z\|^2, \tag{8.4}$$

ce qui implique

$$\|S_p z\| \leq \|z\| \forall z \in V.$$

Comme $S_p z$ est engendré par les $(e_n)_{1\leq n\leq p}$, et que $(z - S_p z)$ est orthogonal à chacun des $(e_n)_{1\leq n\leq p}$, on vérifie facilement que

$$S_p z = \sum_{n=1}^{p} \langle z, e_n\rangle e_n.$$

Pour p suffisamment grand, on a $S_p y = y$ car y est une combinaison linéaire finie des $(e_n)_{n\geq 1}$. Par conséquent

$$\|S_p x - x\| \leq \|S_p(x - y)\| + \|y - x\| \leq 2\|x - y\| \leq 2\epsilon.$$

On en déduit la convergence de $S_p x$ vers x. De cette convergence et de l'équation (8.4) on tire

$$\lim_{p\to+\infty} \|S_p x\|^2 = \|x\|^2,$$

qui n'est rien d'autre que la formule de sommation (8.3), dite de Parseval. □

Définition 8.1.9 *Soit V et W deux espaces de Hilbert réels. Une application linéaire A de V dans W est dite continue s'il existe une constante C telle que*

$$\|Ax\|_W \leq C\|x\|_V \quad \forall x \in V.$$

La plus petite constante C qui vérifie cette inégalité est la norme de l'application linéaire A, autrement dit

$$\|A\| = \sup_{x\in V, x\neq 0} \frac{\|Ax\|_W}{\|x\|_V}.$$

Souvent on utilisera la dénomination équivalente d'opérateur au lieu d'application entre espaces de Hilbert (on parlera ainsi d'opérateur linéaire continu plutôt que d'application linéaire continue). Si V est de dimension finie, alors toutes les applications linéaires de V dans W sont continues, mais ce n'est plus vrai si V est de dimension infinie.

Définition 8.1.10 *Soit V un espace de Hilbert réel. Son dual V' est l'ensemble des formes linéaires* **continues** *sur V, c'est-à-dire l'ensemble des applications linéaires continues de V dans $\mathbb{R}$. Par définition, la norme d'un élément $L \in V'$ est*

$$\|L\|_{V'} = \sup_{x\in V, x\neq 0} \frac{|L(x)|}{\|x\|}.$$

Dans un espace de Hilbert la dualité a une interprétation très simple grâce au théorème de Riesz (voir le théorème 10.3 de [16]) qui permet d'identifier un espace de Hilbert à son dual par isomorphisme.

Théorème 8.1.11 (de représentation de Riesz) *Soit V un espace de Hilbert réel, et soit V' son dual. Pour toute forme linéaire continue $L \in V'$ il existe un unique $y \in V$ tel que*

$$L(x) = \langle y, x\rangle \quad \forall x \in V.$$

De plus, on a $\|L\|_{V'} = \|y\|$.

Démonstration. Soit $M = \mathrm{Ker} L$. Il s'agit d'un sous-espace fermé de V car L est continue. Si $M = V$, alors L est identiquement nulle et seul $y = 0$ convient. Si $M \neq V$, alors il existe $z \in V \setminus M$. Soit alors $z_M \in M$ sa projection sur M. Comme z n'appartient pas à M, $z - z_M$ est non nul et, par le Théorème 8.1.3, est orthogonal à tout élément de M. Soit finalement

$$z_0 = \frac{z - z_M}{\|z - z_M\|}.$$

Tout vecteur $x \in V$ peut s'écrire

$$x = w + \lambda z_0 \text{ avec } \lambda = \frac{L(x)}{L(z_0)}.$$

On vérifie aisément que $L(w) = 0$, donc $w \in M$. Ceci prouve que $V = \text{Vect}(z_0) \oplus M$. Par définition de z_M et de z_0, on a $\langle w, z_0 \rangle = 0$, ce qui implique

$$L(x) = \langle x, z_0 \rangle L(z_0),$$

d'où le résultat désiré avec $y = L(z_0)z_0$ (l'unicité est évidente). D'autre part, on a

$$\|y\| = |L(z_0)|,$$

et

$$\|L\|_{V'} = \sup_{x \in V, x \neq 0} \frac{|L(x)|}{\|x\|} = L(z_0) \sup_{x \in V, x \neq 0} \frac{\langle x, z_0 \rangle}{\|x\|}.$$

Le maximum dans le dernier terme de cette égalité est atteint par $x = z_0$, ce qui implique que $\|L\|_{V'} = \|y\|$. □

Un résultat essentiel pour pouvoir démontrer le Lemme de Farkas 2.5.18 (utile en optimisation) est la propriété géométrique suivante qui est tout à fait conforme à l'intuition.

Théorème 8.1.12 (Séparation d'un point et d'un convexe) *Soit K une partie convexe non vide et fermée d'un espace de Hilbert V, et $x_0 \notin K$. Alors il existe un hyperplan fermé de V qui sépare strictement x_0 et K, c'est-à-dire qu'il existe une forme linéaire $L \in V'$ et $\alpha \in \mathbb{R}$ tels que*

$$L(x_0) < \alpha < L(x) \quad \forall\, x \in K\ . \tag{8.5}$$

Démonstration. Notons x_K la projection de x_0 sur K. Puisque $x_0 \notin K$, on a $x_K - x_0 \neq 0$. Soit L la forme linéaire définie pour tout $y \in V$ par $L(y) = \langle x_K - x_0, y \rangle$, et soit $\alpha = (L(x_k) + L(x_0))/2$. D'après (8.1), on a $L(x) \geq L(x_K) > \alpha > L(x_0)$ pour tout $x \in K$, ce qui achève la démonstration. □

Nous aurons enfin besoin pour démontrer le Théorème de Minkowski 4.3.2 d'une variante du théorème de séparation, faisant intervenir la notion importante d'hyperplan d'appui. Si K est un convexe d'un espace de Hilbert V, on appelle **hyperplan d'appui** de K en un point x un hyperplan affine $H = \{y \in V \mid L(y) = \alpha\}$, avec $L \in V'$, $L \neq 0$, et $\alpha \in \mathbb{R}$, tel que $\alpha = L(x) \leq L(y)$, pour tout $y \in C$.

Corollaire 8.1.13 (Hyperplan d'appui) *Il existe un hyperplan d'appui en tout point frontière d'un convexe fermé K d'un espace de Hilbert de dimension finie.*

Démonstration. Soit x un point frontière de K : il existe alors une suite $x_n \in V \setminus K$, avec $x_n \to x$. Le Théorème de séparation 8.1.12 fournit pour tout n une forme linéaire L_n non nulle telle que $L_n(x_n) \leq L_n(y)$ pour tout $y \in K$. On peut choisir L_n de norme

1. Comme V est de dimension finie, la sphère unité de V' est compacte, et quitte à remplacer L_n par une sous-suite, on peut supposer que L_n converge vers une forme linéaire L, qui est non nulle, car de norme 1. Il suffit maintenant de passer à la limite dans $L_n(x_n) \leq L_n(y)$, ce que l'on justifie en écrivant $L_n(x_n) = L_n(x_n - x) + L_n(x)$ et en notant que $|L_n(x_n - x)| \leq \|L_n\| \|x_n - x\| = \|x_n - x\|$, pour obtenir $L(x) \leq L(y)$ quel que soit $y \in K$. Ainsi, $H = \{y \in V \mid L(y) = L(x)\}$ est un hyperplan d'appui de K en x. □

Remarque 8.1.14 La preuve du Corollaire 8.1.13 ne s'étend pas en dimension infinie : dans ce cas, la suite L_n a bien une valeur d'adhérence L pour la topologie faible, mais rien ne dit que $L \neq 0$. Comme contre exemple, considérons l'ensemble K des suites de ℓ_2 à termes positifs ou nuls, qui est un convexe fermé de ℓ_2 d'intérieur vide. Tout point de K est donc point frontière, mais si x est une suite de ℓ_2 à termes strictement positifs, il n'existe pas d'hyperplan d'appui en x. •

8.2 Notion de sélection mesurable

L'objectif de cette section est d'apporter quelques compléments sur les résultats de sélection mesurable qui sont parfois invoqués dans ce cours pour justifier de manière mathématiquement rigoureuse certains résultats de contrôle optimal. Ces résultats font appel à des notions relativement fines de théorie de la mesure, et ne seront donc qu'esquissés ici. Une présentation complète peut être trouvée dans le chapitre 14 du livre [24]. Le contenu de cette section est inspiré de ce chapitre.

Commençons par présenter la problématique. On pose $I = [0,T]$. On considère une application $\Phi : [0,T] \times \mathbb{R}^k \to \overline{\mathbb{R}} = [-\infty, +\infty]$. Pour tout $t \in I$, on considère le sous-ensemble

$$\overline{U}(t) = \arg\min_{u \in \mathbb{R}^k} \Phi(t,u) \subset \mathbb{R}^k, \tag{8.6}$$

et on pose $J = \{t \in I \mid \overline{U}(t) \neq \emptyset\}$. On souhaite savoir s'il existe une application $\overline{u} : J \to \mathbb{R}^k$ qui soit **mesurable** et telle que $\overline{u}(t) \in \overline{U}(t)$ pour tout $t \in J$. Une telle application est appelée une **sélection mesurable**. Un résultat simple et utile est que si l'application Φ est **mesurable** par rapport à t (à u fixé) et si elle est **convexe et continue** par rapport à u (à t fixé), alors il existe une telle sélection mesurable.

Le reste de cette section a pour objectif d'apporter une réponse mathématique un peu plus complète au problème de la sélection mesurable. Dans un premier temps, on considère des applications définies sur I à valeurs dans les sous-ensembles de $\mathbb{R}^k$. On note $S : I \rightrightarrows \mathbb{R}^k$ une telle application (le symbole $\rightrightarrows$ est là pour nous rappeler que $S(t)$ est un sous-ensemble de $\mathbb{R}^k$ qui n'est pas forcément réduit à un point). On équipe I d'une σ-algébre notée $\mathcal{A}$ (par exemple, la tribu borélienne de $\mathbb{R}$ restreinte à I).

Définition 8.2.1 (Mesurabilité) *On dit que l'application $S : I \rightrightarrows \mathbb{R}^k$ est mesurable si pour tout ouvert $O \subset \mathbb{R}^k$, l'image réciproque*

$$S^{-1}(O) = \bigcup_{u \in O} S^{-1}(u) = \{t \in I \mid S(t) \cap O \neq \emptyset\} \tag{8.7}$$

est mesurable, i.e., si $S^{-1}(O) \in \mathcal{A}$. *En particulier, le domaine de* S, $\operatorname{dom} S = S^{-1}(\mathbb{R}^k)$, *est donc mesurable (on notera que si* $S(t) = \emptyset$, *alors* $t \notin \operatorname{dom} S$*).*

Si l'application S ne prend comme valeurs que des singletons, on retrouve la définition usuelle de la mesurabilité d'une application de I dans $\mathbb{R}^k$.

Théorème 8.2.2 (Représentation de Castaing) *La mesurabilité d'une application* $S : I \rightrightarrows \mathbb{R}^k$ *à* ***valeurs fermées*** *(cela signifie que pour tout* $t \in I$, $S(t)$ *est un fermé) est équivalente à l'existence d'une représentation de Castaing, i.e., à l'existence d'une famille dénombrable de fonctions mesurables* $s_n : \operatorname{dom} S \to \mathbb{R}^k$, $\forall n \in \mathbb{N}$, *telles que pour tout* $t \in \operatorname{dom} S$, $S(t) = \overline{\{s_n(t)\}_{n\in\mathbb{N}}}$.

Corollaire 8.2.3 (Sélection mesurable) *Une application* $S : I \rightrightarrows \mathbb{R}^k$ *mesurable à valeurs fermées admet une sélection mesurable, i.e., il existe une application mesurable* $s : \operatorname{dom} S \to \mathbb{R}^k$ *telle que* $s(t) \in S(t)$ *pour tout* $t \in \operatorname{dom} S$.

Considérons à nouveau une application $\Phi : [0,T] \times \mathbb{R}^k \to \overline{\mathbb{R}}$. L'application-épigraphe $\mathcal{E}_\Phi : I \rightrightarrows \mathbb{R}^k \times \mathbb{R}$ et l'application-domaine $\mathcal{D}_\Phi : I \rightrightarrows \mathbb{R}^k$, associées à Φ, sont telles que, pour tout $t \in I$,

$$\mathcal{E}_\Phi(t) = \{(u,\alpha) \in \mathbb{R}^k \times \mathbb{R} \mid \Phi(t,u) \le \alpha\}, \tag{8.8a}$$

$$\mathcal{D}_\Phi(t) = \{u \in \mathbb{R}^k \mid \Phi(t,u) < +\infty\}. \tag{8.8b}$$

Définition 8.2.4 (Intégrande normal) *On dit que l'application* $\Phi : [0,T] \times \mathbb{R}^k \to \overline{\mathbb{R}}$ *est un intégrande normal si son application-épigraphe* $\mathcal{E}_\Phi : I \rightrightarrows \mathbb{R}^k \times \mathbb{R}$ *est mesurable à valeurs fermées.*

Proposition 8.2.5 (Ensembles de niveau) *L'application* $\Phi : [0,T] \times \mathbb{R}^k \to \overline{\mathbb{R}}$ *est un intégrande normal si et seulement si pour tout* $\alpha \in \overline{\mathbb{R}}$, *l'application ensemble de niveau* $N_\alpha : I \rightrightarrows \mathbb{R}^k$ *telle que* $N_\alpha(t) = \{u \in \mathbb{R}^k \mid \Phi(t,u) \le \alpha\}$ *est mesurable à valeurs fermées.*

On rappelle qu'une fonction $f : \mathbb{R}^k \to \overline{\mathbb{R}}$ est semi-continue inférieurement (sci en abrégé) si son épigraphe $\{(u,\alpha) \in \mathbb{R}^k \times \mathbb{R} \mid f(u) \le \alpha\}$ est fermé ; de manière équivalente, pour tout $u \in \mathbb{R}^k$ et tout $\epsilon > 0$, il existe un voisinage U de u tel que pour tout $v \in U$, on a $f(v) \ge f(u) - \epsilon$.

Proposition 8.2.6 (Conséquences de la normalité d'un intégrande) *On suppose que l'application* $\Phi : [0,T] \times \mathbb{R}^k \to \overline{\mathbb{R}}$ *est un intégrande normal. Alors,*
(i) *l'application-domaine* $\mathcal{D}_\Phi : I \rightrightarrows \mathbb{R}^k$ *est mesurable ;*
(ii) *pour toute fonction mesurable* $I \ni t \mapsto u(t) \in \mathbb{R}^k$, *la fonction* $t \mapsto \Phi(t,u(t))$ *est mesurable ;*
(iii) *l'application* Φ *est mesurable par rapport à* t *(à* u *fixé) et elle est sci par rapport à* u *(à* t *fixé) ; en revanche, toute application qui est mesurable par rapport à* t *et sci par rapport* u *n'est pas nécessairement un intégrande normal.*

Proposition 8.2.7 (Fonction de Carathéodory) *Toute fonction de Carathéodory, i.e., toute fonction qui est mesurable par rapport à* t *(à* u *fixé) et continue par rapport à* u *(à* t *fixé) est un intégrande normal.*

Exemple 8.2.8 [Indicatrice] On suppose que l'application $S : I \rightrightarrows \mathbb{R}^k$ est mesurable et à valeurs fermées. Alors, la fonction indicatrice $\delta_S : I \times \mathbb{R}^k \to \overline{\mathbb{R}}$ telle que

$$\delta_S(t,u) = \begin{cases} 0 & \text{si } u \in S(t), \\ +\infty & \text{sinon,} \end{cases}$$

est un intégrande normal.

Venons-en au résultat principal lié à la notion d'intégrande normal.

Théorème 8.2.9 (Mesurabilité de minimiseurs et du minimum) *On suppose que l'application $\Phi : [0,T] \times \mathbb{R}^k \to \overline{\mathbb{R}}$ est un intégrande normal. On pose pour tout $t \in I$,*

$$\varphi(t) = \inf_{u \in \mathbb{R}^k} \Phi(t,u), \quad \overline{U}(t) = \arg\min_{u \in \mathbb{R}^k} \Phi(t,u). \tag{8.9}$$

Alors, l'application $\varphi : I \to \overline{\mathbb{R}}$ est mesurable et l'application $\overline{U} : I \rightrightarrows \mathbb{R}^k$ est mesurable à valeurs fermées. Par conséquent, le sous-ensemble $J = \{t \in I \mid \overline{U}(t) \neq \emptyset\} \subset I$ est mesurable et pour tout $t \in J$, on peut choisir un minimiseur $\overline{u}(t)$ dans $\overline{U}(t)$ de sorte que l'application $t \mapsto \overline{u}(t)$ soit mesurable.

Proposition 8.2.10 (Convexité) *Soit $\Phi : [0,T] \times \mathbb{R}^k \to \overline{\mathbb{R}}$ une application mesurable par rapport à t et sci par rapport à u. Alors, si Φ est convexe par rapport à u (à t fixé), Φ est un intégrande normal.*

8.3 Rappels sur les équations différentielles ordinaires

Dans cette section nous rappelons quelques résultats importants dans l'étude des équations différentielles ordinaires. On fixe un horizon temporel $T > 0$ et une condition initiale $x_0 \in \mathbb{R}^d$. On considère le **problème de Cauchy** qui consiste à chercher une fonction $x : [0,T] \to \mathbb{R}^d$ telle que

$$\dot{x}(t) = F(t,x(t)), \quad \forall t \in [0,T], \qquad x(0) = x_0, \tag{8.10}$$

pour une application donnée $F : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$. Commençons par rappeler un résultat bien connu.

Théorème 8.3.1 (Cauchy–Lipschitz, cas continu et Lipschitz global) *On suppose que :*
(i) *L'application F est **continue** en t et en x, i.e., $F \in C^0([0,T] \times \mathbb{R}^d; \mathbb{R}^d)$;*
(ii) *L'application F est **globalement lipschitzienne** en x, i.e.,*

$$\exists C_0 \in \mathbb{R}_+, \quad \forall t \in [0,T],\ \forall x_1, x_2 \in \mathbb{R}^d, \quad |F(t,x_1) - F(t,x_2)|_{\mathbb{R}^d} \leq C_0 |x_1 - x_2|_{\mathbb{R}^d}. \tag{8.11}$$

Alors, il existe une unique solution au problème de Cauchy telle que

$$x \in C^1([0,T]; \mathbb{R}^d). \tag{8.12}$$

Cette solution satisfait donc le système différentiel (8.10) *pour tout $t \in [0,T]$.*

Démonstration. Le principe de la preuve consiste à observer que x est solution du problème de Cauchy (8.10) si et seulement si

$$x(t) = x_0 + \int_0^t F(s, x(s))\,\mathrm{d}s, \quad \forall t \in [0,T].$$

On introduit l'espace $Y = C^0([0,T];\mathbb{R}^d)$; il s'agit d'un espace de Banach (espace vectoriel normé complet) équipé de la norme de la convergence uniforme $\|y\|_Y = \sup_{t\in[0,T]} |y(t)|_{\mathbb{R}^d}$ pour tout $y \in Y$. Résoudre le problème de Cauchy revient à chercher un point fixe de l'application $\Phi : Y \to Y$ où pour tout $y \in Y$, $\Phi(y)$ est tel que

$$\Phi(y)(t) = x_0 + \int_0^t F(s, y(s))\,\mathrm{d}s, \quad \forall t \in [0,T].$$

Montrons que l'application Φ est strictement contractante de Y dans Y. On considère la norme $\|y\|_{Y*} = \sup_{t\in[0,T]}(e^{-C_0 t}|y(t)|_{\mathbb{R}^d})$ où C_0 est la constante intervenant dans la propriété de Lipschitz globale de l'application F. Il est clair que la norme $\|\cdot\|_{Y*}$ est équivalente à la norme $\|\cdot\|_Y$ sur Y. On constate que pour tout $y_1, y_2 \in Y$, on a

$$\begin{aligned}
\|\Phi(y_1) - \Phi(y_2)\|_{Y*} &= \sup_{t\in[0,T]} \left(e^{-C_0 t} |\Phi(y_1)(t) - \Phi(y_2)(t)|_{\mathbb{R}^d} \right) \\
&\leq \sup_{t\in[0,T]} \left(e^{-C_0 t} \int_0^t |F(s, y_1(s)) - F(s, y_2(s))|_{\mathbb{R}^d}\,\mathrm{d}s \right) \\
&\leq \sup_{t\in[0,T]} \left(e^{-C_0 t} C_0 \int_0^t |y_1(s) - y_2(s)|_{\mathbb{R}^d}\,\mathrm{d}s \right) \\
&= \sup_{t\in[0,T]} \left(e^{-C_0 t} C_0 \int_0^t e^{C_0 s} e^{-C_0 s} |y_1(s) - y_2(s)|_{\mathbb{R}^d}\,\mathrm{d}s \right) \\
&\leq \left(\sup_{t\in[0,T]} e^{-C_0 t} C_0 \int_0^t e^{C_0 s}\,\mathrm{d}s \right) \|y_1 - y_2\|_{Y*} \\
&= \left(\sup_{t\in[0,T]} 1 - e^{-C_0 t} \right) \|y_1 - y_2\|_{Y*} = \left(1 - e^{-C_0 T}\right) \|y_1 - y_2\|_{Y*},
\end{aligned}$$

où on a utilisé le caractère globalement lipschitzien en x de l'application F pour passer de la deuxième à la troisième ligne du calcul. L'application Φ est donc bien strictement contractante de Y dans Y. On conclut par le théorème du point fixe de Picard. □

L'hypothèse de continuité en t de l'application F faite au Théorème 8.3.1 n'est pas vraiment satisfaisante pour l'étude des systèmes de contrôle. En effet, ces systèmes s'écrivent sous la forme

$$\dot{x}(t) = f(t, x(t), u(t)), \quad \forall t \in [0,T], \qquad x(0) = x_0, \tag{8.13}$$

où $u \in L^1([0,T];\mathbb{R}^k)$ et $f : [0,T] \times \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^d$. L'étude du système différentiel (8.13) se ramène à celle du problème de Cauchy (8.10) en posant

$$F(t,x) = f(t, x, u(t)), \quad \forall (t,x) \in [0,T] \times \mathbb{R}^d. \tag{8.14}$$

On voit donc que même si l'application f est régulière en u, le fait que le contrôle ne dépende pas continûment du temps fait que l'application F ne sera pas nécessairement continue en t. Afin de traiter cette situation, on dispose de la variante suivante du théorème 8.3.1 (la preuve utilise des arguments analogues à ceux évoqués ci-dessus). On renvoie le lecteur à la Définition 5.1.2 pour la notion de fonction absolument continue.

Théorème 8.3.2 (Cauchy–Lipschitz, cas mesurable et Lipschitz global) *On suppose que :*
(i) *L'application F est mesurable en t et continue en x, i.e., pour tout $x \in \mathbb{R}^d$, l'application $t \mapsto F(t,x)$ est mesurable et pour presque tout $t \in [0,T]$, l'application $x \mapsto F(t,x)$ est continue ;*
(ii) *L'application F est intégrable en t, i.e.,*

$$\forall x \in \mathbb{R}^d, \quad \exists \beta \in L^1([0,T];\mathbb{R}_+), \quad \forall t \in [0,T], \quad |F(t,x)|_{\mathbb{R}^d} \leq \beta(t); \tag{8.15}$$

(iii) *L'application F est* ***globalement lipschitzienne*** *en x, i.e.,*

$$\begin{aligned} &\exists C_0 \in L^1([0,T];\mathbb{R}_+), \\ &\text{p.p. } t \in [0,T],\ \forall x_1, x_2 \in \mathbb{R}^d, \quad |F(t,x_1) - F(t,x_2)|_{\mathbb{R}^d} \leq C_0(t)|x_1 - x_2|_{\mathbb{R}^d}. \end{aligned} \tag{8.16}$$

Alors, il existe une ***unique solution*** *au problème de Cauchy telle que*

$$x \in AC([0,T];\mathbb{R}^d). \tag{8.17}$$

Cette solution, qui est dérivable p.p. sur $[0,T]$, satisfait le système différentiel (8.10) *pour presque tout $t \in [0,T]$; elle vérifie également*

$$x(t) = x_0 + \int_0^t F(s,x(s))\,\mathrm{d}s, \quad \forall t \in [0,T]. \tag{8.18}$$

Remarque 8.3.3 [Intégrabilité] Grâce à la propriété (iii) du Théorème 8.3.2, il suffit, afin d'établir la propriété (ii), de montrer que $F(t,0) \in L^1([0,T];\mathbb{R}^d)$.

Un cas d'application du Théorème 8.3.2 est le cas linéaire (éventuellement avec un terme de dérive) où on a $F(t,x) = A(t)x + r(t)$ avec $A \in L^1([0,T];\mathbb{R}^{d\times d})$ et $r \in L^1([0,T];\mathbb{R}^d)$; l'application F est alors globalement lipschitzienne de constante $C_0(t) = |A(t)|_{\mathbb{R}^{d\times d}}$ (où $|\cdot|_{\mathbb{R}^{d\times d}}$ désigne la norme matricielle subordonnée à la norme euclidienne). Lorsque l'application F est non-linéaire en x, la propriété d'être globalement lipschitzienne est en général perdue. Dans ce cas, il est bien connu que la solution x du problème de Cauchy (8.10) peut exploser en temps fini.

Exemple 8.3.4 [Explosion en temps fini] Donnons un exemple simple d'explosion en temps fini. On se place dans $\mathbb{R}$ ($d = 1$) et on considère l'application $F(t,x) = 1 - x^2$ (qui ne dépend que de x). Le problème de Cauchy est donc $\dot{x}(t) = 1 - x(t)^2$ avec $x(0) = x_0 \in \mathbb{R}$. Si $|x_0| \leq 1$, il vient $x(t) = \tanh(t + t_0)$ avec $\tanh(t_0) = x_0$ et $\lim_{t\to\infty} x(t) = 1$; on a donc existence globale en temps de la solution. En revanche,

si $|x_0| > 1$, il vient $x(t) = \coth(t + t_0)$ avec $\coth(t_0) = x_0$ et deux situations peuvent se produire : (i) si $x_0 > 1$, alors $t_0 > 0$ et on a $\lim_{t\to\infty} x(t) = 1$, i.e., on a encore existence globale en temps de la solution ; (ii) si $x_0 < -1$, alors $t_0 < 0$ et dans ces conditions, $\lim_{t\uparrow t_0} |x(t)| = +\infty$; on a donc explosion en temps fini.

Remarque 8.3.5 [Non-unicité] Lorsque l'application F est uniquement continue en x, on peut ne pas avoir unicité de la solution du problème de Cauchy. Par exemple, pour le problème de Cauchy $\dot{x}(t) = \sqrt{|x(t)|}$ avec $x(0) = 0$ (i.e., pour $F(t,x) = \sqrt{|x|}$), $x(t) \equiv 0$ est solution, et il en est de même de $x(t) = \frac{1}{4}t^2$ et de $x(t) = \frac{1}{4}\max(t-t_0, 0)^2$ pour tout $t_0 \in \mathbb{R}_+$.

Afin de traiter le cas de dynamiques non-linéaires, on dispose de l'extension suivante du Théorème 8.3.2, où la propriété de Lipschitz globale est remplacée par une propriété locale (pour la preuve, voir par exemple l'annexe C de la référence [26]).

Théorème 8.3.6 (Cauchy–Lipschitz, cas mesurable et Lipschitz local) *On suppose que :*
(i) *L'application F est mesurable en t et continue en x ;*
(ii) *L'application F est intégrable en t, i.e.,*

$$\forall x \in \mathbb{R}^d, \quad \exists \beta \in L^1([0,T];\mathbb{R}_+), \quad \forall t \in [0,T], \quad |F(t,x)|_{\mathbb{R}^d} \leq \beta(t); \tag{8.19}$$

(iii) *L'application F est* ***localement lipschitzienne*** *en x, i.e.,*

$$\begin{aligned} &\forall x \in \mathbb{R}^d, \quad \exists r > 0, \quad \exists C_0 \in L^1([0,T];\mathbb{R}_+), \\ &\text{p.p. } t \in [0,T],\ \forall x_1, x_2 \in B(x,r), \quad |F(t,x_1) - F(t,x_2)|_{\mathbb{R}^d} \leq C_0(t)|x_1 - x_2|_{\mathbb{R}^d}, \end{aligned} \tag{8.20}$$

où $B(x,r)$ désigne la boule ouverte de centre x et de rayon r.
Alors, il existe une ***unique solution maximale*** *au problème de Cauchy* (8.10). *Cette solution est définie sur l'intervalle $J \subseteq [0,T]$ et on a soit $J = [0,T]$ soit $J = [0,T_*[$ avec $T_* < T$ et $\lim_{t\uparrow T_*} |x(t)|_{\mathbb{R}^d} = +\infty$. La solution maximale x est dans $AC(J;\mathbb{R}^d)$, elle satisfait le système différentiel* (8.10) *pour presque tout $t \in J$ et elle vérifie* (8.18) *pour tout $t \in J$.*

Exemple 8.3.7 [Explosion pour un système de contrôle] On se place dans $\mathbb{R}$ ($d = 1$) et on considère le système de contrôle (8.13) avec un contrôle à valeurs scalaires ($k = 1$) et l'application f telle que $f(t,x,u) = x^2 + u$ (qui ne dépend pas de t explicitement). On obtient alors le problème de Cauchy $\dot{x}(t) = x(t)^2 + u(t)$. On considère la donnée initiale $x_0 = 0$ et on suppose que le contrôle est constant en temps égal à $u_0 \in \mathbb{R}_+$. On vérifie sans peine que la trajectoire est donnée par $x(t) = \sqrt{u_0}\tan(\sqrt{u_0}t)$. On a donc explosion au temps fini $T_* = \frac{\pi}{2\sqrt{u_0}}$ qui dépend de la valeur (constante) prise par le contrôle.

Bibliographie

[1] ALLAIRE G., *Analyse numérique et optimisation,* Éditions de l'École Polytechnique, Palaiseau (2005).

[2] AUBIN J.-P., *Mathematical methods of game and economic theory*, volume 7 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam-New York, 1979.

[3] BARDI M., CAPUZZO-DOLCETTA I., *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations*. Systems & Control : Foundations & Applications. Birkhäuser Boston, Inc., Boston, MA, 1997. With appendices by Maurizio Falcone and Pierpaolo Soravia.

[4] BONNANS J., *Optimisation continue,* Mathématiques appliquées pour le Master / SMAI, Dunod, Paris (2006).

[5] BONNANS J., GAUBERT S., *Recherche opérationnelle. Aspects mathématiques et applications,* Éditions de l'École Polytechnique, Palaiseau (2015).

[6] BONNANS J., GILBERT J.-C., LEMARECHAL C., SAGASTIZABAL C., *Optimisation numérique,* Mathématiques et Applications 27, Springer, Paris (1997).

[7] BONNANS J., ROUCHON P., *Commande et optimisation de systèmes dynamiques,* Éditions de l'École Polytechnique, Palaiseau (2005).

[8] BOSCAIN U., MIRRAHIMI M., *Automatic Control with Applications in Robotics and in Quantum Engineering,* cours de 3ème année à l'Ecole Polytechnique (2019).

[9] BREZIS H., *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

[10] CHAMBOLLE A., POCK T., *An introduction to continuous optimization for imaging,* Acta Numerica, pp. 161-319 (2016).

[11] CHVÁTAL V., *Linear programming*, Freeman and Co., New York (1983).

[12] CULIOLI J.-C., *Introduction à l'optimisation,* Éditions Ellipses, Paris (1994).

[13] ERN A., *Contrôle de modèles dynamiques,* cours de 2ème année à l'Ecole Polytechnique (2019).

[14] EKELAND I., TEMAM R., *Analyse convexe et problèmes variationnels,* Dunod, Paris (1974).

[15] FLETCHER R., *Practical methods of optimization. Vol. 1.* John Wiley & Sons, Ltd., Chichester, 1980. Unconstrained optimization, A Wiley-Interscience Publication.

[16] GOLSE F., LASZLO Y., PACARD F., VITERBO C., *Analyse réelle,* cours de 1ère année à l'Ecole Polytechnique (2019).

[17] HERMES H., LASALLE J. P., *Functional analysis and time optimal control.* Academic Press, New York-London, 1969. Mathematics in Science and Engineering, Vol. 56.

[18] ISIDORI A., *Nonlinear control systems.* Communications and Control Engineering Series. Springer-Verlag, Berlin, third edition, 1995.

[19] LEE E. B., MARKUS L., *Foundations of optimal control theory.* Robert E. Krieger Publishing Co., Inc., Melbourne, FL, second edition, 1986.

[20] LIONS P.-L., *Contrôle de modèles dynamiques.* Cours polycopié. École Polytechnique, 2016.

[21] NESTEROV Y., *Lectures on convex optimization,* Springer Optimization and Its Applications, 137, Springer, Cham, 2018.

[22] NOCEDAL J., WRIGHT S., *Numerical optimization,* Springer Series in Operations Research and Financial Engineering, New York (2006).

[23] PADBERG M., *Linear optimization and extensions*, Springer, Berlin (1999).

[24] ROCKAFELLAR R. T., WETS R. J.-B., *Variational analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998.

[25] SCHRIJVER A., *Theory of linear and integer programming*, Wiley, New York (1986).

[26] SONTAG E. D., *Mathematical control theory*, volume 6 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 1998. Deterministic finite-dimensional systems.

[27] TRELAT E., *Contrôle optimal.* Mathématiques Concrètes. Vuibert, Paris, 2005. Théorie & applications.

[28] VINTER R., *Optimal control.* Systems & Control : Foundations & Applications. Birkhäuser Boston, Inc., Boston, MA, 2000.