Optimisation et contrôle

Grégoire ALLAIRE, Alexandre ERN
Ecole Polytechnique

7 janvier 2026

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 discute des aspects théoriques de l'optimisation. Il présente quelques résultats d'existence de solutions de problèmes d'optimisation, notamment à l'aide de notions d'analyse convexe, mais il se concentre surtout sur la dérivation des 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 dans le Chapitre 3. Elles reposent sur des notions fondamentales comme celle des **multiplicateurs de Lagrange**, du **point-selle pour un Lagrangien** ou encore de la **dualité**. Le Chapitre 3 est donc consacré aux algorithmes numériques d'optimisation avec ou sans contraintes. Les algorithmes classiques de type gradient, du premier ordre, ou de type Newton, du second ordre sont étudiés en détails. De nombreux autres algorithmes, plus spécifiques ou spécialisés, sont aussi brièvement discutés afin de montrer au lecteur la richesse des méthodes numériques disponibles en optimisation. 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 [23].

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 février 2025

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 [6] 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 [7], [13], [15], [27].

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 **Hamiltonien** et le **feedback** grâce à l'équation de Riccati. Enfin le Chapitre 7 considère le cas général d'une dynamique 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 [33], ou des ouvrages plus spécialisés (en anglais) comme [3], [4], [16], [20], [22], [30], [32] ou [34].

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 (prix Nobel d'économie) et récemment encore par de très nombreux mathématiciens. Cette théorie du transport, dit optimal, a de très nombreuses applications en gestion, finance, traitement des images, problèmes inverses, intelligence artificielle... 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 (qui assurent que chaque femme trouvera un mari et chaque homme une épouse)

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

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 nous renvoyons au cours de troisième année [6] pour plus de détails. •

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" ou "parcimonieuse". 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) : nous verrons comment résoudre (1.3) dans la Sous-section 3.2.7. 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 l'Exercice 2.5.16 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, dite "logistique", 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 $y = \operatorname{sgn}\Big(h_{w^*,\tau^*}(x)\Big)$.

Le problème (1.5) peut aussi s'interpréter comme une séparation des données en deux classes les plus éloignées possibles au sens de la distance Euclidienne dans $\mathbb{R}^d$: le rapport $2/\|w\|_2$ est la distance entre les deux hyperplans affines $w \cdot x - \tau = +1$ et $w \cdot x - \tau = -1$ qui séparent les données si celles-ci vérifient $y_i(w \cdot x_i - \tau) \geq 1$. Dans ce cas, cette technique relève de ce que l'on appelle les machines à vecteurs de support ou séparateurs à vaste marge (en anglais, support vector machine, ou SVM). •

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], [15]). •

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 le temps final, appelé horizon temporel. Celui-ci 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 [14].

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)

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

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 (robot à vitesse constante ou véhicule de Dubins) Passons à un premier exemple de système de contrôle non-linéaire : le véhicule, dit de Dubins, qui se déplace à vitesse constante mais orientation variable. 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 du véhicule 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'un contrôle $u : [0,T] \to [-1, 1]$ ($k = 1$) qui prescrit la vitesse angulaire de l'axe des roues. La dynamique du système est régie par le système différentiel suivant :

$$\begin{cases} \dot{X}(t) = v\cos(\theta(t)), \\ \dot{Y}(t) = v\sin(\theta(t)), \\ \dot{\theta}(t) = u(t), \end{cases}$$

où v est la vitesse du robot, supposée constante. La donnée initiale pour ce système prescrit, non seulement la position initiale $(X(0), Y(0))$, mais aussi la vitesse initiale qui dépend de $\theta(0)$. On peut envisager plusieurs problèmes de contrôle pour ce véhicule. Comme dans le cas du tram, on peut vouloir atteindre une cible prescrite, et le faire en un minimum de temps. Comme dans le cas du système linéaire-quadratique, on peut minimiser une fonctionnelle de coût résultant d'une pondération entre l'adéquation à une trajectoire cible et des valeurs modérées pour le contrôle. Nous renvoyons au Chapitre 7 pour des éléments de résolution. •

Exemple 1.3.4 (ruche d'abeilles) La dynamique des populations est une source de nombreux modèles de contrôle non-linéaire. Par souci de simplicité nous considérons un modèle de ruche d'abeilles qui ne comprend que deux espèces : les abeilles et les reines. On suppose que dans la ruche, la population d'abeilles $a(t)$ et celle des reines $r(t)$ évoluent selon la dynamique

$$\dot{x}(t) = \begin{pmatrix} \dot{a}(t) \\ \dot{r}(t) \end{pmatrix} = \begin{pmatrix} \varphi(u(t))a(t) \\ u(t)a(t) \end{pmatrix} \quad \text{pour } t \in [0,T],$$

avec la donnée initiale $a(0) > 0$, $r(0) \geq 0$ et un contrôle $u(t)$ à valeurs dans $U = [0,1]$, qui représente l'effort des abeilles pour fournir des reines (on vérifie bien que plus u est grand, plus la croissance de r est importante). La fonction φ est définie par

$$\varphi(u) = \alpha(1-u) - \beta,$$

avec des paramètres $\alpha > \beta > 0$ strictement positifs. L'interprétation de ce modèle de fonction φ est la suivante :

- si $u = 1$, alors $\varphi(u) = -\beta$ et la population d'abeilles décroît (exponentiellement) ;
- si $u = 0$, alors $\varphi(u) = \alpha - \beta > 0$ et la population d'abeilles croît (exponentiellement).

Dans l'objectif d'obtenir un maximum de reines r, il ne faut pas non plus que la population d'abeilles diminue trop car la production de reines est aussi proportionnelle à a. Il y a donc un équilibre à trouver, un contrôle optimal u qui ne doit pas être tout le temps constant égal à 0 ou à 1, pour favoriser la croissance des reines. L'objectif est de trouver un contrôle optimal qui maximise la population de reines au temps final T. Autrement dit, on veut résoudre le problème de contrôle optimal :

$$\inf_{u(t)\in[0,1]} J(u) = -r(T).$$

•

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.

Contre-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.

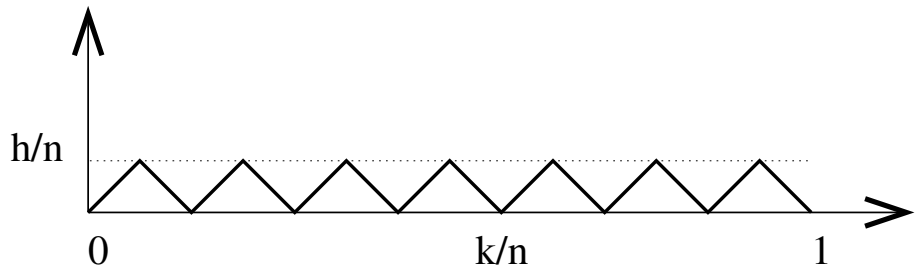

FIGURE 2.1 – Suite minimisante u^n pour l'Exemple 2.3.2.

Contre-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, à l'intérieur de la boule unité, s'obtient par un argument similaire sur une plus petite boule qui est un voisinage de ce point. □

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 (ou α-convexe) s'il existe $\alpha > 0$ tel que, pour tout $u, v \in K$ et tout $\theta \in [0,1]$,*

$$J(\theta u + (1-\theta)v) \leq \theta J(u) + (1-\theta)J(v) - \frac{\alpha\theta(1-\theta)}{2}\|u-v\|^2 . \tag{2.9}$$

Evidemment, les fonctions fortement convexes sont également strictement convexes. Dans la Définition 2.3.6, il est possible de ne tester la forte convexité de J que pour la valeur $\theta = 1/2$, comme le montre l'exercice suivant.

Exercice 2.3.3 Soit J une fonction continue sur un ensemble convexe K et $\alpha > 0$ tels que, pour tout $u, v \in K$,

$$J\left(\frac{u+v}{2}\right) \leq \frac{J(u)+J(v)}{2} - \frac{\alpha}{8}\|u-v\|^2 . \tag{2.10}$$

Montrer que J est fortement ou α-convexe.

Exercice 2.3.4 Montrer qu'une fonction J, définie sur un ensemble convexe K, est fortement ou α-convexe si et seulement si la fonction $J(v) - \frac{\alpha}{2}\|v\|^2$ est convexe sur K.

Exercice 2.3.5 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 2.3.6 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) pour $\theta = 1/2$ 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. □

Remarque 2.3.8 La démonstration de la Proposition 2.3.7 peut paraitre un peu compliquée puisqu'elle fait appel à un résultat abstrait, le Théorème 8.1.12 de séparation d'un point et d'un convexe. Nous verrons une démonstration beaucoup plus simple pour les fonctions J différentiables dans les Propositions 2.4.4 et 2.4.5. •

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 2.3.9 (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*

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

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

Démonstration. Soit $(u^n)_{n\in\mathbb{N}}$ 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 (2.10) 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 (2.10),

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

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

Il est possible de généraliser en grande partie le Théorème 2.3.9 au cas de fonctions J qui sont seulement convexes (et non pas fortement convexes). Cependant, autant la démonstration du Théorème 2.3.9 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 2.3.10 (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 2.3.11 Le Théorème 2.3.10 donne l'existence d'un minimum comme le précédent Théorème 2.3.9, 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 2.3.12 Le Théorème 2.3.10 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 2.3.10 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 [9]).

Définition 2.3.13 *On dit qu'une suite $(u^n)_{n\in\mathbb{N}}$ 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 2.3.13 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 2.3.14 *De toute suite $(u^n)_{n\in\mathbb{N}}$ 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 2.3.15 *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 2.3.16 *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 2.3.15 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 2.3.10. D'après (2.2), toute suite minimisante (u^n) est bornée. On déduit alors du Lemme 2.3.14 qu'il existe une sous-suite $(u^{n'})$ convergeant faiblement vers une limite $u \in V$. Mais, d'après le Lemme 2.3.16, $u \in K$ et

$$J(u) \leq \liminf_{n'\to+\infty} J(u^{n'}) = \inf_{v\in K} J(v) \ .$$

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*

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

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

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

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 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*

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

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

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

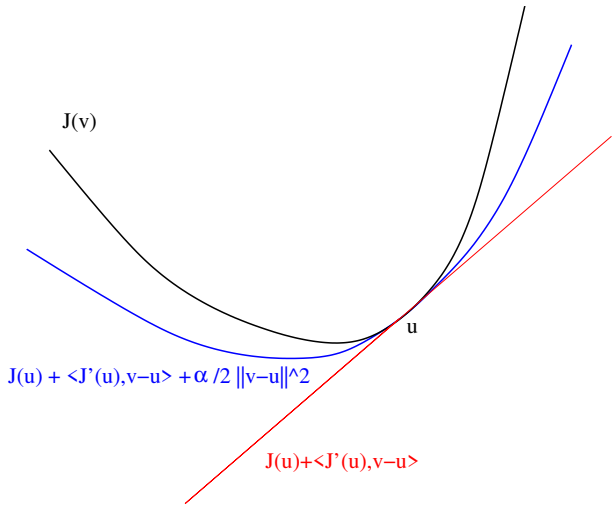

FIGURE 2.2 – Fonction fortement convexe : fonction quadratique minorante (en bleu) et plan tangent (en rouge).

Remarque 2.4.6 Les conditions (2.20) et (2.23) ont une interprétation géométrique simple : la première signifie que la fonction convexe $J(v)$ est toujours au dessus de son plan tangent en u (considéré comme une fonction affine de v), tandis que la deuxième affirme que $J(v)$ est au dessus d'une fonction quadratique en v qui lui est tangente en u (voir la Figure 2.2). Les conditions (2.21) et (2.24) sont des propriétés de monotonie ou de croissance de J'. •

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, elle vérifie

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

dans laquelle on fait tendre θ vers 0 pour obtenir (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

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

Définissons maintenant 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

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

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*

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

Telle qu'elle est définie la dérivée seconde est difficile à évaluer en pratique car $w \to J''(u)w$ est un opérateur linéaire continu de V dans V'. Heureusement, en faisant agir $J''(u)w$ 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 $(J''(u)w)\, v$. Au vu de la formule de Taylor d'ordre deux ci-dessous, en pratique on calcule plutôt $J''(u)(w, w)$.

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

$$J(u+w) = J(u) + \langle J'(u), w\rangle + \frac{1}{2}J''(u)(w,w) + o(\|w\|^2), \tag{2.27}$$

avec $\lim_{w\to 0} \frac{o(\|w\|^2)}{\|w\|^2} = 0$ où $J''(u)$ est identifiée à une forme bilinéaire continue sur $V \times V$.

Démonstration. On écrit un développement de Taylor avec reste exact

$$J(u+w) = J(u) + \int_0^1 \langle J'(u+sw), w\rangle ds$$

et on y insère la définition de la dérivée seconde

$$J'(u+sw) = J'(u) + sJ''(u)w + o(sw).$$

Nous laissons au lecteur le soin de finir les détails pour arriver à (2.27). □

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

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

Terminons cette section en donnant la définition d'une fonction Lipschitzienne.

Définition 2.4.9 *Soit F une fonction de V dans W où V et W sont deux espaces de Hilbert. On dit que F est Lipschitzienne de constante $L > 0$ si*

$$\|F(v) - F(u)\| \leq L\|v - u\| \qquad \textit{pour tout } v, u \in V.$$

Lemme 2.4.10 *Soit J une fonction différentiable de V dans $\mathbb{R}$ telle que sa dérivée J' est Lipschitzienne de constante $L > 0$. Alors*

$$J(v) \leq J(u) + \langle J'(u), v - u\rangle + \frac{L}{2}\|v - u\|^2 \qquad \textit{pour tout } v, u \in V.$$

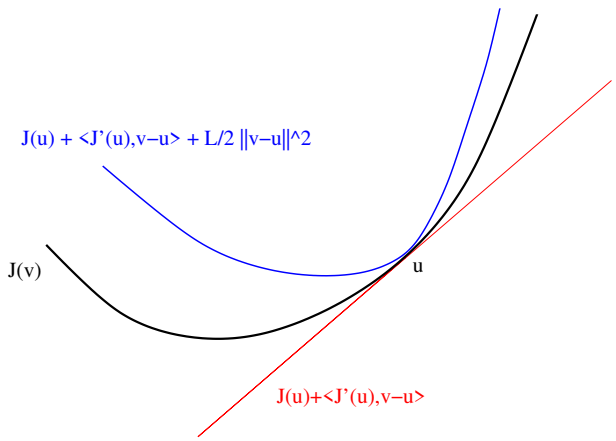

FIGURE 2.3 – Fonction L-Lipschitzienne : fonction quadratique majorante (en bleu) et hyperplan tangent (en rouge).

Remarque 2.4.11 Une interprétation du Lemme 2.4.10 est qu'une fonction différentiable à dérivée L-Lipschitzienne peut être majorée par une fonction quadratique qui lui est tangente au point u (voir la Figure 2.3). C'est en quelque sorte une version opposée (majoration au lieu de minoration) de la Remarque 2.4.6 pour une fonction α-convexe différentiable (notons néanmoins qu'il n'y a pas d'hypothèse de convexité dans le Lemme 2.4.10). Par conséquent, les fonctions fortement convexes à dérivée Lipschitzienne sont remarquables car elles sont encadrées par deux fonctions quadratiques qui sont tangentes au point u. Cette propriété très utile sera exploitée dans les preuves de convergence des algorithmes d'optimisation. •

Démonstration. Ecrivons une formule de Taylor avec reste exact

$$J(v) = J(u) + \int_0^1 \langle J'(u + t(v-u)), v-u\rangle dt.$$

Par conséquent,

$$\Big|J(v) - J(u) - \langle J'(u), v-u\rangle\Big| \leq \int_0^1 \|J'(u+t(v-u)) - J'(u)\|\|v-u\|dt,$$

d'où la conclusion en utilisant l'hypothèse de Lipschitziannité de J'. □

Exercice 2.4.8 Soit J une fonction de classe C^2 de V dans $\mathbb{R}$. On suppose qu'il existe une constante $L > 0$ telle que $J''(u)(w,w) \leq L\|w\|^2$ pour tout $u, w \in V$. Montrer que J' est Lipschitzienne de constante $L > 0$.

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. •

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

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

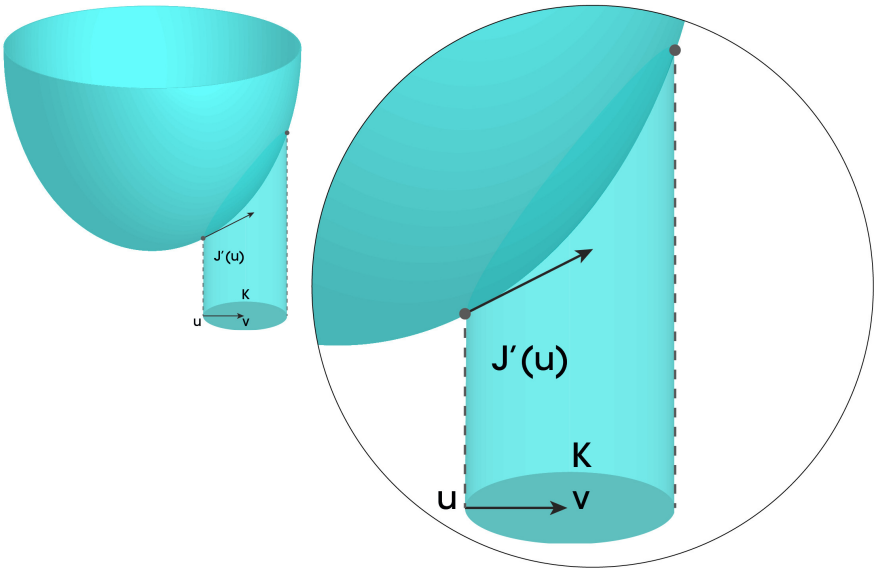

FIGURE 2.4 – Inéquation d'Euler : l'angle entre la dérivée $J'(u)$ et la direction rentrante $(v-u)$ est aigu.

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. Montrer aussi que la solution est unique si et seulement si $\mathrm{Ker}A = \{0\}$.

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

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

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 2.5.18.

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.

Autrement dit, $K(v)$ est l'ensemble de toutes les directions possibles de variations à partir de v en restant dans K. Ce cône $K(v)$ est aussi appelé **cône tangent** à K. En effet, $K(v)$ est l'ensemble de tous les vecteurs qui sont tangents en v à une courbe contenue dans K et passant par v. Par exemple, si K est une variété régulière, $K(v)$ est simplement l'espace tangent à K en v.

Il est facile de vérifier que $0 \in K(v)$ (prendre la suite constante $v^n = v$), 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 $\lambda \geq 0$ (changer la suite ε^n en ε^n/λ pour $\lambda > 0$) et que $K(v)$ est fermé (par un argument de suite diagonale).

Exercice 2.5.9 Montrer que $K(v) = V$ si v est intérieur à K. Montrer que $K(v_0) = \{0\}$ si $K = K_0 \cup \{v_0\}$ avec la distance $d(v_0, K_0) > 0$.

L'intérêt du cône des directions admissibles réside dans le résultat suivant, qui donne une condition **nécessaire** d'optimalité.

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*

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

Démonstration. Soit une direction admissible $w \in K(u)$. Il existe donc une suite

$$\lim_{n\to+\infty} v^n = u\ ,\ \lim_{n\to+\infty} \varepsilon^n = 0\ ,\ \lim_{n\to+\infty} \frac{v^n - u}{\varepsilon^n} = w$$

Pour n suffisamment grand,

$$J(u) \leq J(v^n) = J\big(u + \varepsilon^n w + o(\varepsilon^n)\big) = J(u) + \varepsilon^n \langle J'(u), w\rangle + o(\varepsilon^n).$$

En soustrayant $J(u)$, en divisant par ε^n, puis en passant à la limite $n \to +\infty$, on obtient le résultat. □

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 Exercices 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 $\lambda_1, \dots, \lambda_M \in \mathbb{R}$, appelés **multiplicateurs de Lagrange**, *tels que*

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

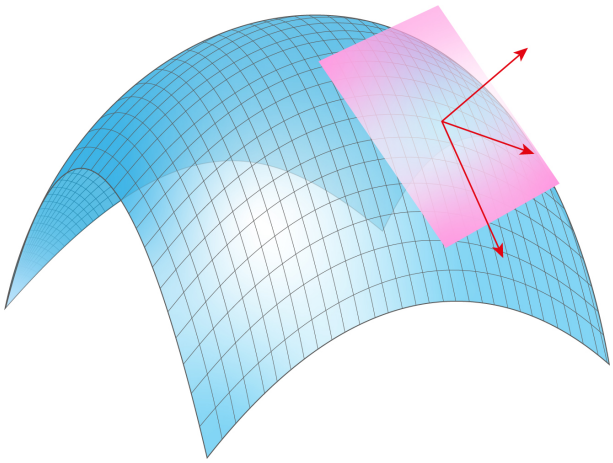

FIGURE 2.5 – Variété K (en bleu) et son hyperplan tangent $K(u)$ en un point u (en rose). Ce dessin correspond à $V = \mathbb{R}^3$ et $M = 1$ contrainte.

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 2.5), c'est-à-dire

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

avec, par définition

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

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

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

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 $v=u$ qui vérifie $\mathcal{F}(\varepsilon,u)=F(u+\varepsilon w)=o(\varepsilon)$ puisque $\langle F'(u),w\rangle=0$. Le Lemme 2.5.7 fournit alors une suite v_ε telle que $\mathcal{F}(\varepsilon,v_\varepsilon)=F(v_\varepsilon+\varepsilon w)=0$ et $\|u-v_\varepsilon\|_V=o(\varepsilon)$. On en déduit que la suite $v_\varepsilon+\varepsilon w$ 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 2.5.14. □

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*

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

Une démonstration du Lemme 2.5.7 peut être trouvée, dans un cadre plus général (théorème de l'application surjective), dans [5].

Remarque 2.5.8 Lorsque les vecteurs $\big(F_i'(u)\big)_{1\le i\le M}$ 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 $m\le n$. On note $(b_i)_{1\le i\le m}$ 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

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

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 $F(v) = 0$.*

Une propriété importante du Lagrangien, qui justifie en quelque sorte sa définition, est qu'il 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 $J(v) = \mathcal{L}(v,\mu)$ 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 3.4.2 lorsque sera présenté l'algorithme d'Uzawa. •

La notion de Lagrangien permet d'écrire le Théorème 2.5.6 sous la forme de la stationnarité du Lagrangien (c'est-à-dire que son gradient par rapport aux deux variables v et μ s'annule).

Corollaire 2.5.12 *On se place sous les hypothèses du Théorème 2.5.6. Alors, si $u \in K$ est un minimum local de J sur K, il existe $\lambda \in \mathbb{R}^M$ tel que*

$$\frac{\partial \mathcal{L}}{\partial v}(u,\lambda) = 0 \quad , \quad \frac{\partial \mathcal{L}}{\partial \mu}(u,\lambda) = 0 \, , \tag{2.41}$$

où $\mathcal{L}(v,\mu)$ est le Lagrangien introduit dans la Définition 2.5.9.

Démonstration. Tout d'abord, pour tout $\mu \in \mathbb{R}^M$, $\dfrac{\partial \mathcal{L}}{\partial \mu}(u,\mu) = F(u)$ qui s'annule si $u \in K$, c'est-à-dire si $F(u) = 0$. D'autre part, on a

$$\frac{\partial \mathcal{L}}{\partial v}(v,\mu) = J'(v) + \mu F'(v),$$

qui s'annule pour $(v,\mu) = (u,\lambda)$ car il est égal à la condition d'optimalité (2.37). Autrement dit, le Corollaire 2.5.12 est équivalent au Théorème 2.5.6, au sens où on a montré que la stationnarité du Lagrangien (2.41) est équivalente à la condition d'optimalité et à la satisfaction de la contrainte. □

Le résultat suivant permet de donner une interprétation concrète à la notion de multiplicateur de Lagrange. Plus précisément, les multiplicateurs de Lagrange λ_i dans le Théorème 2.5.6 donnent la sensibilité de la valeur minimale de J aux variations du niveau des contraintes F_i (au signe près). Par exemple, en économie, les multiplicateurs de Lagrange mesurent des prix ou des coûts marginaux, alors qu'en mécanique ils fournissent des forces de réaction correspondant à des déplacements imposés.

Pour $\epsilon \in \mathbb{R}^M$, on considère le problème d'optimisation

$$\inf_{v \in V,\ F(v)=\epsilon} J(v), \tag{2.42}$$

où le niveau des contraintes est paramétré par ϵ. On note $\mathcal{M}(\epsilon)$ la valeur minimum dans (2.42) ($\mathcal{M}$ est une fonction de $\mathbb{R}^M$ dans $\mathbb{R}$).

Lemme 2.5.13 *Soit $J : V \mapsto \mathbb{R}$ et $F : V \mapsto \mathbb{R}^M$ des fonctions de classe C^1. On suppose que, pour ϵ petit, le problème (2.42) admet une solution unique, notée u_ϵ, qui est supposée différentiable par rapport à ϵ en 0. Pour $\epsilon = 0$ on note u cette solution et on suppose que les vecteurs $\big(F_i'(u)\big)_{1 \le i \le M}$ sont linéairement indépendants. Soit $\lambda \in \mathbb{R}^M$ le multiplicateur de Lagrange associé à u, donné par le Théorème 2.5.6. Alors, pour $1 \le i \le M$, on a*

$$\lambda_i = -\frac{\partial \mathcal{M}}{\partial \epsilon_i}(0).$$

Démonstration. On introduit le Lagrangien de (2.42), où l'on fait intervenir explicitement le niveau de contraintes donné par la variable ϵ. Autrement dit, pour $(v,\mu,\epsilon) \in V \times \mathbb{R}^M \times \mathbb{R}^M$, on définit

$$\mathcal{L}(v,\mu,\epsilon) = J(v) + \mu \cdot (F(v) - \epsilon).$$

Comme la solution u_ϵ vérifie les contraintes, on a $\mathcal{L}(u_\epsilon, \mu, \epsilon) = J(u_\epsilon) = \mathcal{M}(\epsilon)$. En dérivant cette relation en $\epsilon = 0$, on obtient

$$\frac{\partial \mathcal{M}}{\partial \epsilon_i}(0) = \langle \frac{\partial \mathcal{L}}{\partial v}(u, \mu, 0), \frac{\partial u_\epsilon}{\partial \epsilon_i}(0) \rangle + \frac{\partial \mathcal{L}}{\partial \epsilon_i}(0) = \langle J'(u) + \mu \cdot F'(u), \frac{\partial u_\epsilon}{\partial \epsilon_i}(0) \rangle - \mu_i.$$

On remplace alors μ par λ et on utilise la condition d'optimalité du Théorème 2.5.6 pour conclure. □

Exercice 2.5.16 On reprend l'Exemple 1.2.5 sur la consommation des ménages en économie. Si on note $x \in \mathbb{R}^n_+$ le vecteur des biens consommés, un ménage cherche à maximiser

$$\max_{x \in \mathbb{R}^n_+,\, x \cdot p \leq b} u(x), \tag{2.43}$$

avec le vecteur des prix $p \in \mathbb{R}^n_+$ dont toutes les composantes sont strictement positives, $b > 0$ le budget du ménage et $u(x)$ la fonction d'utilité supposée C^1 et croissante par rapport à chaque composante de x. Montrer que (2.43) admet une solution x^* et que la contrainte d'inégalité peut être remplacée par une contrainte d'égalité. Déduire de la condition d'optimalité que le rapport entre chaque prix p_i et l'utilité marginale $\frac{\partial u}{\partial x_i}(x^*)$ est constant pour $1 \leq i \leq n$.

On suppose désormais que la fonction d'utilité $u(x)$ est strictement concave et de classe C^2. Montrer que la solution x^* et le multiplicateur de Lagrange λ sont uniques et que leur dépendance par rapport au budget b est de classe C^1 (on pourra utiliser un théorème de fonction implicite). Interpréter λ en terme de variation marginale, par rapport à b, du maximum d'utilité.

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

Proposition 2.5.14 *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*

$$\left(J''(u) + \sum_{i=1}^{M} \lambda_i F_i''(u) \right)(w, w) \geq 0 \quad \forall\, w \in K(u) = \bigcap_{i=1}^{M} [F_i'(u)]^\perp . \tag{2.44}$$

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

$$J'(u) + \sum_{i=1}^{M} \lambda_i F_i'(u) = 0. \tag{2.45}$$

En sommant les conditions $f_i''(0) = 0$, multipliées par λ_i, on obtient

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

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 (2.45) 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 (2.44) lorsque $u'(0)$ parcourt $K(u)$.

L'existence de tels chemins admissibles $u(t)$ pour lesquels $u'(0)$ 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 [5]). □

Exercice 2.5.17 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

$$K = \{v \in V \ , \quad F_i(v) \le 0 \quad \text{pour} \quad 1 \le i \le M\} \ , \tag{2.46}$$

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 (2.46) 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 (2.46) 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 2.5.15 *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 2.5.16 *On dit que les contraintes (2.46) 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 $i \in I(u)$*

$$\begin{array}{ll} \textit{ou bien} & \langle F_i'(u), \overline{w} \rangle < 0\ , \\ \textit{ou bien} & \langle F_i'(u), \overline{w} \rangle = 0 \quad \textit{et} \quad F_i \quad \textit{est affine}\ . \end{array} \tag{2.47}$$

Remarque 2.5.17 La direction $\overline{w}$ est en quelque sorte une "direction rentrante" puisque on déduit de (2.47) 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 2.5.16 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 (2.46).

Théorème 2.5.18 *On suppose que K est donné par (2.46), 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*

$$J'(u)+\sum_{i=1}^{M} \lambda_i F_i'(u) = 0, \quad \lambda_i \geq 0, \quad \lambda_i = 0 \ \textit{si}\ F_i(u) < 0, \quad \forall\, i \in \{1, \ldots, M\}. \tag{2.48}$$

Remarque 2.5.19 On peut réécrire la condition (2.48) sous la forme suivante

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

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

$$\tilde{K}(u) = \{w \in V \text{ tel que } \langle F_i'(u), w\rangle \leq 0, \quad \forall\, i \in I(u)\} \ . \tag{2.49}$$

(On peut montrer que $\tilde{K}(u)$ n'est autre que le cône $K(u)$ des directions admissibles, voir [5]). Soit $\overline{w}$ une direction admissible satisfaisant (2.47), $w \in \tilde{K}(u)$, 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

$$\begin{aligned} F_i(u + \varepsilon(w + \delta\overline{w})) &= F_i(u) + \varepsilon\langle F_i'(u), w + \delta\overline{w}\rangle + o(\varepsilon) \\ &\leq \varepsilon\delta\langle F_i'(u), \overline{w}\rangle + o(\varepsilon) \ < \ 0 \ , \end{aligned} \tag{2.50}$$

   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

$$F_i(u + \varepsilon(w + \delta\overline{w})) = F_i(u) + \varepsilon\langle F_i'(u), w + \delta\overline{w}\rangle = \varepsilon\langle F_i'(u), w\rangle \leq 0 \ . \tag{2.51}$$

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 2.5.20 ci-dessous. □

Lemme 2.5.20 (de Farkas) *Soient $a_1, \ldots, a_M$ des éléments fixés de V. On considère les ensembles*

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

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 $(a_i)_{1\le i\le M}$ 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 $\lambda_i^n \ge 0\ \forall\, i\, \forall\, n$), convergeant vers une limite $q \in V$. Alors il est clair que chaque suite (λ_i^n) converge dans $\mathbb{R}_+$ vers une limite $\lambda_i \ge 0$ (pour $1 \le i \le M$) puisque les vecteurs $(a_i)_{1\le i\le M}$ 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 $(a_i)_{1\le i\le M}$ 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 $(a_i)_{1\le i\le M}$ 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 $t \le 0$, on peut aussi écrire $q = -\displaystyle\sum_{i=1}^M (\lambda_i + t\mu_i)a_i$, et on peut choisir $t \le 0$ pour que

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

Ce raisonnement montre que

$$\hat{\mathcal{K}} = \bigcup_{i_0=1}^M \left\{ q \in V\ ,\ \exists\, \lambda_1,\dots,\lambda_M \ge 0\ ,\ q = -\sum_{i\ne i_0} \lambda_i a_i \right\}. \tag{2.52}$$

Par notre hypothèse de récurrence, chacun des ensembles apparaissant dans le membre de droite de (2.52) est fermé, et il en est donc de même de $\hat{\mathcal{K}}$.

Raisonnons maintenant par l'absurde : supposons que $\langle p, w\rangle \ge 0\ \forall\, w \in \mathcal{K}$ 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 $w \ne 0$ dans V et $\alpha \in \mathbb{R}$ tels que

$$\langle p, w\rangle < \alpha < \langle q, w\rangle\ \forall\, q \in \hat{\mathcal{K}}\ . \tag{2.53}$$

Mais alors, on doit avoir $\alpha < 0$ puisque $0 \in \hat{\mathcal{K}}$; d'autre part, pour tout $i \in \{1,\dots,M\}$ nous pouvons choisir dans (2.53) $q = -\lambda a_i$ avec λ arbitrairement grand, ce qui montre que $\langle a_i, w\rangle \le 0$. 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 2.5.18 grâce à la notion de Lagrangien.

Définition 2.5.21 *On appelle* **Lagrangien** *du problème de minimisation de $J(v)$, sous les contraintes d'inégalité $F(v) \le 0$, 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 $F(v) \le 0$.*

On remarquera que, par rapport à la Définition 2.5.9, la variable μ est ici contrainte à être positive ou nulle. C'est ce qui distingue les contraintes d'égalité des contraintes d'inégalité. Comme dans le Lemme 2.5.10 la maximisation du Lagrangien permet de faire "disparaître" la contrainte.

Lemme 2.5.22 *Le problème de minimisation sous contrainte d'inégalité est équivalent à un problème de min-max*

$$\inf_{v\in V,\ F(v)\leq 0} J(v) = \inf_{v\in V} \sup_{\mu\in(\mathbb{R}^+)^M} \mathcal{L}(v,\mu). \tag{2.54}$$

La démonstration du Lemme 2.5.22 est très similaire à celle du Lemme 2.5.10 et laissée au lecteur en guise d'exercice (noter qu'il est essentiel de tenir compte de la limitation $\mu \geq 0$).

Remarque 2.5.23 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 3.4.2 lorsque sera présenté l'algorithme d'Uzawa. •

La notion de Lagrangien permet d'écrire le Théorème 2.5.18 sous la forme de la stationnarité du Lagrangien.

Corollaire 2.5.24 *On se place sous les hypothèses du Théorème 2.5.18. Alors, si $u \in K$ est un minimum local de J sur K, il existe $\lambda \in (\mathbb{R}^+)^M$ tel que*

$$\frac{\partial \mathcal{L}}{\partial v}(u,\lambda) = 0\,, \quad \frac{\partial \mathcal{L}}{\partial \mu}(u,\lambda)\cdot(\mu-\lambda) \leq 0 \quad \forall \mu \in (\mathbb{R}^+)^M\,, \tag{2.55}$$

où $\mathcal{L}(v,\mu)$ est le Lagrangien introduit dans la Définition 2.5.21.

Démonstration. Les conditions (2.55) sont appelées stationnarité du Lagrangien. La deuxième partie de (2.55) est l'inéquation d'Euler (2.29) pour la maximisation par rapport à μ dans le convexe fermé $(\mathbb{R}^+)^M$ de $\mathcal{L}(u,\mu)$. On vérifie que la condition nécessaire d'optimalité (2.48) du Théorème 2.5.18 est équivalente à (2.55) puisque, d'une part,

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

et, d'autre part, la condition $\lambda \geq 0$, $F(u) \leq 0$, $\lambda \cdot F(u) = 0$ est équivalente à

$$F(u)\cdot(\mu-\lambda) \leq 0 \quad \forall \mu \in (\mathbb{R}^+)^M$$

et on rappelle que

$$\frac{\partial \mathcal{L}}{\partial \mu}(u,\lambda) = F(u).$$

Autrement dit, (2.55) est la condition nécessaire d'optimalité pour que (u, λ) soit un point selle du Lagrangien $\mathcal{L}(v, \mu)$ dans $V \times (\mathbb{R}^+)^M$. □

Exercice 2.5.18 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 2.5.18 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 2.5.19 Soit f une fonction définie sur un ouvert borné Ω. Pour $\epsilon > 0$ on considère le problème de régularisation suivant

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

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

$$K = \{v \in V\,, \quad G(v) = 0\,, \quad F(v) \leq 0\}\,, \tag{2.56}$$

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 $I(u) = \{i \in \{1, \ldots, M\}\ ,\ F_i(u) = 0\}$ l'ensemble des contraintes d'inégalité actives en $u \in K$.

Définition 2.5.25 *On dit que les contraintes (2.56) 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 $i \in I(u)$*

$$\langle F_i'(u), \overline{w} \rangle < 0\,. \tag{2.57}$$

Nous pouvons alors énoncer les conditions **nécessaires** d'optimalité sur l'ensemble (2.56).

Théorème 2.5.26 *Soit $u \in K$ où K est donné par (2.56). 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 2.5.25). Alors, si u est un minimum local de J sur K, il existe des multiplicateurs de Lagrange $\mu_1, \ldots, \mu_N$, et $\lambda_1, \ldots, \lambda_M \geq 0$, tels que*

$$J'(u) + \sum_{i=1}^{N} \mu_i G_i'(u) + \sum_{i=1}^{M} \lambda_i F_i'(u) = 0 \ , \quad \lambda \geq 0 \ , \ F(u) \leq 0 \ , \ \lambda \cdot F(u) = 0 \ . \ (2.58)$$

La démonstration du Théorème 2.5.26 est une simple adaptation de celles des Théorèmes 2.5.6 et 2.5.18, 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 2.5.16 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 2.5.27 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 2.5.16. 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 $i \in I(u)$)

$$(F_i'(u))_{i \in I(u)} \text{ est une famille libre} \qquad (2.59)$$

entraîne (2.47), 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 (2.47) n'implique pas (2.59). •

Remarque 2.5.28 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 2.5.25. Cette condition "forte" (c'est-à-dire moins souvent vérifiée) de qualification est

$$\big(G_i'(u)\big)_{1 \leq i \leq N} \bigcup (F_i'(u))_{i \in I(u)} \text{ est une famille libre.} \qquad (2.60)$$

On vérifie facilement que (2.60) entraîne (2.57), c'est-à-dire que les contraintes sont qualifiées. •

Remarque 2.5.29 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 $i \in \{1, \ldots, M\}$,

$$\begin{array}{l} \text{les fonctions } F_i \text{ sont convexes et,} \\ \quad \text{ou bien} \quad F_i(\overline{v}) < 0 \ , \\ \quad \text{ou bien} \quad F_i(\overline{v}) = 0 \quad \text{et} \quad F_i \quad \text{est affine.} \end{array} \tag{2.61}$$

L'hypothèse (2.61) entraîne que les contraintes sont qualifiées en $u \in K$ au sens de la Définition 2.5.16. En effet, si $i \in I(u)$ et si $F_i(\overline{v}) < 0$, alors, d'après la condition de convexité (2.20)

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

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 2.5.16 de qualification des contraintes est satisfaite avec $\overline{w} = \overline{v} - u$. L'avantage de l'hypothèse (2.61) est de ne pas nécessiter de connaître le point de minimum u ni de calculer les dérivées des fonctions $F_1, \ldots, F_M$. •

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 2.5.18 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 $P = (\mathbb{R}_+)^M$.

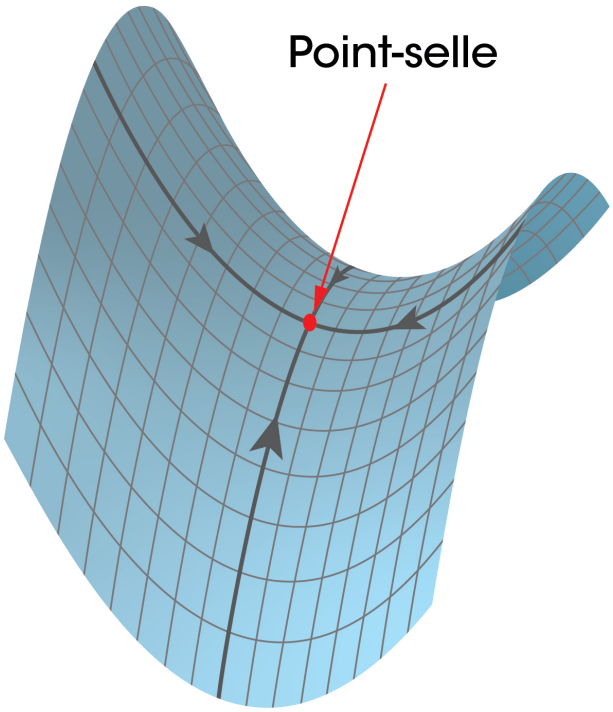

FIGURE 2.6 – Point selle ou col pour un Lagrangien.

Donnons maintenant la définition d'un point-selle, souvent appelé également min-max ou col (voir la Figure 2.6).

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*

$$\forall\, q \in P \qquad \mathcal{L}(u,q) \le \mathcal{L}(u,p) \le \mathcal{L}(v,p) \quad \forall\, v \in U\ . \tag{2.62}$$

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 (2.46) étudiés dans la section précédente. Pour simplifier, nous utiliserons de nouveau des inégalités entre vecteurs, notant parfois $q \ge 0$ au lieu de $q \in (\mathbb{R}_+)^M$.

Proposition 2.6.2 *On suppose que les fonctions $J, F_1, \dots, F_M$ sont continues sur V, et que l'ensemble K est défini par (2.36) ou (2.46). 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 (2.46). 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 $F_1, \dots, F_M$ sont dérivables en u, on a

$$J'(u) + \sum_{i=1}^{M} p_i F_i'(u) = 0\ . \tag{2.63}$$

Démonstration. Écrivons la condition de point-selle

$$\forall\, q \in P \quad J(u) + q \cdot F(u) \le J(u) + p \cdot F(u) \le J(v) + p \cdot F(v) \quad \forall\, v \in U\ . \tag{2.64}$$

Examinons d'abord le cas de contraintes d'égalité. Puisque $P = \mathbb{R}^M$, la première inégalité dans (2.64) montre que $F(u) = 0$, i.e. $u \in K$. Il reste alors $J(u) \leq J(v) + p \cdot F(v)$, pour tout $v \in U$, 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 (2.64) 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 (2.64) 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 (2.46), c'est-à-dire

$$K = \{v \in V \text{ tel que } F_i(v) \leq 0 \text{ pour } 1 \leq i \leq M\} \ . \tag{2.65}$$

Le Théorème 2.5.18 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 2.3.10 dès que K est non vide et que la condition "infinie à l'infini" (2.10) 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 2.5.18 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, défini par (2.65). 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 2.5.16. 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*

$$F(u) \leq 0 \ , \ p \geq 0 \ , \ p \cdot F(u) = 0 \ , \ J'(u) + \sum_{i=1}^{M} p_i F_i'(u) = 0 \ . \tag{2.66}$$

Démonstration. Si u est un minimum de J sur K, on peut appliquer le Théorème 2.5.18, qui donne exactement la condition d'optimalité (2.66), 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. •

Comme dans le cas des contraintes d'égalité (voir le Lemme 2.5.13), on peut donner une interprétation concrète des multiplicateurs de Lagrange λ_i dans le Théorème 2.5.18 : ils donnent la sensibilité de la valeur minimale de J aux variations du niveau des contraintes F_i (au signe près). Pour un niveau des contraintes $\epsilon \in \mathbb{R}^M$, on considère le problème d'optimisation

$$\inf_{v \in V, \ F(v) \leq \epsilon} J(v). \tag{2.67}$$

On note $J_{\min}(\epsilon)$ la valeur minimum dans (2.67).

Lemme 2.6.6 *On se place sous les hypothèses du Théorème 2.6.4 de Kuhn et Tucker et on suppose que, pour* ϵ *petit, le problème* (2.67) *admet une solution unique, notée* u_ϵ. *On suppose aussi que la valeur minimale* $J_{\min}(\epsilon)$ *de* (2.67) *est différentiable par rapport à* ϵ *en 0. Alors, si on note* u *la solution pour* $\epsilon = 0$ *et* $\lambda \in \mathbb{R}^M$ *le multiplicateur de Lagrange associé, on a, pour* $1 \leq i \leq M$,

$$\lambda_i = -\frac{\partial J_{\min}}{\partial \epsilon_i}(0).$$

Démonstration. D'après le Théorème 2.5.18, la solution u du problème (2.67) non perturbé, c'est-à-dire pour $\epsilon = 0$, vérifie la condition d'optimalité pour un certain multiplicateur de Lagrange $\lambda \geq 0$

$$J'(u) + \lambda \cdot F'(u) = 0, \quad \lambda \cdot F(u) = 0. \tag{2.68}$$

Comme les fonctions J et F sont convexes et que $\lambda \geq 0$, pour tout v, on a

$$J(v) + \lambda \cdot F(v) - J(u) - \lambda \cdot F(u) \geq \langle J'(u) + \lambda \cdot F'(u), v - u \rangle.$$

D'après la condition d'optimalité (2.68), on a donc

$$J(v) + \lambda \cdot F(v) - J(u) \geq 0.$$

On prend alors $v = u_\epsilon$, la solution du problème perturbé (2.67), et comme $F(u_\epsilon) \leq \epsilon$ et $\lambda \geq 0$, on en déduit que

$$J_{\min}(\epsilon) + \lambda \cdot \epsilon - J_{\min}(0) \geq 0. \tag{2.69}$$

Comme on a supposé que l'application $\epsilon \mapsto J_{\min}(\epsilon)$ est dérivable, on a

$$J_{\min}(\epsilon) = J_{\min}(0) + \frac{\partial J_{\min}}{\partial \epsilon}(0) \cdot \epsilon + o(\epsilon).$$

L'inégalité (2.69) devient, pour tout ϵ,

$$\left(\frac{\partial J_{\min}}{\partial \epsilon}(0) + \lambda \right) \cdot \epsilon + o(\epsilon) \geq 0.$$

En divisant cette inéquation par la norme de ϵ et en faisant tendre ϵ vers 0, on obtient que pour tout vecteur e de norme unité,

$$\left(\frac{\partial J_{\min}}{\partial \epsilon}(0) + \lambda \right) \cdot e \geq 0.$$

En appliquant cette inégalité à $-e$ au lieu de e, on en déduit l'égalité recherchée. Remarquons que, puisque $\lambda \geq 0$, la valeur minimum $J_{\min}(\epsilon)$ ne peut que décroitre lorsque ϵ croît, ce qui est normal puisque l'ensemble des solutions admissibles croît. □

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 2.6.7 *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 $U \times P$*

$$\forall\, q \in P \quad \mathcal{L}(u, q) \leq \mathcal{L}(u, p) \leq \mathcal{L}(v, p) \quad \forall\, v \in U\ . \tag{2.70}$$

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

$$\mathcal{J}(v) = \sup_{q \in P} \mathcal{L}(v,q) \qquad \mathcal{G}(q) = \inf_{v \in U} \mathcal{L}(v,q) \ . \tag{2.71}$$

On appelle problème primal le problème de minimisation

$$\inf_{v \in U} \mathcal{J}(v) \ , \tag{2.72}$$

et problème dual le problème de maximisation

$$\sup_{q \in P} \mathcal{G}(q) \ . \tag{2.73}$$

Remarque 2.6.8 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 2.6.7 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 (2.70) 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 2.6.9 (de dualité) *Le couple (u,p) est un point-selle de $\mathcal{L}$ sur $U \times P$ si et seulement si*

$$\mathcal{J}(u) = \min_{v \in U} \mathcal{J}(v) = \max_{q \in P} \mathcal{G}(q) = \mathcal{G}(p) \ . \tag{2.74}$$

Remarque 2.6.10 Par la Définition (2.71) de $\mathcal{J}$ et $\mathcal{G}$, (2.74) est équivalent à

$$\mathcal{J}(u) = \min_{v \in U} \left(\sup_{q \in P} \mathcal{L}(v,q) \right) = \max_{q \in P} \left(\inf_{v \in U} \mathcal{L}(v,q) \right) = \mathcal{G}(p) \ . \tag{2.75}$$

Si le sup et l'inf sont atteints dans (2.75) (c'est-à-dire qu'on peut les écrire max et min, respectivement), on voit alors que (2.75) 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 (2.71) que $\mathcal{J}(v) \geq \mathcal{L}(v,p)$, d'où $\mathcal{J}(v) \geq \mathcal{L}^*$ d'après (2.70). 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 (2.74) a lieu et posons $\mathcal{L}^* = \mathcal{J}(u)$. La définition (2.71) de $\mathcal{J}$ montre que

$$\mathcal{L}(u,q) \leq \mathcal{J}(u) = \mathcal{L}^* \quad \forall \, q \in P \ . \tag{2.76}$$

De même, on a aussi :

$$\mathcal{L}(v,p) \geq \mathcal{G}(p) = \mathcal{L}^* \quad \forall \, v \in U \ , \tag{2.77}$$

et on déduit facilement de (2.76)-(2.77) que $\mathcal{L}(u,p) = \mathcal{L}^*$, ce qui montre que (u,p) est point-selle. □

Remarque 2.6.11 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**

$$\inf_{v\in U}\left(\sup_{q\in P}\mathcal{L}(v,q)\right) \geq \sup_{q\in P}\left(\inf_{v\in U}\mathcal{L}(v,q)\right). \tag{2.78}$$

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 (2.78). La différence (positive) entre les deux membres de l'inégalité (2.78) est appelée **saut de dualité**. •

Exercice 2.6.1 Soit F une fonction bornée non constante de $\mathbb{R}$ dans $\mathbb{R}$. On définit sur $\mathbb{R}\times\mathbb{R}$ le Lagrangien $\mathcal{L}(v,q) = F(v+q)$. Vérifier que pour ce Lagrangien l'inégalité (2.78) est stricte avec ses deux membres finis.

Le Théorème 2.6.9 de dualité affirme que l'existence d'un point selle pour un Lagrangien $\mathcal{L}(v,q)$ est équivalente à l'échange du min en v et du max en q pour son optimisation. On peut alors se poser la question des conditions sur le Lagrangien pour l'existence d'un point selle. Le Théorème 2.6.4 de Kuhn et Tucker donne un premier exemple d'existence d'un point selle pour un Lagrangien du type $\mathcal{L}(v,q) = J(v) + q \cdot F(v)$, défini sur $V \times (\mathbb{R}_+)^M$, avec J fortement convexe et $F_1, \ldots, F_M$ convexes. En effet, dans ce cas il existe un unique minimum global de $J(v)$ sous les contraintes $F(v) \leq 0$ (si l'ensemble K, défini par (2.65) n'est pas vide). Le résultat suivant donne une réponse plus générale sous l'hypothèse cruciale que le Lagrangien est convexe en v et concave en q.

Proposition 2.6.12 *Soit U (respectivement P) un convexe compact non vide de V (respectivement Q). Soit un Lagrangien $\mathcal{L}(v,q)$ défini et continu sur $U \times P$. On suppose que, pour tout $q \in P$, $v \mapsto \mathcal{L}(v,q)$ est strictement convexe sur U et que, pour tout $v \in U$, $q \mapsto \mathcal{L}(v,q)$ est concave sur P. Alors il existe un point selle de $\mathcal{L}$ sur $U \times P$.*

Démonstration. Pour tout $q \in P$, il existe un unique point de minimum, noté $v^{\,\mathrm{opt}}(q) \in U$, de l'application $v \mapsto \mathcal{L}(v,q)$ qui est strictement convexe et continue sur le compact U. On pose

$$\mathcal{G}(q) = \mathcal{L}(v^{\,\mathrm{opt}}(q), q) = \min_{v\in U}\mathcal{L}(v,q). \tag{2.79}$$

L'application $\mathcal{G}$ est concave comme minimum d'une famille de fonctions concaves continues. Comme elle ne prend pas de valeurs infinies, elle est continue d'après le Lemme 2.3.5. Comme P est compact et que $\mathcal{G}$ est continue, $\mathcal{G}$ admet au moins un point de maximum sur P noté q^*. On note $v^* = v^{\,\mathrm{opt}}(q^*)$. On va montrer que (v^*, q^*) est un point selle de $\mathcal{L}$ sur $U \times P$, c'est-à-dire que

$$\mathcal{L}(v^*, q) \leq \mathcal{L}(v^*, q^*) \leq \mathcal{L}(v, q^*) \tag{2.80}$$

pour tout couple $(v,q) \in U \times P$. La deuxième inégalité de (2.80) est évidente et découle simplement de la définition de $v^* = v^{\,\mathrm{opt}}(q^*)$. Il reste à prouver la première

inégalité de (2.80). Pour tout $\theta \in [0,1]$ et tout $q \in P$, on pose

$$v_\theta^{\rm opt} = v^{\rm opt}((1-\theta)q^* + \theta q)$$

et, puisque q^* maximise $\mathcal{G}$ sur P, on a

$$\mathcal{G}(q^*) \geq \mathcal{G}((1-\theta)q^* + \theta q) = \mathcal{L}(v_\theta^{\rm opt}, (1-\theta)q^* + \theta q).$$

La concavité de $q \mapsto \mathcal{L}(v_\theta^{\rm opt}, q)$ conduit alors à

$$\mathcal{G}(q^*) \geq \mathcal{L}(v_\theta^{\rm opt}, (1-\theta)q^* + \theta q) \geq (1-\theta)\mathcal{L}(v_\theta^{\rm opt}, q^*) + \theta\mathcal{L}(v_\theta^{\rm opt}, q).$$

Or, par définition (2.79) de $\mathcal{G}$, on a $\mathcal{L}(v_\theta^{\rm opt}, q^*) \geq \mathcal{G}(q^*)$, d'où

$$\mathcal{G}(q^*) \geq (1-\theta)\mathcal{G}(q^*) + \theta\mathcal{L}(v_\theta^{\rm opt}, q),$$

ce qui donne par soustraction et division par $\theta \neq 0$

$$\mathcal{G}(q^*) \geq \mathcal{L}(v_\theta^{\rm opt}, q) \quad \text{ pour tout } 0 < \theta \leq 1.$$

On ne peut malheureusement pas prendre $\theta = 0$ dans l'inégalité ci-dessus et conclure simplement que $\mathcal{G}(q^*) = \mathcal{L}(v^*, q^*) \geq \mathcal{L}(v^*, q)$, puisque $v_0^{\rm opt} = v^*$. Il faut donc utiliser un argument de "passage à la limite" quand θ tend vers zéro. Comme U est compact, il existe une suite θ_n qui tend vers zéro tel que $v_{\theta_n}^{\rm opt}$ converge vers une limite notée $\tilde{v}$. En passant à la limite dans l'inégalité précédente, on obtient

$$\mathcal{G}(q^*) = \mathcal{L}(v^*, q^*) \geq \lim_{n\to+\infty} \mathcal{L}(v_{\theta_n}^{\rm opt}, q) = \mathcal{L}(\tilde{v}, q).$$

Pour conclure, il suffit donc de prouver que $\tilde{v} = v^*$ et ainsi obtenir la première inégalité de (2.80). Par définition (2.79) de $\mathcal{G}$, on a pour tout $v \in U$

$$\mathcal{L}(v_{\theta_n}^{\rm opt}, (1-\theta_n)q^* + \theta_n q) \leq \mathcal{L}(v, (1-\theta_n)q^* + \theta_n q).$$

Par continuité du Lagrangien, on peut passer à la limite $\theta_n \to 0$ pour obtenir

$$\mathcal{L}(\tilde{v}, q^*) \leq \mathcal{L}(v, q^*).$$

Ainsi, $\tilde{v}$ est un minimiseur de $v \mapsto \mathcal{L}(v, q^*)$. Comme cette dernière application est strictement convexe, elle admet au plus un minimiseur et $\tilde{v} = v^{\rm opt}(q^*) = v^*$. □

Application

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

$$\inf_{v\in V\ ,\ F(v)\leq 0} J(v) \tag{2.81}$$

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 $v \in V$,

$$\mathcal{J}(v) = \sup_{q \in (\mathbb{R}_+)^M} \mathcal{L}(v,q) = \begin{cases} J(v) & \text{si } F(v) \leq 0 \\ +\infty & \text{sinon ,} \end{cases} \tag{2.82}$$

ce qui montre que le problème primal $\inf_{v \in V} \mathcal{J}(v)$ est exactement le problème d'origine (2.81) ! D'autre part, la fonction $\mathcal{G}(q)$ du problème dual est bien définie par (2.71), car (2.71) 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 (2.81) 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 2.6.9 nous donne le résultat suivant.

Corollaire 2.6.13 *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 2.5.16. 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*

$$F(u) \leq 0 \ , \ p \geq 0 \ , \ p \cdot F(u) = 0 \ , \ J'(u) + p \cdot F'(u) = 0 \ . \tag{2.83}$$

L'application la plus courante du Corollaire 2.6.13 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 2.6.13 il suffit d'ajouter une hypothèse de forte convexité ou de comportement infini à l'infini sur J.

Remarque 2.6.14 Pour illustrer le Corollaire 2.6.13 et l'intérêt de la dualité, nous considérons un problème de minimisation quadratique dans $\mathbb{R}^N$ avec contraintes d'inégalité affines

$$\min_{v\in\mathbb{R}^N,\ F(v)=Bv-c\leq 0} \left\{ J(v) = \frac{1}{2}Av\cdot v - b\cdot v \right\} , \tag{2.84}$$

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

$$\mathcal{L}(v,q) = \frac{1}{2}Av\cdot v - b\cdot v + q\cdot(Bv-c) \quad \forall\,(v,q)\in\mathbb{R}^N\times(\mathbb{R}_+)^M\ . \tag{2.85}$$

Nous avons déjà fait dans (2.82) le calcul de $\mathcal{J}$, et dit que le problème primal est exactement (2.84). 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

$$\sup_{q\geq 0}\left(-\frac{1}{2}q\cdot BA^{-1}B^*q + (BA^{-1}b-c)\cdot q - \frac{1}{2}A^{-1}b\cdot b\right)\ . \tag{2.86}$$

Certes, la fonctionnelle à maximiser dans (2.86) n'a pas une allure particulièrement sympathique. Il s'agit encore d'un problème avec fonctionnelle quadratique et contraintes affines. Cependant, le Corollaire 2.6.13 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 (2.86) 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 3.4 que cet avantage peut être utilisé pour mettre au point un algorithme de calcul de la solution du problème primal. •

Remarque 2.6.15 Une autre manière, très simple, d'éliminer des contraintes d'inégalité non-linéaires est l'introduction de variables supplémentaires, dites d'écart (ou "slack variables" en anglais). On remplace les contraintes $F_i(v)\leq 0$, pour $1\leq i\leq M$, par

$$F_i(v) + z_i = 0 \quad \text{et} \quad z_i\geq 0\,,$$

puis on optimise par rapport au couple $(v,z)\in V\times(\mathbb{R}_+)^M$ avec ces nouvelles contraintes d'égalité et d'inégalité (mais très simples pour z). Cette manière de faire n'est pas très économe car on augmente le nombre de variables et de contraintes (nous reverrons cette idée en programmation linéaire, voir le chapitre 4). •

Exercice 2.6.2 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

$$G(\tau) = \frac{1}{2}\int_0^L |\tau|^2 dx. \tag{2.87}$$

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

$$\inf_{-\tau'=f \text{ dans } (0,L)} \left\{ G(\tau) = \frac{1}{2}\int_0^L |\tau|^2 dx \right\}. \tag{2.88}$$

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 (2.88) admette un point de minimum σ, montrer que (σ, u) est un point selle du Lagrangien et que $\sigma = u'$.

Chapitre 3

ALGORITHMES D'OPTIMISATION

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. Le lecteur trouvera qu'il y a peut-être beaucoup trop d'algorithmes présentés ici, mais il doit savoir que chacun a son utilité et son efficacité en pratique sur tel ou tel problème. Aucun algorithme n'est universellement meilleur qu'un autre et le choix d'un algorithme, plutôt qu'un autre, dépend en particulier de la régularité de la fonction à minimiser (et des contraintes) et de la taille du problème à résoudre.

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 $(u^n)_{n\in\mathbb{N}}$ vers u quel que soit le choix de la donnée initiale u^0), nous étudierons leur vitesse de convergence.

Dans toute ce chapitre nous supposerons que la fonction objectif à minimiser J est convexe. Le rôle de la convexité est absolument essentiel : si on applique ces algorithmes à la minimisation de fonctions non convexes (ce qui arrive souvent en pratique), on peut se heurter à des difficultés sérieuses. 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. Très souvent nous demanderons que la fonction J soit α-convexe, différentiable et que sa dérivée soit L-Lipschitzienne. Ces hypothèses sont assez fortes, mais nous verrons qu'elles sont cruciales pour les démonstrations de convergence des algorithmes et qu'elles conditionnent les vitesses de convergence que l'on peut obtenir. Rappelons que ces hypothèses permettent d'encadrer en tout point u la fonction J par deux fonctions quadratiques qui lui sont tangentes en u (voir les Remarques 2.4.6 et 2.4.11, ainsi que les Figures 2.2 et 2.3).

Remarque 3.0.1 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 schématiser 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). Nous n'évoquons pas, non plus, les algorithmes qui n'utilisent pas de dérivées car ils sont très peu efficaces en pratique. •

3.1 Algorithmes de type gradient (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 2.3.9 il existe une unique solution u, caractérisée d'après la Remarque 2.5.2 par l'équation d'Euler $J'(u) = 0$. Pour démontrer la convergence des algorithmes nous utiliserons très souvent une hypothèse de Lipschitziannité de la dérivée J' (voir la Définition 2.4.9).

3.1.1 Algorithme de gradient à pas optimal

L'algorithme de gradient construit une suite minimisante u^n à l'aide d'une récurrence à un seul pas, c'est-à-dire que la nouvelle itérée u^{n+1} ne dépend que de la précédente u^n. On passe de u^n à u^{n+1} 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).

Ce simple constat conduit à choisir parmi les méthodes du type (3.2), qui sont appelées "méthodes de descente", la méthode de gradient ou de plus grande pente (*steepest descent* en anglais) dans laquelle w^n est positivement proportionnelle à

$J'(u^n)$. Le choix du coefficient de proportionnalité, ou pas de descente, distingue plusieurs algorithmes. Ici, nous présentons 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 $\mu^n \in \mathbb{R}$, appelé **pas de descente**, est choisi à chaque étape tel que

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

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

Théorème 3.1.1 *On suppose que J est α-convexe différentiable et que J' est Lipschitzien sur tout borné de V, c'est-à-dire que*

$$\forall\, M > 0,\ \exists\, L_M > 0,\ \|v\| + \|w\| \le M \ \Rightarrow\ \|J'(v) - J'(w)\| \le L_M \|v - w\| . \tag{3.5}$$

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

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

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

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

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

$$\|u^n\| \le M .$$

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

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

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

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

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

Remarque 3.1.2 L'hypothèse (3.5) est plus faible que la L-Lipschitziannité de la dérivée J' (voir la Définition 2.4.9) car elle est locale et non globale. Bien sûr, si J est de classe C^2 sur V, alors elle est automatiquement vérifiée. •

Remarque 3.1.3 De manière générale on dira qu'un algorithme d'optimisation converge s'il converge quelle que soit l'initialisation u^0. En effet, cela n'aurait pas de sens de développer une théorie de convergence sous des conditions restrictives sur le vecteur initial u^0 (par exemple qu'il appartienne à un certain sous-ensemble de V) car en pratique les inévitables erreurs d'arrondi du calcul numérique sur ordinateur empêcherait que ces conditions soient vérifiées exactement et la suite des itérées u^n pourrait ne pas converger. Néanmoins nous verrons parfois des algorithmes qui ne convergent que si u^0 est proche de la solution (comme l'algorithme de Newton). •

Remarque 3.1.4 Dans (3.2) la direction w^n, positivement proportionnelle à $J'(u^n)$, n'est pas la seule direction de descente possible, c'est-à-dire qui garantisse la décroissance de la fonction coût J pour un pas μ^n petit. Pour s'en rendre compte de manière simple, 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 3.1.5 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 pour sa résolution (voir [26]). En pratique, on n'a souvent pas besoin de trouver le pas optimal μ^n et on se contente d'un pas $\tilde{\mu}^n$ qui fait "suffisamment" décroitre la fonction $\mu \mapsto J\big(u^n - \mu\, J'(u^n)\big)$. Cette idée est explorée dans la Sous-section 3.1.3 ci-dessous. •

On peut se demander si la suite des pas optimaux μ^n reste bornée, loin de 0 et de $+\infty$, ou pas. Les trois exercices suivants montrent, d'une part que c'est le cas pour une fonction objectif quadratique ou bien α-convexe à dérivée Lipschitzienne et, d'autre part, que le pas peut tendre vers $+\infty$ pour une fonction objectif très "plate" près de son minimum.

Exercice 3.1.1 On applique l'algorithme du gradient à pas optimal à la fonction $J(x) = \frac{1}{2}Ax \cdot x - b \cdot x$, définie de $\mathbb{R}^N$ dans $\mathbb{R}$, avec $b \in \mathbb{R}^N$ et A une matrice symétrique définie positive de valeurs propres $0 < \lambda_1 \leq \cdots \leq \lambda_N$. Montrer qu'à chaque itération n le pas optimal vérifie $1/\lambda^N \leq \mu^n \leq 1/\lambda^1$.

Exercice 3.1.2 On applique l'algorithme du gradient à pas optimal à une fonction J α-convexe différentiable et telle que sa dérivée J' est L-Lipschitzienne sur V (voir la Définition 2.4.9). En utilisant l'encadrement de J par deux fonctions quadratiques (cf. la Remarque 2.4.11), montrer que la suite des pas optimaux vérifie $0 < c \leq \mu^n \leq C < +\infty$ pour deux constantes c, C indépendantes de n.

Exercice 3.1.3 Soit $J(x) = \frac{1}{2p}(Ax \cdot x)^p$, définie de $\mathbb{R}^N$ dans $\mathbb{R}$, pour un entier $p \geq 2$ et A une matrice symétrique définie positive. Montrer que J est strictement convexe, que son minimum est atteint en 0 et qu'elle n'est pas fortement convexe en 0. On applique l'algorithme du gradient à pas optimal à J et on suppose que la suite des itérées vérifie $\|x^n\| = \epsilon^n \to 0$ avec $\epsilon^n > 0$. Montrer que la suite des pas optimaux tend vers l'infini comme $\mu^n = \mathcal{O}\left((\epsilon^n)^{-2(p-1)}\right)$.

3.1.2 Algorithme de gradient à pas fixe

L'algorithme de gradient à pas fixe est une variante du précédent où l'on utilise un pas de descente fixe $\mu > 0$, indépendant du numéro d'itération n. Autrement dit, on construit une suite u^n définie par

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

à partir d'une initialisation $u^0 \in V$. 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 affirme que, si le pas de descente μ est suffisamment petit, alors l'algorithme (3.8) converge.

Théorème 3.1.6 *On suppose que J est α-convexe différentiable et que J' est Lipschitzien sur V, c'est-à-dire qu'il existe une constante $L > 0$ telle que*

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

Alors, si $0 < \mu < 2\alpha/L^2$, 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)

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

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

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

d'après (3.9) et l'α-convexité. Si $0 < \mu < 2\alpha/L^2$, il est facile de voir que $1 - 2\alpha\mu + L^2\mu^2 \in]0,1[$, et la convergence se déduit de (3.10). En fait, le calcul qui conduit à (3.10), s'il est appliqué à deux vecteurs quelconques à la place de u^n et u montre que l'application $v \mapsto v - \mu J'(v)$ est strictement contractante lorsque $0 < \mu < 2\alpha/L^2$, donc elle admet un unique point fixe u vers lequel converge la suite u^n. □

Remarque 3.1.7 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. Par contre, on ne connait en général pas de bonne estimation des constantes α et L et on ne sait pas a priori si un choix de pas de descente $\mu > 0$ est suffisamment petit. •

Remarque 3.1.8 Le Théorème 3.1.6 donne une estimation de la **vitesse de convergence** de l'algorithme qui est dite **géométrique** car la quantité $\|u^n - u\|^{1/n}$ admet une limite finie, inférieure à $\gamma < 1$, lorsque n tend vers $+\infty$. 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. On peut optimiser le choix du pas de descente μ dans l'intervalle $]0, 2\alpha/L^2[$ qui minimise la valeur de γ. On trouve facilement que le pas optimal est $\mu = \alpha/L^2$ qui conduit à $\gamma = \sqrt{1 - \alpha^2/L^2}$. Mais cette valeur de γ n'est pas la vitesse de convergence optimale de l'algorithme du gradient à pas fixe car on a privilégié une démonstration simple du Théorème 3.1.6. Une meilleure constante sera donnée dans la Proposition 3.2.1. Notons que, de toute façon en pratique, on ne connait souvent pas les valeurs de α et L et donc qu'il est difficile de choisir le pas optimal μ. L'Exercice 3.1.5 montre, sur un exemple quadratique, que l'algorithme du gradient à pas fixe ne peut pas converger plus vite que cette convergence géométrique (mais avec une meilleure valeur de γ). On verra par la suite d'autres algorithmes où on peut améliorer la constante γ 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 3.1.4 On se place sous les hypothèses du Théorème 3.1.6. 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,

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

Exercice 3.1.5 Soit A une matrice symétrique définie positive de taille $N \times N$, de valeurs propres ordonnées $0 < \lambda_1 \leq \cdots \leq \lambda_N$, et 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'algorithme

converge si et seulement si le pas fixe vérifie $0 < \mu < 2/\lambda_N$, et que la meilleure vitesse de convergence de l'algorithme s'obtient pour le choix du pas de descente $\mu_B = 2/(\lambda_1+\lambda_N)$, qui conduit à

$$\|x^n - x^*\| \leq \gamma_B^n \|x^0 - x^*\| \quad \text{avec} \quad \gamma_B = \frac{1 - \lambda_1/\lambda_N}{1 + \lambda_1/\lambda_N}.$$

Comparer avec la valeur optimisée de γ dans la Remarque 3.1.8 (identifier les valeurs L et α pour la fonction $J(x)$).

La proposition suivante est une version affaiblie du Théorème 3.1.6 où on ne fait plus l'hypothèse que la fonction J est α-convexe mais seulement qu'elle est strictement convexe. De plus, on se restreint à la dimension finie car on utilise un argument de compacité simple (mais le résultat reste vraie en dimension infinie avec une preuve plus compliquée).

Proposition 3.1.9 *Soit une fonction $J(u)$ définie sur $V = \mathbb{R}^N$, strictement convexe, différentiable, infinie à l'infini et telle que J' est Lipschitzien, c'est-à-dire vérifie (3.9) pour une constante $L > 0$. Alors, si $0 < \mu < 2/L$, 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. Puisque $u^{n+1} = u^n - \mu J'(u^n)$ on applique le Lemme 3.1.10 avec $v = u^n$ et $v^* = u^{n+1}$, ce qui conduit à

$$\mu(1 - L\mu/2)\|J'(u^n)\|^2 \leq J(u^n) - J(u^{n+1}).$$

On en déduit que, pour $0 < \mu < 2/L$, la suite $J(u^n)$ est décroissante et, comme J est infinie à l'infini, elle est minorée. Donc la suite $J(u^n)$ converge vers une limite finie et la suite $\|J'(u^n)\|$ tends vers 0. Puisque J est infinie à l'infini, la suite u^n est bornée et, comme $V = \mathbb{R}^N$, il existe une sous-suite convergente, notée $u^{n'}$. Alors, J' étant continue (car Lipschitzien), la limite de la sous-suite $u^{n'}$ est un zéro de J'. Or, J est strictement convexe donc admet au plus un point de minimum, c'est-à-dire que J' ne s'annule qu'en un seul point u. Par conséquent, toutes les sous-suites convergentes de u^n convergent vers la même limite u. Autrement dit, toute la suite u^n converge vers u qui est l'unique point de minimum de J. □

Lemme 3.1.10 *Soit J une fonction différentiable, de dérivée L-Lipschitzienne. Soit $v \in V$ quelconque et $v^* = v - \mu J'(v)$. Alors*

$$J(v) - J(v^*) \geq \mu(1 - L\mu/2)\|J'(v)\|^2.$$

Démonstration. En vertu du Lemme 2.4.10 la fonction $J(v^*)$ est majorée par une fonction quadratique de v^*

$$J(v^*) \leq J(v) + \langle J'(v), v^* - v\rangle + \frac{L}{2}\|v^* - v\|^2.$$

On remplace $v^* - v$ par $-\mu J'(v)$ pour obtenir le résultat. □

Remarque 3.1.11 Le Lemme 3.1.10 est une garantie de décroissance d'un pas de l'algorithme du gradient, sans aucune condition de convexité sur la fonction. •

Exercice 3.1.6 Démontrer la Proposition 3.1.9 dans le cas d'un espace de Hilbert V de dimension infinie. Indication : on utilisera la notion de convergence faible et le Lemme 2.3.16.

Si on ne suppose même plus que la fonction J est convexe, alors on ne peut pas garantir que la suite u^n converge vers un point de minimum.

Exercice 3.1.7 Soit une fonction $J(u)$ différentiable, infinie à l'infini et telle que J' est Lipschitzien, c'est-à-dire vérifie (3.9) pour une constante $L > 0$. On ne suppose pas que J est convexe. A l'aide des mêmes raisonnements que pour la Proposition 3.1.9, montrer que l'algorithme du gradient à pas fixe $0 < \mu < 2/L$ vérifie que $J(u^n)$ décroit et $J'(u^n)$ tend vers 0.

L'exercice suivant donne un exemple de fonction J pour lequel l'algorithme de gradient converge seulement vers un point d'inflexion et pas vers un point de minimum.

Exercice 3.1.8 Soit la fonction $J(x) = \frac{1}{2}x_1^2 + \frac{1}{4}x_2^4 - \frac{1}{2}x_2^2$ définie sur $\mathbb{R}^2$. Vérifier que J n'est pas convexe et que l'ensemble de ses points critiques (où son gradient s'annule) sont $O = (0,0)$, $P = (0,-1)$ et $Q = (0,1)$. Montrer que P et Q sont des points de minimum tandis que O est un point selle. Montrer que, si on initialise un algorithme de gradient au point $x^0 = (1,0)$, alors quelque soit la règle de choix du pas de descente $\mu^n > 0$ (fixe, optimal ou autre) qui assure la convergence la limite est le point O qui n'est pas un minimum, même local, de J.

3.1.3 Algorithme de gradient à pas variable

L'algorithme du gradient à pas fixe est nettement plus simple que celui à pas optimal puisqu'il ne nécessite pas de recherche linéaire du meilleur pas à chaque itération. Néanmoins, l'algorithme du gradient à pas fixe souffre d'une difficulté majeure car il ne converge que si le pas μ est plus petit qu'une valeur prescrite par les constantes α de forte convexité et L de Lipschitziannité. Or, en pratique, ces constantes ne sont pas connues et il est donc tout à fait possible que le pas choisi se révèle trop grand et que l'algorithme ne converge pas. A contrario, un choix trop prudent d'un pas trop petit conduit, certes à la convergence de l'algorithme, mais de manière beaucoup trop lente puisque la vitesse de convergence est d'autant plus faible que le pas est petit (le coefficient $\gamma < 1$ dans le Théorème 3.1.6 converge vers 1 quand μ tends vers 0).

Pour cette raison, il est absolument nécessaire **d'adapter** le pas de descente afin de corriger une possible erreur de choix initial et d'avoir à chaque itération, non pas la valeur optimale du pas, mais une valeur garantissant une décroissance suffisante de fonction coût J. On présente donc une variante des algorithmes précédents, appelée gradient à pas variable. Cette variante est basée sur une condition de décroissance suffisante, appelée **règle d'Armijo**

Définition 3.1.12 *Soit $v \in V$ et $v^* = v - \mu\, J'(v)$. On dit que le pas $\mu > 0$ vérifie la règle d'Armijo si*

$$J(v^*) \leq J(v) - \frac{\mu}{2}\|J'(v)\|^2. \tag{3.11}$$

La condition (3.11) garantit bien une décroissance stricte de J, sauf si $J'(v) = 0$ (auquel cas v est le point de minimum si J est convexe). Le lemme suivant montre que les pas suffisamment petits vérifient la règle d'Armijo.

Lemme 3.1.13 *Soit J une fonction différentiable, de dérivée L-Lipschitzienne. Soit $v \in V$ quelconque et $v^* = v - \mu J'(v)$. Si $0 < \mu \leq \frac{1}{L}$, alors le pas μ vérifie la règle d'Armijo.*

Démonstration. La démonstration est similaire à celle du Lemme 3.1.10. On majore la fonction $J(v^*)$ grâce au Lemme 2.4.10

$$J(v^*) \leq J(v) + \langle J'(v), v^* - v\rangle + \frac{L}{2}\|v^* - v\|^2 \leq J(v) - \mu\left(1 - \frac{L\mu}{2}\right)\|J'(v)\|^2$$

d'où le résultat si $0 < \mu \leq 1/L$. □

On peut maintenant définir **l'algorithme du gradient à pas variable** A partir de u^0 quelconque dans V et d'un pas $\mu > 0$, on construit la suite $(u^n)_{n\geq 0}$ définie par

$$u^{n+1} = u^n - \mu^n\, J'(u^n)\,, \tag{3.12}$$

où $\mu^n = \frac{\mu}{2^{i_n}}$ avec $i_n \geq 0$ le plus petit entier tel que μ^n vérifie la règle d'Armijo. Autrement dit, à chaque itération on teste successivement les pas μ, $\frac{\mu}{2}$, $\frac{\mu}{4}$, $\frac{\mu}{8}$, ... jusqu'à ce que l'un d'entre eux vérifie la condition de décroissance (3.11).

En vertu du Lemme 3.1.13 on sait qu'il existe un tel entier i_n mais on n'a pas besoin de connaitre la valeur de L pour tester la règle d'Armijo. Par rapport à l'algorithme du gradient à pas fixe, le coût supplémentaire à chaque itération pour l'algorithme du gradient à pas variable est de faire i_n évaluations supplémentaires de la fonction J (en général, on calcule toujours $J(u^n)$ pour vérifier la convergence des algorithmes).

Le résultat suivant affirme que, pour n'importe quel pas de descente initial $\mu > 0$, l'algorithme (3.12) converge. C'est donc une amélioration significative du Théorème 3.1.6 puisqu'il n'y a pas besoin de connaitre les valeurs de α et L pour choisir μ et garantir la convergence.

Théorème 3.1.14 *On suppose que J est α-convexe différentiable et que J' est Lipschitzien sur V de constante $L > 0$, c'est-à-dire que (3.9) est vérifié. Alors l'algorithme de gradient à pas variable converge : quel que soit u^0, la suite (u^n) définie par (3.12) converge vers la solution u de (3.1)*

$$\|u^n - u\| \leq \gamma^n \|u^0 - u\| \quad \textit{avec} \quad \gamma = \sqrt{1 - \alpha/(2L)} < 1.$$

Démonstration. Posons $v^n = u^n - u$ qui vérifie $v^{n+1} = v^n - \mu^n J'(u^n)$, d'où il vient

$$\|v^{n+1}\|^2 = \|v^n\|^2 + 2\mu^n\langle J'(u^n), u - u^n\rangle + (\mu^n)^2\big\|J'(u^n)\big\|^2. \tag{3.13}$$

Pour majorer le deuxième terme du membre de droite de (3.13), on utilise l'α-convexité (2.23) de J en u^n

$$\langle J'(u^n), u - u^n\rangle \leq J(u) - J(u^n) - \frac{\alpha}{2}\|u^n - u\|^2.$$

Pour majorer le troisième terme du membre de droite de (3.13), on utilise la règle d'Armijo

$$\frac{\mu^n}{2}\big\|J'(u^n)\big\|^2 \leq J(u^n) - J(u^{n+1}).$$

On déduit alors de (3.13)

$$\begin{aligned}\|v^{n+1}\|^2 &\leq \|v^n\|^2 + 2\mu^n\big(J(u) - J(u^n)\big) - \alpha\mu^n\|v^n\|^2 + 2\mu^n\big(J(u^n) - J(u^{n+1})\big)\\ &\leq (1 - \alpha\mu^n)\,\|v^n\|^2\,,\end{aligned}$$

car $J(u) - J(u^{n+1}) \leq 0$. On sait que $\mu^n = \mu/2^{i_n}$ où i_n est le plus petit entier qui permet de vérifier la règle d'Armijo. Donc $\mu/2^{i_n-1} = 2\mu^n$ ne vérifie pas la règle d'Armijo. En raison du Lemme 3.1.13 ce pas de descente ne peut pas être plus petit que $1/L$, c'est-à-dire que $2\mu^n > 1/L$. Autrement dit, $1 - \alpha\mu^n < 1 - \alpha/(2L) < 1$, ce qui donne la vitesse de convergence annoncée. □

Remarque 3.1.15 La démonstration du Théorème 3.1.14 ressemble à celle du Théorème 3.1.6, sauf qu'on n'y utilise pas explicitement le caractère L-Lipschitzien de J'. Celui-ci est néanmoins présent (et nécessaire) à travers l'utilisation du Lemme 3.1.13 sur la règle d'Armijo. •

Remarque 3.1.16 On peut se demander d'où vient la forme particulière de la condition (3.11) de la règle d'Armijo. Il est clair que cette condition doit exprimer une décroissance de la fonction coût mais pourquoi cette dernière doit-elle être proportionnelle à la norme au carré du gradient ? La démonstration du Théorème 3.1.14 fournit une justification (ça marche) mais on peut donner une meilleure motivation de l'origine de (3.11). On définit sur $\mathbb{R}^+$ la fonction $j(\mu) = J\big(u^n - \mu\, J'(u^n)\big)$. Un calcul facile donne $j'(0) = -\|J'(u^n)\|^2$ et la condition d'Armijo est équivalente à

$$j(\mu) \leq j(0) + \frac{\mu}{2}j'(0).$$

Autrement dit, puisque $j(\mu) = j(0) + \mu j'(0) + o(\mu)$, on demande au pas μ de suivre une pente plus faible (en valeur absolue) que celle de la tangente mais de vérifier une inégalité sans la présence du reste $o(\mu)$. Par ailleurs, dans l'algorithme (3.12) μ est choisi comme le plus grand possible dans une suite discrète de pas, ce qui veut dire qu'une inégalité inverse a lieu pour le pas discret juste supérieur. De ce point de vue, la règle d'Armijo dans l'algorithme (3.12) est parfois appelée règle d'Armijo-Goldstein qui, pour des paramètres $0 < p_- < p_+ < 1$ fixés, demande que le pas $\mu > 0$ vérifie

$$j(0) + p_+\mu j'(0) \leq j(\mu) \leq j(0) + p_-\mu j'(0).$$

Pour finir sur la règle d'Armijo, notons qu'il est possible de changer la suite discrète de pas $(\frac{\mu}{2^i})_{i\geq 0}$ dans l'algorithme (3.12) par une autre suite $(\frac{\mu}{c^i})_{i\geq 0}$ avec $c > 1$. •

Remarque 3.1.17 Il existe d'autres règles du même type, comme celle de Wolfe (voir [7], [26]), afin d'adapter le pas de descente. Elles peuvent varier selon le type de problème et avoir un caractère assez heuristique. Mais l'essentiel est d'éviter les biais dans le choix du pas initial causés par l'absence de connaissance des paramètres α et L. •

3.2 Généralisations et autres algorithmes de type gradient

La section précédente a donné les bases de ce qu'il faut savoir sur les algorithmes de gradient pour l'optimisation sans contraintes. Cette section, qui peut être omise en première lecture, va généraliser ce que nous venons de voir dans deux directions. D'une part, nous allons voir comment les algorithmes de gradient doivent être modifiés et comment ils convergent lorsque la fonction coût à minimiser est moins régulière que ce que nous avons supposé jusqu'ici. D'autre part, nous présentons d'autres algorithmes de descente, qu'on dit du premier ordre car à chaque itération ils n'utilisent que l'information du gradient et pas celle de dérivées d'ordre supérieur (contrairement à la méthode de Newton que nous verrons plus tard). Typiquement ces nouveaux algorithmes diffèrent des deux 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 multi-pas). Nous en donnons quelques exemples simples pour illustrer la richesse de l'algorithmique en optimisation, sachant que certains algorithmes sont spécialisés pour des classe de fonctions coûts bien particulières.

3.2.1 Vitesse de convergence

Commençons par dire quelques mots de la vitesse de convergence de l'algorithme du gradient à pas fixe. En effet, cette question est intéressante en elle-même mais, en plus, elle motive la conception d'autres algorithmes pour l'améliorer. Comme nous l'avions dit à la Remarque 3.1.8 la vitesse de convergence du Théorème 3.1.6 n'est pas optimale. Nous donnons donc une nouvelle démonstration du Théorème 3.1.6, mais en nous plaçant en dimension finie et en supposant la fonction J de classe C^2 pour simplifier l'analyse, ce qui nous permet d'obtenir la vitesse de convergence optimale pour l'algorithme du gradient à pas fixe.

Proposition 3.2.1 *On ajoute aux hypothèses du Théorème 3.1.6 le fait que $V = \mathbb{R}^N$ et que la fonction J est de classe C^2. Alors l'algorithme de gradient à pas fixe converge pour $0 < \mu < 2/L$, c'est-à-dire que quel que soit u^0, la suite (u^n) définie par (3.8) converge vers la solution u de (3.1), et la vitesse de convergence optimale s'obtient pour le pas $\mu_{\text{opt}} = 2/(L+\alpha)$*

$$\|u^n - u\| \leq \gamma_{\text{opt}}^n \|u^0 - u\| \quad \textit{avec} \quad \gamma_{\text{opt}} = \frac{1-\alpha/L}{1+\alpha/L} < 1.$$

Démonstration. On écrit à nouveau

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

et on utilise un développement de Taylor avec reste exact

$$u^{n+1} - u = B^n(u^n - u) \quad \text{avec} \quad B^n = \left(\mathrm{Id} - \mu \int_0^1 J''(u + t(u^n - u))dt\right).$$

L'hypothèse (3.9) sur J' Lipschitzien implique $J'' \leq L\,\mathrm{Id}$, tandis que la forte convexité de J implique $J'' \geq \alpha\,\mathrm{Id}$. Par conséquent, la matrice symétrique B^n vérifie

$$(1 - \mu L)\,\mathrm{Id} \leq B^n \leq (1 - \mu\alpha)\,\mathrm{Id}.$$

Soit $\rho(B^n)$ le rayon spectral de B^n, c'est-à-dire le maximum des valeurs absolues de ses valeurs propres. On a $\rho(B^n) \leq \max(|1 - \mu L|, |1 - \mu\alpha|)$ dont on déduit $\rho(B^n) \leq \gamma < 1$ avec γ indépendant de n si $0 < \mu < 2/L$. Autrement dit, $\|u^n - u\| \leq \gamma^n \|u^0 - u\|$ et l'algorithme du gradient à pas fixe converge si $0 < \mu < 2/L$. Si maintenant on choisit μ tel que $(1 - \mu L) \leq 0 \leq (1 - \mu\alpha)$, alors

$$\rho(B^n) \leq \max(\mu L - 1, 1 - \mu\alpha)$$

et dans cette majoration la valeur optimale du pas est $\mu_{\mathrm{opt}} = 2/(L + \alpha)$ qui conduit à $\gamma_{\mathrm{opt}} = \frac{L-\alpha}{L+\alpha}$. □

Remarque 3.2.2 Comme l'a montré l'Exercice 3.1.5 sur une fonction quadratique, on ne peut pas espérer améliorer la vitesse de convergence de l'algorithme du gradient à pas fixe obtenue à la Proposition 3.2.1, qui est donc bien optimale. De même, l'Exercice 3.1.5 montre que l'on ne peut pas élargir l'ensemble des pas, $0 < \mu < 2/L$, pour lesquels l'algorithme du gradient à pas fixe converge. •

Remarque 3.2.3 Si on cherche à calculer la solution u du problème de minimisation (3.1) avec une précision relative ε, il faut faire (au plus) n itérations avec

$$\frac{\|u^n - u\|}{\|u^0 - u\|} \leq \gamma_{\mathrm{opt}}^n \leq \varepsilon,$$

c'est-à-dire $n \geq \log\varepsilon / \log\gamma_{\mathrm{opt}}$. Or, dans la formule $\gamma_{\mathrm{opt}} = \frac{1-\alpha/L}{1+\alpha/L}$ le rapport α/L est souvent très petit, ce qui conduit à l'approximation $\log\gamma_{\mathrm{opt}} \approx -2\alpha/L$. Par conséquent, une estimation du nombre maximal d'itérations nécessaire pour atteindre la précision ε est

$$n \approx |\log\varepsilon| \frac{L}{2\alpha}.$$

Le nombre $|\log\varepsilon|$ n'est pas très grand en pratique. Par contre, le rapport L/α, appelé conditionnement de la fonction J, peut être très grand. C'est pour réduire ce coefficient, et donc accélérer la convergence, que l'on va étudier divers algorithmes dans la suite (Nesterov, boule pesante, gradient conjugué) qui tous conduisent à remplacer le rapport L/α par $\sqrt{L/\alpha}$, ce qui constitue une amélioration très significative en pratique. •

Une propriété intéressante des fonctions J, α-convexes à dérivée L-Lipschitzienne, est que la décroissance de la norme de la dérivée $J'(u^n)$ est équivalente à la décroissance de la norme de l'erreur $(u^n - u)$ où u est l'unique point de minimum de J et u^n est la suite des itérées produite par n'importe quel algorithme d'optimisation. Autrement dit, il suffit d'étudier la convergence vers zéro de la dérivée $J'(u^n)$ pour avoir une idée de la convergence vers zéro de l'erreur $(u^n - u)$, sans même connaitre le point de minimum. C'est une conséquence du résultat suivant.

Lemme 3.2.4 *Soit J une fonction différentiable α-convexe à dérivée L-Lipschitzienne de V dans $\mathbb{R}$. Soit $u \in V$ son unique point de minimum. Alors, pour tout $v \in V$,*

$$\alpha\|v - u\| \leq \|J'(v)\| \leq L\|v - u\|.$$

Démonstration. Notons que $J'(u) = 0$. La première inégalité est alors une conséquence de (2.24) et de l'inégalité de Cauchy-Schwarz, tandis que la deuxième inégalité est simplement l'application de la Définition 2.4.9 sur la L-Lipschitzianité de J'. □

Si la fonction J que l'on minimise n'est pas fortement (ou α-) convexe, alors la vitesse de convergence est bien moindre, devenant algébrique au lieu de géométrique.

Proposition 3.2.5 *Soit une fonction $J(u)$ définie sur $V = \mathbb{R}^N$, strictement convexe, de classe C^2, infinie à l'infini et telle que J' est Lipschitzien, c'est-à-dire vérifie (3.9) pour une constante $L > 0$. Alors l'algorithme de gradient à pas fixe converge pour $0 < \mu < 2/L$, c'est-à-dire que quel que soit u^0, la suite (u^n) définie par (3.8) converge vers la solution u de (3.1), et on a*

$$0 \leq J(u^n) - J(u) \leq \frac{\kappa}{n+1} \qquad \text{avec } \kappa = \frac{\|u^0 - u\|^2}{\mu(1 - L\mu/2)}.$$

Remarque 3.2.6 Noter que la Proposition 3.2.5 ne dit rien de la vitesse de convergence de $\|u^n - u\|$ et que la vitesse de convergence est ici définie comme la vitesse de convergence de $J(u^n)$ vers $J(u)$. Cette dernière est **algébrique** (comme $1/n$), ce qui est beaucoup moins rapide que la convergence géométrique du cas fortement convexe (comme γ^{2n} avec $0 < \gamma < 1$, d'après l'Exercice 3.1.4). On ne peut pas avoir une estimation de $\|u^n - u\|$, indépendante de J, car les fonctions convexes, mais pas fortement convexes, peuvent être très plates près de leur minimum et la convergence de u^n vers u peut être aussi lente que l'on veut selon la forme de J. •

Démonstration. On a déjà démontré à la Proposition 3.1.9 que u^n converge vers l'unique point de minimum u de J. On calcule la norme au carré de la relation

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

où l'on a utilisé la propriété $J'(u) = 0$. On obtient

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

et l'on majore le dernier terme à l'aide du Lemme 3.2.7 ci-dessous. Comme $J'(u) = 0$, on en déduit

$$\|u^{n+1} - u\|^2 \leq \|u^n - u\|^2 - 2\mu(1 - L\mu/2)\|J'(u^n)\|^2,$$

et donc que la suite $\|u^n - u\|$ est décroissante pour $0 < \mu < 2/L$. Par ailleurs, la convexité de J en u^n donne

$$J(u^n) + \langle J'(u^n), u - u^n\rangle \leq J(u),$$

et par Cauchy-Schwarz ainsi que par la décroissance de $\|u^n - u\|$

$$0 \leq J(u^n) - J(u) \leq \|J'(u^n)\| \, \|u^n - u\| \leq \|J'(u^n)\| \, \|u^0 - u\|. \tag{3.14}$$

Or, lors de la preuve la Proposition 3.1.9 on a montré que

$$J(u^{n+1}) \leq J(u^n) - \mu(1 - L\mu/2)\|J'(u^n)\|^2,$$

c'est-à-dire, en posant $\Delta^n = J(u^n) - J(u) > 0$, que

$$\mu(1 - L\mu/2)\|J'(u^n)\|^2 \leq \Delta^n - \Delta^{n+1} \tag{3.15}$$

En combinant (3.14) et (3.15) on obtient

$$\Delta^{n+1} \leq \Delta^n - \kappa^{-1}(\Delta^n)^2 \quad \text{ avec } \kappa = \frac{\|u^0 - u\|^2}{\mu(1 - L\mu/2)}. \tag{3.16}$$

Montrons alors par récurrence sur n que $\Delta^n \leq \frac{\kappa}{n+1}$, ce qui est le résultat recherché. Pour $n = 0$, on reprend (3.14) et on majore

$$\|J'(u^0)\| = \|J'(u^0) - J'(u)\| \leq L\|u^0 - u\| \leq \frac{\|u^0 - u\|}{\mu(1 - L\mu/2)}$$

car $0 < \mu < 2/L$, donc $\Delta^0 \leq \kappa$. De (3.16) on a immédiatement $\Delta^{n+1} \leq \Delta^n \leq \Delta^0 \leq \kappa$. Supposons qu'à l'ordre $(n-1)$ on ait bien $n\Delta^{n-1} \leq \kappa$. On calcule alors

$$(n+1)\Delta^n \leq (n+1)\Delta^{n-1} - (n+1)\kappa^{-1}(\Delta^{n-1})^2 \leq n\Delta^{n-1} + \Delta^{n-1}\left(1 - (n+1)\kappa^{-1}\Delta^{n-1}\right)$$

$$\leq \kappa + \kappa^{-1}\Delta^{n-1}\left(\kappa - (n+1)\Delta^n\right)$$

puisque $\Delta^n \leq \Delta^{n-1}$. On regroupe les termes pour en déduire

$$0 \leq \left(\kappa - (n+1)\Delta^n\right)\left(1 + \kappa^{-1}\Delta^{n-1}\right)$$

qui prouve que $\kappa - (n+1)\Delta^n \geq 0$ comme voulu. □

Lemme 3.2.7 *Soit une fonction $J(u)$ définie sur $V = \mathbb{R}^N$, strictement convexe, de classe C^2 et telle que J' est L-Lipschitzien au sens de la Définition 2.4.9 pour une constante $L > 0$. Alors*

$$\|J'(v) - J'(w)\|^2 \leq L\,\langle J'(v) - J'(w), v - w\rangle \quad \textit{pour tout } v, w \in V.$$

Démonstration. On écrit la formule de Taylor avec reste exact

$$J'(v) = J'(w) + \int_0^1 J''(w + t(v-w))(v-w)\,dt.$$

Comme J est strictement convexe et J' L-Lipschitzien, la matrice $N \times N$ des dérivées secondes est inversible et vérifie $0 < J''(u) \leq L\,\mathrm{Id}$ pour tout $u \in \mathbb{R}^N$. Par conséquent, il en est de même de la matrice

$$A = \int_0^1 J''(w + t(v-w))\,dt$$

qui vérifie $A^{-1} \geq 1/L\,\mathrm{Id}$. On déduit alors de la formule de Taylor ci-dessus

$$A^{-1}\left(J'(v) - J'(w)\right) = v - w,$$

donc, pour tout $v, w \in \mathbb{R}^N$,

$$\frac{1}{L}\|J'(v) - J'(w)\|^2 \leq \langle A^{-1}\left(J'(v) - J'(w)\right), J'(v) - J'(w)\rangle = \langle J'(v) - J'(w), v - w\rangle.$$

En fait, le résultat reste vrai pour une fonction $J(u)$ convexe mais pas strictement convexe : il suffit de lui ajouter $\epsilon\|u\|^2$ avec $\epsilon > 0$ puis de passer à la limite $\epsilon \to 0$ dans le résultat. □

Remarque 3.2.8 Un aspect essentiel des estimations de vitesse de convergence que nous venons de voir est qu'elles sont toutes indépendantes de la dimension de l'espace V dans lequel on minimise. C'est particulièrement évident lorsque les preuves sont faites pour un espace de Hilbert quelconque, possiblement de dimension infinie. C'est une information très importante lorsqu'on travaille sur des problèmes en très grande dimension (voire infinie). Par contre, en général le coût de calcul d'une itération dépend de la dimension. Penser, par exemple, à la minimisation d'une fonction quadratique comme dans l'Exercice 3.1.5, où le coût d'une itération de la méthode du gradient à pas fixe est dominée par le calcul du gradient qui correspond à un produit matrice vecteur et est donc de l'ordre de N^2, avec N la dimension. Néanmoins, il faut se méfier de certains problèmes où les constantes α et L qui apparaissent dans la vitesse de convergence dépendent, de manière plus ou moins cachée, de la dimension (par exemple, pour une fonction quadratique où la matrice est issue de la discrétisation d'une équation différentielle). •

3.2.2 Algorithme de Nesterov

Nesterov [25] a proposé un algorithme qui améliore la vitesse de convergence de la méthode du gradient à pas fixe pour les fonctions convexes. Pour cette raison l'algorithme de Nesterov est parfois appelé **algorithme du gradient accéléré**. On l'initialise avec une constante $a_0 = 1$ et un vecteur $v^0 = u^{-1} \in V$. Ensuite, pour $n \geq 0$,

$$\begin{cases} u^n = v^n - \mu\, J'(v^n) \\ a_{n+1} = \dfrac{1}{2}(1 + \sqrt{4a_n^2 + 1}) \\ v^{n+1} = u^n + \dfrac{a_n - 1}{a_{n+1}}(u^n - u^{n-1}) \end{cases} \tag{3.17}$$

avec un pas fixe $\mu > 0$. L'idée de cet algorithme est d'extrapoler un nouveau vecteur v^{n+1} au delà du segment (u^{n-1}, u^n) où l'on calculera le gradient qui donnera la nouvelle itérée u^{n+1} (on parle aussi de sur-relaxation). La formule de récurrence pour le coefficient a_n vient d'une majoration précise dans la démonstration de convergence. On vérifie facilement que, pour n grand, $a_n \approx n/2$ et le coefficient d'extrapolation vaut $(a_n - 1)/a_{n+1} \approx 1$.

Proposition 3.2.9 *Soit J une fonction convexe différentiable, telle que J' est L-Lipschitzienne, c'est-à-dire vérifie (3.9). Soit un pas de descente μ tel que $0 < \mu \leq 1/L$. On suppose qu'il existe un point de minimum u de J sur V. Alors la suite u^n de l'agorithme de Nesterov, définie par* (3.17)*, vérifie*

$$0 \leq J(u^n) - J(u) \leq \frac{\kappa}{(n+2)^2} \qquad \text{avec } \kappa = 2\mu^{-1}\|u^0 - u\|^2 + 4(J(u^0) - J(u)).$$

Remarque 3.2.10 Avec quelques hypothèses supplémentaires on peut montrer que la suite u^n converge vers u (voir l'Exercice 3.2.1 ci-dessous). La vitesse de convergence de la Proposition 3.2.9 améliore sensiblement ($1/n^2$ au lieu de $1/n$) celle obtenue dans la Proposition 3.2.5. Comme on le verra à la Remarque 3.2.11 cette vitesse est optimale pour les fonctions convexes. Néanmoins, pour des fonctions fortement convexes, quitte à changer la définition précise du coefficient d'extrapolation a_n, on peut encore améliorer cette vitesse de convergence et retrouver une convergence géométrique (voir la Proposition 3.2.12). •

Démonstration. Soit $p^n = (a_n - 1)(u^{n-1} - u^n)$ de sorte que $v^{n+1} = u^n - p^n/a_{n+1}$. L'astuce de Nesterov est d'étudier la convergence de la suite $u^n - p^n$ plutôt que celle de u^n ou v^n. On calcule

$$u^{n+1} - p^{n+1} = u^{n+1} - (a_{n+1} - 1)(u^n - u^{n+1}) = a_{n+1}u^{n+1} - (a_{n+1} - 1)u^n$$

$$= a_{n+1}(v^{n+1} - \mu J'(v^{n+1})) - (a_{n+1} - 1)u^n = u^n - p^n - a_{n+1}\mu J'(v^{n+1}).$$

On en déduit

$$\begin{aligned} \|u^{n+1} - p^{n+1} - u\|^2 &= \|u^n - p^n - u\|^2 + a_{n+1}^2\mu^2\|J'(v^{n+1})\|^2 \\ &\quad +2a_{n+1}\mu\langle J'(v^{n+1}), u - v^{n+1}\rangle \\ &\quad +2(a_{n+1} - 1)\mu\langle J'(v^{n+1}), p^n\rangle, \end{aligned} \tag{3.18}$$

en ayant utilisé $a_{n+1}(v^{n+1} + p^n - u^n) = (a_{n+1} - 1)p^n$ dans la dernière ligne. On majore la deuxième ligne de (3.18) par convexité de J

$$\langle J'(v^{n+1}), u - v^{n+1}\rangle \leq J(u) - J(v^{n+1})$$

et la troisième ligne de même car

$$\langle J'(v^{n+1}), p^n\rangle = a_{n+1}\langle J'(v^{n+1}), u^n - v^{n+1}\rangle \leq a_{n+1}(J(u^n) - J(v^{n+1})).$$

Ainsi (3.18) devient

$$\begin{aligned}\|u^{n+1}-p^{n+1}-u\|^2 &\leq \|u^n-p^n-u\|^2+a_{n+1}^2\mu^2\|J'(v^{n+1})\|^2\\ &\quad +2a_{n+1}\mu(J(u)-J(v^{n+1}))\\ &\quad +2(a_{n+1}-1)\mu a_{n+1}(J(u^n)-J(v^{n+1})).\end{aligned}\tag{3.19}$$

On majore les deux dernières lignes de (3.19) en écrivant

$$\begin{aligned}J(u)-J(v^{n+1}) &= J(u)-J(u^{n+1})+J(u^{n+1})-J(v^{n+1})\\ J(u^n)-J(v^{n+1}) &= J(u^n)-J(u^{n+1})+J(u^{n+1})-J(v^{n+1})\end{aligned}$$

puis en appliquant le Lemme 3.1.10 avec $v=v^{n+1}$ et $v^*=u^{n+1}$

$$J(u^{n+1})-J(v^{n+1})\leq -\mu(1-L\mu/2)\|J'(v^{n+1})\|^2.\tag{3.20}$$

Par conséquent, (3.19) devient

$$\begin{aligned}\|u^{n+1}-p^{n+1}-u\|^2 &\leq \|u^n-p^n-u\|^2+a_{n+1}^2\mu^2\|J'(v^{n+1})\|^2(L\mu-1)\\ &\quad +2\mu a_{n+1}(J(u)-J(u^{n+1}))\\ &\quad +2\mu(a_{n+1}-1)a_{n+1}(J(u^n)-J(u^{n+1})).\end{aligned}\tag{3.21}$$

On remarque que la définition de la suite a_n implique que $(a_{n+1}-1)a_{n+1}=a_n^2$. En supposant que $0<\mu\leq 1/L$ et en introduisant $J(u)$ dans la dernière ligne de (3.21), on obtient

$$\|u^{n+1}-p^{n+1}-u\|^2+2\mu a_{n+1}^2(J(u^{n+1})-J(u))\leq \|u^n-p^n-u\|^2+2\mu a_n^2(J(u^n)-J(u))$$

donc, comme $a_0=1$ et $p^0=0$,

$$2\mu a_n^2(J(u^n)-J(u))\leq \|u^0-u\|^2+2\mu(J(u^0)-J(u)).$$

Il est facile de vérifier que $a_n\geq 1+n/2$ ce qui donne la majoration attendue. □

Exercice 3.2.1 On se place sous les hypothèses de la Proposition 3.2.9 et dans un espace de dimension finie $V=\mathbb{R}^N$. On suppose de plus que la fonction J est strictement convexe et infinie à l'infini. On considère la suite u^n donnée par l'algorithme (3.17) de Nesterov.

1. Montrer que la suite u^n est bornée.
2. Montrer que toute sous-suite convergente de u^n converge vers une limite où J' s'annule.
3. En déduire que toute la suite u^n converge vers l'unique point de minimum de J.

Exercice 3.2.2 Soit la suite réelle a_n définie par l'algorithme (3.17) de Nesterov avec l'initialisation $a_0=1$.

1. Vérifier que $n/2+1 \leq a_n \leq n/2+\sqrt{n}/2$.
2. En déduire que $a_n = \frac{n}{2}(1+o(1))$.

Exercice 3.2.3 Une variante de l'algorithme (3.17) de Nesterov consiste à remplacer la suite scalaire a_n, donnée par une relation de récurrence, par la borne inférieure $n/2+1$. Pour une initialisation $v^0 = u^{-1} \in V$, et pour $n \geq 0$, on obtient

$$\begin{cases} u^n = v^n - \mu\, J'(v^n) \\ v^{n+1} = u^n + \dfrac{n}{n+3}(u^n - u^{n-1}) \end{cases} \tag{3.22}$$

Montrer que la Proposition 3.2.9 reste valide pour cette variante. (On reprendra la démonstration en identifiant le seul passage où la formule explicite pour a_n est utilisée.)

Remarque 3.2.11 La vitesse de convergence de la Proposition 3.2.9 est optimale pour des fonctions convexes, mais pas fortement convexes, et à dérivée Lipschtzienne. Précisément, pour tout entier n (plus petit que la dimension de l'espace V), on peut construire une fonction J et un point initial u^0 tel qu'à l'itération n la différence $J(u^n) - J(u)$ soit précisément de l'ordre de la borne supérieure donnée par la Proposition 3.2.9. En effet, pour un entier $p \leq N$, considérons l'exemple suivant d'une fonction J définie sur $\mathbb{R}^N$

$$J(v) = \frac{L}{8}\left((v_1-1)^2 + \sum_{i=2}^{p}(v_i - v_{i-1})^2 + \sum_{i=p+1}^{N} v_i^2\right)$$

dont l'unique point de minimum est $u = (1,...,1,0,...,0)$, qui a ses p premières composantes égales à 1 et les suivantes égales à 0, avec $J(u) = 0$. On initialise un algorithme du gradient avec $u^0 = 0$. Tous les algorithmes de gradient que nous étudions dans cette section (pas optimal ou fixe, Nesterov, boule pesante ou gradient conjugué) vérifient la propriété suivante : à chaque itération n, la solution approchée u^n appartient à l'espace affine $u^0 + E_n$ où E_n est le sous-espace vectoriel engendré par $\{J'(u^0), J'(u^1), ..., J'(u^{n-1})\}$. Par conséquent, une borne inférieure est

$$J(u^n) \geq \min_{v \in u^0 + E_n} J(v).$$

Or, dans l'exemple ci-dessus, il est facile de voir, par récurrence sur n, que $J'(u^{n-1})$ appartient à l'espace engendré par les n premiers vecteurs de la base canonique, $\{e_1, ..., e_n\}$, et par conséquent que $E_n \subset \mathrm{Vect}\{e_1, ..., e_n\}$ donc

$$J(u^n) \geq \min_{v \in u^0 + E_n} J(v) \geq \min_{v \in u^0 + \mathrm{Vect}\{e_1,...,e_n\}} J(v).$$

Autrement dit, sur cet exemple (très défavorable !) l'algorithme de gradient (quelle que soit sa version) ne peut propager l'information que la solution u a ses p premières composantes égales à 1, à partir d'une initialisation nulle $u^0 = 0$, qu'à la "vitesse" d'une composante par itération. Un calcul facile permet de calculer le minimum sur le sous-espace affine $u^0 + \mathrm{Vect}\{e_1, ..., e_n\}$: ce point de minimum, noté w^n, a ses

dernières $N-n$ composantes nulles tandis que les n premières sont $w_i^n = 1 - \frac{i}{n+1}$ pour $1 \leq i \leq n$. Par conséquent $(w_1^n - 1)^2 = (w_i^n - w_{i-1}^n)^2 = (w_n^n)^2 = \frac{1}{(n+1)^2}$ pour $2 \leq i \leq n$, ce qui donne, pour $0 \leq n \leq p$,

$$\min_{v \in u^0 + \mathrm{Vect}\{e_1,\dots,e_n\}} J(v) = J(w^n) = \frac{L}{8(n+1)}.$$

Comme $J(u) = 0$ et $\|u^0 - u\|^2 = p$, on a donc

$$J(u^n) - J(u) \geq \frac{L}{8(n+1)} = \frac{L}{8p(n+1)}\|u^0 - u\|^2,$$

ce qui, à l'itération $n = p-1$, donne

$$J(u^n) - J(u) \geq \frac{L}{8(n+1)^2}\|u^0 - u\|^2.$$

Or la borne supérieure de la Proposition 3.2.9 est

$$J(u^n) - J(u) \leq \frac{\kappa}{(n+2)^2} \quad \text{avec } \kappa = 2\mu^{-1}\|u^0 - u\|^2 + 4(J(u^0) - J(u)),$$

on en déduit qu'à une constante multiplicative près (indépendante de n) on ne peut pas améliorer cette borne supérieure. •

L'algorithme de Nesterov [25] peut être amélioré pour des fonctions fortement convexes en changeant les formules d'extrapolation, du moins si l'on connait les valeurs des constantes α (convexité) et L (Lipschitziannité), ce qui est malheureusement rarement le cas en pratique. En posant $q = \alpha/L \in]0,1[$, l'algorithme de Nesterov devient, pour $n \geq 0$,

$$\left\{ \begin{array}{l} a_{n+1} = \dfrac{1 - qa_n^2 + \sqrt{(1-qa_n^2)^2 + 4a_n^2}}{2} \\ v^n = u^n + \dfrac{(a_n - 1)(1 - a_{n+1}\alpha\mu)}{a_{n+1}(1-\alpha\mu)}(u^n - u^{n-1}) \\ u^{n+1} = v^n - \mu\, J'(v^n), \end{array} \right. \tag{3.23}$$

que l'on initialise avec $u^0 = u^{-1} \in V$ et $1 \leq a_0 \leq 1/\sqrt{q}$.

Proposition 3.2.12 *On suppose que J est α-convexe différentiable, de dérivée L-Lipschitzienne et que le pas de descente vérifie $0 < \mu \leq 1/L$. Alors la suite u^n de l'agorithme de Nesterov, définie par* (3.23)*, converge vers la solution u de (3.1), et, pour $\mu = 1/L$, la vitesse de convergence est*

$$\|u^n - u\| \leq \frac{2}{\alpha}\gamma_{\mathrm{Nest}}^n \|u^0 - u\| \quad \textit{avec} \quad \gamma_{\mathrm{Nest}} = \left(1 - \sqrt{\frac{\alpha}{L}}\right).$$

Remarque 3.2.13 La vitesse de convergence de l'algorithme de Nesterov est asymptotiquement (c'est-à-dire quand le rapport α/L est petit) meilleure que celle du

gradient à pas optimal. Nous verrons d'autres algorithmes (boule pesante, gradient conjugué) qui ont le même ordre de grandeur asymptotique en $q = \alpha/L$ de leur vitesse de convergence. Lorsque n tend vers l'infini, la suite a_n tend vers $1/\sqrt{q}$ et, pour $\mu = 1/L$, le coefficient d'extrapolation qui donne v^n dans (3.23) converge vers $(1-\sqrt{q})/(1+\sqrt{q})$. •

Démonstration. On applique l'inégalité (3.25) du Lemme 3.2.14 à

$$w = \frac{t-1}{t}u^n + \frac{1}{t}u, \quad v = v^n, \quad v^* = u^{n+1} = v^n - \mu\, J'(v^n),$$

avec $t \geq 1$. En utilisant aussi la forte convexité de J

$$J(w) \leq \frac{t-1}{t}J(u^n) + \frac{1}{t}J(u) - \frac{\alpha}{2}\frac{(t-1)}{t^2}\|u^n - u\|^2,$$

en multipliant (3.25) par t^2 il vient

$$t(t-1)J(u^n) + tJ(u) - \frac{\alpha}{2}(t-1)\|u^n - u\|^2 + \frac{1-\alpha\mu}{2\mu}\|(t-1)u^n + u - tv^n\|^2$$

$$\geq t^2 J(u^{n+1}) + \frac{1}{2\mu}\|tu^{n+1} - (t-1)u^n - u\|^2$$

que l'on réécrit

$$t(t-1)(J(u^n) - J(u)) - \frac{\alpha}{2}(t-1)\|u^n - u\|^2 + \frac{1-\alpha\mu}{2\mu}\|u - u^n + t(u^n - v^n)\|^2$$

$$\geq t^2(J(u^{n+1}) - J(u)) + \frac{1}{2\mu}\|u^{n+1} - u - (t-1)(u^n - u^{n+1})\|^2.$$

Or

$$-\frac{\alpha}{2}(t-1)\|u^n - u\|^2 + \frac{1-\alpha\mu}{2\mu}\|u - u^n + t(u^n - v^n)\|$$

$$= \frac{1-t\alpha\mu}{2\mu}\|u^n - u\|^2 + \frac{1-\alpha\mu}{2\mu}t^2\|u^n - v^n\|^2 + \frac{1-\alpha\mu}{\mu}t\langle u - u^n, u^n - v^n\rangle$$

$$= \frac{1-t\alpha\mu}{2\mu}\|u - u^n + \frac{1-\alpha\mu}{1-t\alpha\mu}t(u^n - v^n)\|^2 + \frac{t^2(1-t)\mu(1-\alpha\mu)}{2(1-t\alpha\mu)}\|u^n - v^n\|^2.$$

Par conséquent

$$\begin{aligned} &t(t-1)(J(u^n) - J(u)) + \frac{1-t\alpha\mu}{2\mu}\|u - u^n + \frac{1-\alpha\mu}{1-t\alpha\mu}t(u^n - v^n)\|^2 \\ &\geq t^2(J(u^{n+1}) - J(u)) + \frac{1}{2\mu}\|u^{n+1} - u - (t-1)(u^n - u^{n+1})\|^2 \\ &+\frac{t^2(t-1)\mu(1-\alpha\mu)}{2(1-t\alpha\mu)}\|u^n - v^n\|^2. \end{aligned} \tag{3.24}$$

On choisit alors $t = a_{n+1}$, où a_n est la suite définie par (3.23). Grâce aux propriétés du Lemme 3.2.15, on a $a_{n+1} \geq 1$ ainsi que

$$a_{n+1}(a_{n+1} - 1) \leq a_n^2(1 - qa_{n+1}) \leq a_n^2(1 - a_{n+1}\alpha\mu),$$

puisque $\alpha\mu \leq q$, et le dernier terme de (3.24) est positif. Prenant en compte la définition de v^n

$$v^n = u^n + \frac{(a_n - 1)(1 - a_{n+1}\alpha\mu)}{a_{n+1}(1 - \alpha\mu)}(u^n - u^{n-1})$$

on déduit alors de (3.24)

$$(1 - a_{n+1}\alpha\mu)\Delta_n \geq \Delta_{n+1},$$

avec

$$\Delta_n = a_n^2 \left(J(u^n) - J(u)\right) + \frac{1}{2\mu}\|u - u^n - (a_n - 1)(u^n - u^{n-1})\|^2,$$

c'est-à-dire que

$$\Delta_n \leq \Delta_0 \prod_{k=1}^{n}(1 - a_k\alpha\mu).$$

Comme $0 \leq 1 - a_k\alpha\mu \leq 1$ et que $\lim_{n\to+\infty} a_n = 1/\sqrt{q}$, on en déduit que Δ_n tend vers 0 et l'algorithme converge. De plus, si on choisit le pas $\mu = 1/L$, comme $1 - \frac{1}{a_n} = (1 - qa_n)\frac{a_{n-1}^2}{a_n^2}$, on a

$$\prod_{k=1}^{n}(1 - a_k\alpha\mu) = \prod_{k=1}^{n}(1 - qa_k) = \frac{a_n^2}{a_0^2}\prod_{k=1}^{n}(1 - \frac{1}{a_k}) \leq a_n^2(1 - \sqrt{q})^n,$$

d'où l'on déduit en particulier $J(u^n) - J(u) \leq (1 - \sqrt{q})^n$. Alors la forte convexité de J et le fait que $J'(u) = 0$ donne

$$\frac{\alpha}{2}\|u^n - u\|^2 \leq J(u^n) - J(u) \leq (1 - \sqrt{q})^n,$$

ce qui conclut la démonstration. □

Lemme 3.2.14 *Soit J une fonction α-convexe sur V dont la dérivée J' est Lipschitzienne de constante L. Pour tout $v, w \in V$ et pour tout $0 < \mu \leq 1/L$, en notant $v^* = v - \mu J'(v)$, on a*

$$J(w) + \frac{1 - \alpha\mu}{2\mu}\|w - v\|^2 \geq J(v^*) + \frac{1}{2\mu}\|v^* - w\|^2. \tag{3.25}$$

Démonstration. Pour tout $v, w \in V$ la forte convexité de J s'écrit

$$J(w) \geq J(v) + \langle J'(v), w - v\rangle + \frac{\alpha}{2}\|w - v\|^2 \tag{3.26}$$

auquel on ajoute un terme $\|w - v\|^2/(2\mu)$ de part et d'autre pour obtenir

$$J(w) + \frac{1 - \alpha\mu}{2\mu}\|w - v\|^2 \geq J(v) + \langle J'(v), w - v\rangle + \frac{1}{2\mu}\|w - v\|^2. \tag{3.27}$$

On note $Q(w)$ la fonction convexe quadratique à droite de (3.27)

$$Q(w) = J(v) + \langle J'(v), w - v\rangle + \frac{1}{2\mu}\|w - v\|^2$$

qui admet $v^* = v - \mu J'(v)$ comme unique point de minimum. Puisque $Q(w)$ est quadratique et que $Q'(v^*) = 0$, on a

$$Q(w) = Q(v^*) + \frac{1}{2\mu}\|v^* - w\|^2. \tag{3.28}$$

En combinant (3.27) et (3.28) on obtient

$$J(w)+\frac{1-\alpha\mu}{2\mu}\|w-v\|^2 \geq J(v)+\langle J'(v), v^*-v\rangle+\frac{1}{2\mu}\|v^*-v\|^2+\frac{1}{2\mu}\|v^*-w\|^2. \tag{3.29}$$

Mais comme J' est L-Lipschitzien, en vertu du Lemme 2.4.10 on a

$$J(v^*) \leq J(v) + \langle J'(v), v^* - v\rangle + \frac{L}{2}\|v^* - v\|^2. \tag{3.30}$$

Donc (3.29) et (3.30) conduisent au résultat

$$\begin{aligned} J(w) + \frac{1-\alpha\mu}{2\mu}\|w - v\|^2 &\geq J(v^*) - \frac{L}{2}\|v^* - v\|^2 + \frac{1}{2\mu}\|v^* - v\|^2 + \frac{1}{2\mu}\|v^* - w\|^2 \\ &\geq J(v^*) + \frac{1}{2\mu}\|v^* - w\|^2 \end{aligned}$$

car $\mu \leq 1/L$. □

Lemme 3.2.15 *Soit $0 < q < 1$. On définit la suite*

$$a_{n+1} = \frac{1 - qa_n^2 + \sqrt{(1 - qa_n^2)^2 + 4a_n^2}}{2},$$

avec $a_0\sqrt{q} \leq 1$. Alors a_n est une suite croissante et majorée pas $1/\sqrt{q}$, qui converge vers $1/\sqrt{q}$. En particulier, si $a_0 \geq 1$, alors $a_n \geq 1$ pour tout n. Par ailleurs, elle vérifie $a_{n+1}(a_{n+1} - 1) \leq a_n^2(1 - qa_{n+1})$.

Remarque 3.2.16 Si $q = 0$, on retrouve la suite a_n définie par (3.17) qui a un comportement différent puisque $a_n \approx 1 + n/2$. Par ailleurs, si $a_0 = 1/\sqrt{q}$, alors la suite est constante $a_n = 1/\sqrt{q}$. •

Démonstration. Montrons par récurrence que $a_n\sqrt{q} \leq 1$. La propriété est vraie pour $n = 0$. Pour $n \geq 0$, le coefficient a_{n+1} est la racine positive du polynôme

$$a_{n+1}^2 - (1 - qa_n^2)a_{n+1} - a_n^2 = 0.$$

En multipliant par q on en déduit

$$qa_{n+1}^2 = qa_{n+1} + (1 - qa_{n+1})qa_n^2. \tag{3.31}$$

Si $qa_{n+1} \geq 1$, on en déduirait $qa_{n+1}^2 \leq qa_{n+1}$, c'est-à-dire $qa_{n+1} \leq q < 1$, ce qui est contradictoire avec l'hypothèse. Donc $qa_{n+1} < 1$ et (3.31) implique que qa_{n+1}^2 est une combinaison convexe de 1 et qa_n^2. Alors, par concavité de la racine carrée, on obtient que $\sqrt{q}a_{n+1} \leq 1$ si $\sqrt{q}a_n \leq 1$. Par ailleurs, on peut aussi réécrire (3.31) sous la forme

$$a_{n+1}^2 = a_n^2 + (1 - qa_n^2)a_{n+1} \geq a_n^2,$$

donc la suite a_n est croissante. Comme elle est majorée, elle converge et sa limite se déduit facilement de (3.31). Finalement,

$$a_{n+1}(a_{n+1} - 1) = a_n^2(\frac{1}{a_n} - qa_{n+1}) \leq a_n^2(1 - qa_{n+1})$$

car $a_n \geq a_0 \geq 1$. □

3.2.3 Algorithme de la boule pesante

Nous présentons ici un algorithme qui est intéressant en lui-même mais qui est surtout l'occasion de montrer une source d'inspiration dans la conception des algorithmes d'optimisation en les interprétant comme des schémas de discrétisation en temps de systèmes dynamiques. Par exemple, l'algorithme de gradient à pas fixe (3.8) peut se voir comme un schéma explicite pour l'équation différentielle

$$\dot{u}(t) = -J'(u(t)), \tag{3.32}$$

avec la donnée initiale $u(0) = u^0$ (la notation $\dot{u}$ signifie la dérivée par rapport à la variable de temps t). En effet, si on assimile le pas de descente $\mu > 0$ à un pas de temps, alors u^n peut être interprété comme une approximation au temps $t^n = n\mu$ de la solution $u(t)$ (voir [1] pour des détails sur cette notion). L'équation différentielle ordinaire (3.32) 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 est fortement convexe, on peut aussi montrer que la trajectoire $u(t)$ converge, quand t tend vers l'infini, vers l'unique point de minimum de J.

Expliquons maintenant la motivation et l'origine de l'algorithme de la boule pesante dans ce contexte [28]. Si la fonction J n'est pas convexe et présente des minima locaux, la solution $u(t)$ peut converger vers un tel minimum local, loin d'un minimum global. Pour améliorer cette situation défavorable, 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 $m > 0$ serait la masse de cette boule. Par analogie avec une discrétisation explicite en temps (qui nécessite 3 points en temps pour approcher la dérivée seconde), on obtiendrait

$$u^{n+1} = u^n - \mu\, J'(u^n) + \nu(u^n - u^{n-1}), \tag{3.33}$$

où $\mu > 0$ et $\nu > 0$ sont deux paramètres positifs. Bien sûr, il faut joindre à cette récurrence pour $n \geq 1$ deux initialisations $u^1, u^0 \in V$. Pour simplifier l'analyse on se place en dimension finie.

Théorème 3.2.17 *On suppose que J est convexe, de classe C^2 sur $V = \mathbb{R}^N$, tel que, pour $0 < \alpha \leq L$,*

$$\alpha \,\mathrm{Id} \leq J''(u) \leq L \,\mathrm{Id}. \tag{3.34}$$

Alors, si u^1, u^0 sont suffisamment proches de l'unique point de minimum u de J, pour $0 \leq \nu < 1$ et $0 < \mu < 2(1+\nu)/L$, l'algorithme de la boule pesante (3.33) *converge et, pour $\nu = (\frac{\sqrt{L}-\sqrt{\alpha}}{\sqrt{L}+\sqrt{\alpha}})^2$ et $\mu = 4/(\sqrt{L}+\sqrt{\alpha})^2$, il existe $C > 0$ tel que*

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

Remarque 3.2.18 L'hypothèse du Théorème 3.2.17, qui demande que l'initialisation u^1, u^0 soit proche de la solution u, rappelle l'hypothèse similaire pour la convergence de l'algorithme de Newton (voir la Proposition 3.3.1) et est malheureusement impossible à vérifier en pratique. Pire encore, si cette hypothèse n'est pas vérifiée, alors il existe des exemples où l'algorithme de la boule pesante ne converge pas... •

Démonstration. Tout d'abord, l'hypothèse (3.34) implique que la fonction J est α-convexe (voir l'Exercice 2.4.7) et que sa dérivée J' est L-Lipschitzienne (voir l'Exercice 2.4.8). En particulier, il existe un unique point de minimum u de J. On utilise alors un argument de stabilité locale, très similaire à celui qui permet d'analyser la convergence de l'algorithme de Newton. Puisque les deux premiers terms u^1, u^0 sont proches de u, on suppose (pour l'instant) que toutes les itérées u^n restent proches de la solution u. Puisque $J'(u) = 0$, un développement de Taylor donne

$$J'(u^n) = J''(u)(u^n - u) + o(u^n - u),$$

c'est-à-dire que

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

Soit z^n le vecteur $(u^n - u, u^{n-1} - u)^*$ et $H = J''(u)$. On a

$$z^{n+1} = Az^n + o(z^n) \quad \text{avec} \quad A = \begin{pmatrix} (1+\nu)\,\mathrm{Id} - \mu H & -\nu\,\mathrm{Id} \\ \mathrm{Id} & 0 \end{pmatrix}.$$

Pour étudier la convergence de la suite z^n on analyse les valeurs propres ρ de A qui vérifient, pour $(x, y) \neq 0$,

$$A \begin{pmatrix} x \\ y \end{pmatrix} = \rho \begin{pmatrix} x \\ y \end{pmatrix}. \tag{3.35}$$

En effet, si toutes les valeurs propres vérifient $|\rho| \leq \rho_{\max} < 1$, alors un lemme classique d'algèbre linéaire (voir la proposition 13.1.7 dans [1]) affirme que, pout tout $\delta > 0$ (aussi petit que l'on veut), il existe une norme matricielle (subordonnée à une

norme vectorielle) telle que $\|A\| \leq \rho_{\max} + \delta$. On déduit alors de $z^{n+1} = Az^n + o(z^n)$ et du fait que, pour n suffisamment grand, $\|o(z^n)\| \leq \delta\|z^n\|$ la majoration

$$\|z^{n+1}\| \leq (\rho_{\max} + 2\delta)\|z^n\|,$$

d'où l'on déduit la convergence de z^n vers 0 pour $\delta < (1 - \rho_{\max})/2$ (et aussi le fait que si u^n est proche de u à un certain rang, cela restera vrai pour les itérations suivantes).

Or ρ est valeur propre dans (3.35) si $Hx = \lambda x$ avec $\lambda = (1 + \nu - \rho - \nu/\rho)/\mu$ où λ est une valeur propre de la Hessienne H. Autrement dit, ρ est une racine du polynome

$$\rho^2 - (1 + \nu - \mu\lambda)\rho + \nu = 0,$$

dont le discriminant est $\Delta = (1+\nu-\mu\lambda)^2 - 4\nu$. On cherche des conditions suffisantes pour que les racines vérifient $|\rho| < 1$. On choisit $\nu \geq 0$ et comme ν est le produit des racines du polynome, il faut que $\nu < 1$. Si $\Delta \leq 0$, alors les deux racines sont complexes et de module égal à $\sqrt{\nu} < 1$. Si $\Delta > 0$, alors un calcul facile montre que les deux racines appartiennent à l'intervalle $(-1, +1)$ si et seulement si $0 < \mu < 2(1 + \nu)/\lambda$ pour toute valeur propre λ de la Hessienne H. A cause de l'hypothèse (3.34) les valeurs propres de H vérifient $\lambda \in [\alpha, L]$. Par conséquent, les conditions $0 < \nu < 1$ et $0 < \mu < 2(1 + \nu)/L$ assurent la convergence de l'algorithme. Pour trouver des valeurs optimales des paramètres, une analyse (un peu longue et pas détaillée ici) montre qu'il est bon de se placer dans le cas des racines complexes de module égal à $\sqrt{\nu}$, ce qui arrive lorsque $\Delta \leq 0$. On vérifie que $\Delta \leq 0$ est équivalent à

$$(1 - \sqrt{\nu})^2 \leq \mu\lambda \leq (1 + \sqrt{\nu})^2,$$

qui doit être vrai pour toute valeur propre $\lambda \in [\alpha, L]$ de H. C'est donc vrai en particulier si

$$\sqrt{\nu} \geq \sqrt{\alpha\mu} - 1, \quad \sqrt{\nu} \geq 1 - \sqrt{L\mu}.$$

On choisit $\sqrt{\mu} = 2/(\sqrt{\alpha}+\sqrt{L})$ pour que les deux bornes inférieures de $\sqrt{\nu}$ coincident et on prend la plus petite valeur qui en résulte, c'est-à-dire $\sqrt{\nu} = \frac{\sqrt{\alpha}-\sqrt{L}}{\sqrt{\alpha}+\sqrt{L}}$. □

Remarque 3.2.19 En écrivant la formule (3.33) comme une récurrence sur la différence $(u^{n+1} - u^n)$ on obtient la forme équivalente de l'algorithme de la boule pesante

$$u^{n+1} = u^n - \mu \sum_{k=1}^{n} \nu^{n-k} J'(u^k) + \nu^n(u^1 - u^0).$$

Autrement dit, si on néglige le terme $\nu^n(u^1 - u^0)$ (ce qui est loisible puisque $0 < \nu < 1$), l'algorithme prend comme direction de descente une pondération de l'ensemble des gradients précédents. Par conséquent le gradient est moyenné ce qui atténue les effets d'éventuelles oscillations. Cette idée de moyenner les gradients se retrouve dans d'autres algorithmes de nature stochastique, comme ADAM, utilisés en apprentissage machine (voir la Remarque 3.2.37). •

Remarque 3.2.20 La vitesse de convergence du Théorème 3.2.17 est bien meilleure que celle de l'algorithme du gradient à pas fixe (voir la Proposition 3.2.1) et un peu meilleure, mais asymptotiquement comparable (pour α/L petit) à celle de l'algorithme de Nesterov (voir la Proposition 3.2.12) qui s'écrit

$$\gamma_{\text{Nest}} = \left(1 - \sqrt{\frac{\alpha}{L}}\right) > \gamma_{\text{BP}} = \frac{1 - \sqrt{\alpha/L}}{1 + \sqrt{\alpha/L}}.$$

On voit donc, comme $0 < \alpha/L < 1$, que la méthode de la boule pesante avec ces paramètres optimaux converge plus vite que les algorithmes précédents. Malheureusement, les valeurs de α et L ne sont souvent pas connues en pratique et il est difficile d'ajuster les valeurs optimales des paramètres μ et ν. •

Exercice 3.2.4 Montrer que la variante (3.22) de l'algorithme de Nesterov correspond formellement à la discrétisation en temps de l'équation différentielle ordinaire suivante

$$\begin{cases} \ddot{u}(t) + \dfrac{3}{t}\dot{u}(t) = -J'(u(t)) \\ u(0) = u^0,\ \dot{u}(0) = 0. \end{cases} \tag{3.36}$$

Pour cela on supposera que les itérées u^n de (3.22) sont la discrétisation au temps $t^n = n\sqrt{\mu}$ (où $\mu > 0$ est le pas de descente) d'une fonction régulière $u(t)$, dont on montrera qu'elle est solution de (3.36) lorsque μ tends vers zéro.

3.2.4 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

$$\begin{cases} u^{n+1} = u^n - \mu^n\, p^n \\ p^{n+1} = J'(u^{n+1}) + \beta^n p^n \end{cases} \tag{3.37}$$

où $\mu^n \in \mathbb{R}$ est le pas optimal qui minimise $\mu \mapsto J(u^n - \mu\, p^n)$, 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. Pour simplifier la présentation, nous supposons à nouveau que l'espace de Hilbert est de dimension finie,

$V = \mathbb{R}^N$. Tout d'abord, la recherche linéaire pour trouver le pas μ^n est classique et reminiscente 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 non-linéaire quelconque, l'algorithme du gradient conjugué fonctionnera néanmoins, approximativement comme dans le cas quadratique, au voisinage d'un point de minimum non dégénéré (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 3.2.5), 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

$$\beta^n = -\frac{J'(u^{n+1}) \cdot Ap^n}{Ap^n \cdot p^n}, \tag{3.38}$$

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

$$Ap^n \cdot p^n = \|J'(u^n)\|^2/\mu^n. \tag{3.39}$$

Par ailleurs,

$$J'(u^{n+1}) \cdot Ap^n = J'(u^{n+1}) \cdot (J'(u^n) - J'(u^{n+1}))/\mu^n, \tag{3.40}$$

d'où l'on déduit la formule de Polak-Ribière en reportant (3.39) et (3.40) dans (3.38). 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 (3.37) se simplifient en

$$\text{pour } n \geq 0 \quad \left\{ \begin{array}{l} x^{n+1} = x^n - \mu^n p^n \\ r^{n+1} = r^n - \mu^n A p^n \\ p^{n+1} = r^{n+1} + \beta^n p^n \end{array} \right. \tag{3.41}$$

avec une initialisation $x^0 \in \mathbb{R}^N$, $p^0 = r^0 = Ax^0 - b$ et

$$\mu^n = \frac{\|r^n\|^2}{Ap^n \cdot p^n} \text{ et } \beta^n = \frac{\|r^{n+1}\|^2}{\|r^n\|^2}.$$

On vérifie facilement que μ^n est précisément l'unique point de minimum du polynôme quadratique $\mu \mapsto J(x^n - \mu p^n)$. On peut vérifier par récurrence que la suite r^n coincide avec le gradient $J'(x^n)$ et donc que β^n est donné par la formule de Fletcher-Rieves.

Exercice 3.2.5 Soit A une matrice symétrique définie positive. Soit $x^0 \in \mathbb{R}^N$. Pour la récurrence (3.41) démontrer que $r^n = Ax^n - b$. Montrer que la suite p^n vérifie $Ap^n \cdot p^j = 0$ pour $0 \leq j < n$ (on dit que la suite p^n 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^n 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^n "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 (voir [28] pour une preuve).

Proposition 3.2.21 *Soit A une matrice symétrique réelle définie positive, de valeurs propres ordonnées $0 < \lambda_1 \leq \cdots \leq \lambda_N$. Soit x la solution exacte du système $Ax = b$. Soit $(x^n)_{n \geq 0}$ la suite de solutions approchées du gradient conjugué, définie par* (3.41)*. Alors*

$$\|x^n - x\| \leq 2\sqrt{\lambda_N/\lambda_1} \left(\frac{1 - \sqrt{\lambda_1/\lambda_N}}{1 + \sqrt{\lambda_1/\lambda_N}} \right)^n \|x^0 - x\|_2.$$

La vitesse de convergence de l'algorithme du gradient conjugué est exactement celle de l'algorithme de la boule pesante et donc bien meilleure que celle de l'algorithme du gradient à pas fixe. Mais ce qui est remarquable dans l'algorithme du gradient conjugué, c'est l'absence de paramètres à régler dans la formule de récurrence (3.41). On n'a pas à connaitre les valeurs propres λ_1, λ_N pour obtenir la vitesse de convergence de la Proposition 3.2.21.

Remarque 3.2.22 Le rapport $\lambda_N/\lambda_1 \geq 1$, qui apparait dans la vitesse de convergence de la méthode du gradient conjugué, est appelé le conditionnement de la matrice A. Pour accélérer encore la convergence de la méthode du gradient conjugué, une idée 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). Cette idée est semblable à celle que nous avons vue à la Remarque 3.1.4. •

3.2.5 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, $V = \mathbb{R}^N$, mais le cas des espaces de Hilbert de dimension infinie n'est pas vraiment plus difficile.

Définition 3.2.23 *Si J est une fonction convexe de $\mathbb{R}^N$ dans $\mathbb{R}$, on appelle* **sous-différentiel** *de J en x l'ensemble*

$$\partial J(x) = \{p \in \mathbb{R}^N \mid J(y) - J(x) \geq p \cdot (y - x), \qquad \forall y \in \mathbb{R}^N\} \ . \tag{3.42}$$

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 3.2.24 *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 3.2.6 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

$$\mathcal{J}(x) = \sup_{\lambda \in \Lambda} J_\lambda(x), \tag{3.43}$$

où chaque fonction $J_\lambda(x)$ est convexe et différentiable sur $\mathbb{R}^N$. On calcule alors aisément un sous-gradient de $\mathcal{J}$.

Lemme 3.2.25 *Soit $\mathcal{J}$ la fonction définie par* (3.43) *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 $\mathcal{J}(y) \geq J_\lambda(y)$ et $\mathcal{J}(x) = J_\lambda(x)$, donc $\mathcal{J}(y) - \mathcal{J}(x) \geq J_\lambda(y) - J_\lambda(x) \geq J'_\lambda(x) \cdot (y - x)$, par convexité de J_λ, ce qui démontre le résultat. □

Remarque 3.2.26 Lorsque l'ensemble des paramètres Λ est supposé fini, on peut améliorer le Lemme 3.2.25 et montre que le sous-différentiel $\partial\mathcal{J}(x)$ est précisément l'enveloppe convexe des gradients $J'_\lambda(x)$ pour $\lambda \in \Gamma(x)$. •

Remarque 3.2.27 Les fonctions non régulières du type (3.43) 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 [6]). 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

$$x^{n+1} = x^n - \frac{\rho_n}{\|p^n\|} p^n \ , \tag{3.44}$$

où p^n est un sous-gradient quelconque de $\mathcal{J}$ au point x^n, et où $\rho_n > 0$ est une suite de réels strictement positifs telle que

$$\rho_n \to 0, \quad \sum_{n\in\mathbb{N}} \rho_n = +\infty, \quad \sum_{n\in\mathbb{N}} \rho_n^2 < +\infty \, . \tag{3.45}$$

Evidemment, la valeur x^{n+1} n'est bien définie que si $p^n \neq 0$. Lorsque $p^n = 0$, l'algorithme s'arrête : x^n est alors le minimum de $\mathcal{J}$ en vertu du Lemme 3.2.24. Un exemple de suite de pas ρ_n vérifiant (3.45) est $\rho_n = 1/(n+1)^{1/2+\epsilon}$ avec $\epsilon > 0$.

Proposition 3.2.28 *Soit une fonction $\mathcal{J}$ convexe, infinie à l'infini, admettant en tout point un sous-différentiel non-vide et localement Lipschitzienne, au sens où, pour tout $M > 0$, il existe $L_M > 0$ tel que*

$$\|x\| + \|y\| \leq M \Rightarrow \|\mathcal{J}(x) - \mathcal{J}(y)\| \leq L_M \|x - y\|.$$

Alors, si les pas vérifient (3.45)*, l'algorithme du sous-gradient* (3.44) *converge.*

Remarque 3.2.29 Les hypothèses de la Proposition 3.2.28 sont vérifiées pour les fonctions $\mathcal{J}$ définies par (3.43) avec un ensemble Λ fini, chacune des fonctions $J_\lambda(x)$ étant convexe différentiable et l'une d'entre elle infinie à l'infini. •

Démonstration. Comme $\mathcal{J}$ est continue et infinie à l'infini, elle admet au moins un point de minimum x^*. En utilisant (3.44) on développe la norme suivante

$$\begin{aligned} \|x^{n+1}-x^*\|^2 &= \|x^n-x^*\|^2 - 2\frac{\rho_n}{\|p^n\|}p^n\cdot(x^n-x^*)+\rho_n^2 \\ &\leq \|x^n-x^*\|^2 - 2\frac{\rho_n}{\|p^n\|}\left(\mathcal{J}(x^n)-\mathcal{J}(x^*)\right)+\rho_n^2 \end{aligned} \tag{3.46}$$

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

$$\|x^{i+1}-x^*\|^2 + 2\min_{0\leq n\leq i}\left(\mathcal{J}(x^n)-\mathcal{J}(x^*)\right)\sum_{n=0}^{i}\frac{\rho_n}{\|p^n\|} \leq \|x^0-x^*\|^2+\sum_{n=0}^{i}\rho_n^2. \tag{3.47}$$

On en déduit tout d'abord que $\|x^{i+1}-x^*\|$ est bornée à cause de la dernière condition de (3.45), donc la suite x^n est bornée par une constante M indépendante de n. On applique alors l'hypothèse que $\mathcal{J}$ est Lipschitzienne à la définition du sous-gradient p^n

$$p^n\cdot(y-x^n)\leq J(y)-J(x^n)\leq L_M\|y-x^n\|$$

d'où l'on déduit en prenant $y=x^n+\varepsilon p^n$, avec $\varepsilon>0$ aussi petit que l'on veut, que $\|p^n\|\leq L_M$. Par conséquent (3.47) implique que

$$\min_{0\leq n\leq i}\left(\mathcal{J}(x^n)-\mathcal{J}(x^*)\right)\leq\frac{L_M}{2}\frac{\|x^0-x^*\|^2+\sum_{n=0}^{\infty}\rho_n^2}{\sum_{n=0}^{i}\rho_n},$$

ce qui démontre que $\liminf_{n\to+\infty}\mathcal{J}(x^n)=\mathcal{J}(x^*)$ puisque la série des ρ_n diverge. Soit $x^{n'}$ une sous-suite de x^n telle que $\lim_{n\to+\infty}\mathcal{J}(x^{n'})=\mathcal{J}(x^*)$. Comme elle est bornée dans $\mathbb{R}^N$, il existe encore une sous-suite, toujours notée $x^{n'}$ par souci de simplicité, qui converge vers une limite x'^* qui, par continuité de J est un point de minimum de J. Cette limite x'^* n'est peut-être pas le point de minimum x^* introduit ci-dessus mais rien ne nous empêche de remplacer x^* par x'^* et tout le raisonnement reste le même car les itérées x^n et p^n ne dépendent pas du choix x^* ou x'^*. Par souci de simplicité dans les notations, on réécrit $x'^*=x^*$. Montrons que toute la suite x^n converge en fait vers x^*.

Pour $\delta>0$ petit fixé, il existe un indice n', suffisamment grand, tel que

$$\|x^{n'}-x^*\|^2\leq\delta/2 \quad \text{et} \quad \sum_{i\geq n'}\rho_i^2\leq\delta/2.$$

Alors, pour tout indice $n\geq n'$, comme (3.46) implique que

$$\|x^{n+1}-x^*\|^2\leq\|x^n-x^*\|^2+\rho_n^2,$$

en sommant on obtient

$$\|x^{n+1} - x^*\|^2 \leq \|x^{n'} - x^*\|^2 + \sum_{i=n'}^{n} \rho_i^2 \leq \delta,$$

ce qui démontre que toute la suite x^n converge vers le point de minimum x^*. □

Remarque 3.2.30 La vitesse de convergence de l'algorithme de sous-gradient (3.44) est très lente puisque pour $\rho_n = 1/(n+1)^{1/2+\epsilon}$ avec $\epsilon > 0$ on trouve que

$$0 \leq \min_{0 \leq n \leq i} \mathcal{J}(x^n) - \mathcal{J}(x^*) \leq \frac{C}{i^{1/2-\epsilon}},$$

ce qui est bien plus lent que la convergence géométrique pour l'algorithme du gradient (voir la Remarque 3.1.8). Si on suppose une propriété du type α-convexité pour J en x^*, alors on obtient la même vitesse de convergence pour $\|x^n - x^*\|^2$. •

Remarque 3.2.31 L'algorithme de sous-gradient (3.44) ne garantit pas la décroissance de la fonction objectif aux points où celle-ci n'est pas différentiable. En effet, considérons l'exemple $\mathcal{J}(x) = \max(J_1(x), J_2(x))$ avec J_1, J_2 convexes différentiables, plaçons nous en un point x où $J_1(x) = J_2(x)$ et prenons le sous-gradient $p = J_1'(x) \neq 0$. Dès que $-J_1'(x)$ n'est pas une direction de descente pour J_2, pour tout pas $\rho > 0$ suffisamment petit, on obtient $\mathcal{J}(x - \rho p/\|p\|) = J_2(x - \rho p/\|p\|) > J_2(x) = \mathcal{J}(x)$. •

Remarque 3.2.32 Il est crucial pour la convergence de l'algorithme de sous-gradient (3.44) que la suite des pas ρ_n tende vers zéro. Si ce n'était pas le cas et si en son point de minimum x^* la fonction $\mathcal{J}$ n'était pas différentiable, alors la suite x^n oscillerait sans converger vers x^*. Pour s'en convaincre, on peut reprendre l'exemple $\mathcal{J}(x) = |x|$ de l'Exercice 3.2.6 et voir que si l'initialisation x^0 n'est pas un multiple entier du pas fixe $\mu > 0$, l'algorithme du gradient à pas fixe ne converge et, à partir d'un certain rang, oscille sans arrêt entre deux points $x^+ > 0$ et $x^- < 0$. •

Remarque 3.2.33 Dans l'algorithme de sous-gradient (3.44) on normalise le sous-gradient p_n. En effet, comme l'a montré l'exemple de l'Exercice 3.2.6, 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. •

3.2.6 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 $V = \mathbb{R}^N$. La spécificité de ce type de problèmes est que la fonction à minimiser est une somme (ou une moyenne) d'un très grand nombre M de fonctions. Par exemple, en apprentissage machine M est le nombre de données

à partir desquelles on veut apprendre les paramètres $x \in \mathbb{R}^N$ du modèle explicatif. Autrement dit, étant données M fonctions f_i, on considère

$$\inf_{x\in\mathbb{R}^N} F(x) = \frac{1}{M}\sum_{i=1}^{M} f_i(x). \tag{3.48}$$

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

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

où $\mu^n > 0$ 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 $x^0 \in \mathbb{R}^N$, on construit la suite

$$x^{n+1} = x^n - \mu^n f_{i_n}'(x^n), \tag{3.49}$$

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

Proposition 3.2.34 *On suppose que $F(x)$ est α-convexe et qu'il existe une constante C, indépendante de M, telle que pour tout $x \in \mathbb{R}^N$*

$$\frac{1}{M}\sum_{i=1}^{M} \|f_i'(x)\|^2 \leq C(1 + \|x\|^2). \tag{3.50}$$

Si le pas de descente est

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

alors l'algorithme du gradient stochastique (3.49) *converge (en moyenne).*

Remarque 3.2.35 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. Soit x^* l'unique point de minimum de $F(x)$. On réécrit (3.49) sous la forme

$$x^{n+1} - x^* = x^n - x^* - \mu^n f_{i_n}'(x^n),$$

dont on calcule la norme au carré

$$\|x^{n+1} - x^*\|^2 = \|x^n - x^*\|^2 - 2\mu^n(x^n - x^*)\cdot f_{i_n}'(x^n) + (\mu^n)^2\|f_{i_n}'(x^n)\|^2. \tag{3.51}$$

On prend l'espérance de (3.51) par rapport à la seule variable aléatoire qui gouverne le choix de l'indice i_n, en notant que x^n ne dépend pas de cette variable aléatoire, pour obtenir

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

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

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

De l'α-convexité de F, voir (2.24), on déduit

$$\mathbb{E}\Big(\|x^{n+1} - x^*\|^2\Big) \leq (1 - 2\mu^n\alpha)\,\|x^n - x^*\|^2 + \frac{(\mu^n)^2}{M}\sum_{i=1}^{M} \|f'_i(x^n)\|^2. \tag{3.52}$$

Puisque

$$1 + \|x\|^2 \leq 1 + 2\|x^*\|^2 + 2\|x - x^*\|^2,$$

l'hypothèse (3.50) implique

$$\frac{1}{M}\sum_{i=1}^{M} \|f'_i(x)\|^2 \leq \tilde{C}(1 + \|x - x^*\|^2)$$

avec $\tilde{C} = \max(2C, (1 + 2\|x^*\|^2)C)$. Par souci de simplicité, on continue de noter $\tilde{C} = C$. Cette dernière inégalité, utilisée dans (3.52), donne

$$\mathbb{E}\Big(\|x^{n+1} - x^*\|^2\Big) \leq \big(1 - 2\mu^n\alpha + C(\mu^n)^2\big)\,\|x^n - x^*\|^2 + C(\mu^n)^2.$$

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

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

Une récurrence facile montre alors que

$$\mathbb{E}\Big(\|x^{n+1} - x^*\|^2\Big) \leq \Pi^n\|x^0 - x^*\|^2 + C\Pi^n\sum_{i=0}^{n} \frac{(\mu^i)^2}{\Pi^i}, \tag{3.53}$$

où, cette fois-ci, $\mathbb{E}$ désigne l'espérance pour toutes les variables aléatoires des indices de i_0 à i_n, et avec

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

On voit immédiatement que pour démontrer la convergence de l'algorithme, non seulement le pas μ^n ne doit pas être trop grand pour que Π^n tende vers zéro mais il doit aussi lui-même tendre vers zéro pour que la série dans (3.53) converge. C'est une situation similaire à celle de l'algorithme du sous-gradient (voir la Proposition 3.2.28). Quitte à diminuer la valeur de α dans la définition de la forte convexité, on peut toujours supposer que $0 < \alpha < 1/2$. Dans ce cas, on vérifie que le choix $\mu^n = 1/(n+1)$ conduit à

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

ce qui prouve la convergence en moyenne, autrement dit, $\mathbb{E}\big(\|x^n - x^*\|^2\big)$ tends vers zéro quand k tends vers l'infini. On peut démontrer la convergence presque sûrement mais cela dépasse la cadre de ce cours. □

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 M est très grand, le coût d'une itération de gradient stochastique est approximativement M plus fois faible que le coût d'une itération du gradient puisqu'on n'évalue qu'une seule dérivée f_i' au lieu de M (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^n 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 3.2.36 Entre l'algorithme du gradient stochastique (3.49) 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

$$x^{n+1} = x^n - \mu^n \frac{1}{m} \sum_{i \in \mathcal{I}_n} f_i'(x^n), \tag{3.54}$$

où $\mu^n > 0$ est un pas de descente et $\mathcal{I}_n$ est une collection de m indices distincts tirés aléatoirement (indépendamment des précédents tirages) et uniformément dans l'ensemble $\{1, ..., M\}$. •

Remarque 3.2.37 Dans la formule (3.54) on retrouve l'idée de moyenner les gradients qui a déjà été évoqué dans l'étude de l'algorithme de la boule pesante (voir la Remarque 3.2.19) et qui permet d'éviter d'éventuelles oscillations de l'algorithme, surtout dans un contexte d'optimisation stochastique avec des données bruitées. Certains algorithmes spécialisés en apprentissage machine exploitent encore plus cette idée, notamment l'algorithme ADAM [21]. •

3.2.7 Algorithme proximal

Pour motiver ce nouvel algorithme nous rappelons l'Exemple 1.2.4 de problème aux moindres carrés avec régularisation ℓ^1 pour obtenir des solutions creuses ou parcimonieuses. Soit $b \in \mathbb{R}^M$ et A une matrice de taille $M \times N$, Pour un paramètre $\tau \geq 0$, on cherche une solution $x \in \mathbb{R}^N$ du problème de minimisation

$$\min_{x \in \mathbb{R}^N} \left\{ \frac{1}{2}\|Ax - b\|^2 + \tau \|x\|_1 \right\} \quad \text{avec} \quad \|x\|_1 = \sum_{i=1}^{N} |x_i|. \tag{3.55}$$

Plus τ est grand et plus la solution x aura des composantes nulles. Au contraire, à la limite $\tau \to 0^+$, il s'agit d'un problème de moindres carrés standard dont on

sait qu'il admet toujours au moins une solution, qui est unique si $\mathrm{Ker}A = \{0\}$ (cf. Exercice 2.5.2).

Remarque 3.2.38 Expliquons sur un cas particulier très simple pourquoi la norme ℓ^1 dans (3.55) conduit à ce que la solution ait de nombreuses composantes nulles. On choisit $M = N$ et $A =$ Id. Dans ce cas, la minimisation de (3.55) se fait "à la main", composante par composante. La fonction de $\mathbb{R}$ dans $\mathbb{R}$,

$$x_i \ \mapsto \ \frac{1}{2}|x_i - b_i|^2 + \tau|x_i|,$$

est fortement convexe, dérivable partout sauf en 0, et admet donc un unique point de minimum x_i^* qui vérifie, soit $x_i^* = 0$, soit la condition d'optimalité

$$x_i^* - b_i + \tau = 0 \text{ si } x_i^* > 0, \quad x_i^* - b_i - \tau = 0 \text{ si } x_i^* < 0.$$

On en déduit facilement l'unique point de minimum qui est donné par l'opérateur de **contraction** (shrinkage, en anglais)

$$x_i^* = S_\tau(b_i) = \begin{cases} 0 & \text{si } |b_i| < \tau, \\ b_i - \ \mathrm{sgn}(b_i)\tau & \text{si } |b_i| \geq \tau. \end{cases} \tag{3.56}$$

L'interprétation est simple si l'on pense que la donnée b est bruitée et que τ correspond au seuil maximal du bruit. Si $|b_i| < \tau$, la donnée est indistinguable du bruit et on obtient $x_i^* = 0$. Sinon, on obtient $x_i^* \approx b_i$ à une correction de l'ordre de τ près. La solution $x^* \in \mathbb{R}^N$ est donc d'autant plus parcimonieuse que le seuil τ est élevé. •

Pour résoudre (3.55) on ne peut pas employer une méthode de gradient classique car la norme ℓ^1, $x \mapsto \|x\|_1$, n'est pas dérivable en tout point dont une des composantes s'annule. On pourrait utiliser l'algorithme du sous-gradient car la fonction objectif est convexe mais on a vu que cet algorithme converge lentement. Il y a donc de la place pour concevoir un algorithme plus efficace. Plus généralement, on considère la minimisation de la somme de deux fonctions convexes sur $\mathbb{R}^N$

$$\min_{x \in \mathbb{R}^N} J_1(x) + \tau J_2(x)$$

avec J_1 différentiable et J_2 convexe mais non différentiable. L'exemple ci-dessus correspond à $J_1(x) = \frac{1}{2}\|Ax - b\|^2$ et $J_2(x) = \|x\|_1$. L'idée est d'appliquer deux algorithmes différents à J_1 et J_2. Evidemment, pour J_1 on a l'embarras du choix car nous avons déjà vu plein d'algorithmes de type gradient ! Par contre, pour J_2 nous ne connaissons que l'algorithme de sous-gradient dont on sait qu'il est peu efficace. C'est pourquoi nous introduisons maintenant l'algorithme **proximal** dont le principe est de faire un algorithme de gradient **implicite**, c'est-à-dire de construire la suite des itérées

$$x^{n+1} = x^n - \tau J_2'(x^{n+1}), \tag{3.57}$$

où $\tau > 0$ s'interprète comme un pas de descente. Evidemment, c'est un algorithme "conceptuel" car il faut résoudre une équation implicite en x^{n+1}, ce qui n'est pas facile. En plus, il faudrait que J_2 soit différentiable, ce qui n'est pas le cas. Néanmoins, on remarque que (3.57) serait la condition d'optimalité de

$$\min_{x\in\mathbb{R}^N} \frac{1}{2}\|x - x^n\|^2 + \tau J_2(x) \tag{3.58}$$

si x^{n+1} en est le point de minimum. Or, si (3.57) n'a pas de sens pour une fonction J_2 non différentiable, au contraire (3.58) admet bien un unique point de minimum, noté x^{n+1}, dès que J_2 est convexe, puisque la fonction à minimiser est fortement convexe. Ainsi, (3.58) permet de définir **l'application proximale** de τJ_2

$$\mathrm{prox}_{\tau J_2}(x^n) = x^{n+1}, \tag{3.59}$$

où x^{n+1} est le point de minimum de (3.58). Avec une initialisation $x^0 \in \mathbb{R}^N$, (3.59) définit un algorithme de minimisation pour J_2, appelé **algorithme proximal**. On vérifie bien qu'il s'agit d'un algorithme de descente car $J_2(x^{n+1}) < J_2(x^n)$, sauf si $x^{n+1} = x^n$ auquel cas la suite x^m est stationnaire pour $m \geq n$. Cependant, cet algorithme est toujours de type conceptuel car pour une fonction convexe quelconque J_2 on ne sait pas résoudre le problème de minimisation (3.58) dont la solution est requise pour définir l'algorithme. En pratique, cet algorithme est donc limité à des fonctions simples pour lesquelles on sait résoudre "à la main" (3.58). C'est précisément le cas pour $J_2(x) = \|x\|_1$, comme on vient de le voir à la Remarque 3.2.38. Dans ce cas, on a $x^{n+1} = \mathrm{prox}_{\tau\|\cdot\|_1}(x^n) = S_\tau(x^n)$, où S_τ est l'opérateur de contraction défini par (3.56).

Finalement, on propose un algorithme de type explicite-implicite pour minimiser $J_1 + \tau J_2$: soit une initialisation $x^0 \in \mathbb{R}^N$, pour $n \geq 0$, on calcule

$$\left\{ \begin{array}{l} x^{n+1/2} = x^n - \tau J_1'(x^n), \\ x^{n+1} = \mathrm{prox}_{\tau J_2}(x^{n+1/2}), \end{array} \right. \tag{3.60}$$

où on a appliqué une itération de l'algorithme du gradient à pas fixe à J_1 suivi d'une itération de l'algorithme proximal à J_2 avec le même pas fixe τ pour les deux. Pour le cas particulier (3.55), cet algorithme proximal est

$$x^{n+1} = S_\tau\Big(x^n - \tau A^*(Ax^n - b)\Big).$$

Il existe de nombreuses variantes de l'algorithme (3.60), notamment en ne choisissant pas le même pas de descente τ pour les deux étapes correspondant aux deux fonctions J_1 et J_2.

Exercice 3.2.7 Soit K un convexe fermé de $\mathbb{R}^N$. On définit la fonction indicatrice de K comme l'application de $\mathbb{R}^N$ dans $\mathbb{R} \cup \{+\infty\}$ définie par

$$I_K(x) = \left\{ \begin{array}{ll} 0 & \text{si } x \in K, \\ +\infty & \text{si } x \notin K. \end{array} \right.$$

Vérifier que I_K est une fonction convexe et que l'application proximale prox_{I_K} est simplement l'opérateur de projection orthogonale sur K.

Exercice 3.2.8 Soit J une fonction convexe minorée par une fonction affine sur $\mathbb{R}^N$. Pour $\tau > 0$, on définit la régularisation de Moreau-Yosida de J, notée J_τ, par

$$J_\tau(x) = \min_{y\in\mathbb{R}^N} \frac{1}{2\tau}\|x-y\|^2 + J(y).$$

Montrer qu'il existe un unique point de minimum, noté $y = \text{prox}_{\tau J}(x)$, dans la définition ci-dessus de J_τ. Vérifier que J_τ est convexe et que $\min_{x\in\mathbb{R}^N} J(x) = \min_{x\in\mathbb{R}^N} J_\tau(x)$. (On peut montrer que J_τ est différentiable, même si J ne l'est pas, mais c'est délicat.) En supposant que J est différentiable, calculer le gradient de J_τ (indépendant du gradient de J !) et vérifier que

$$\text{prox}_{\tau J}(x^n) = x^{n+1} \Leftrightarrow x^{n+1} = x^n - \tau J'_\tau(x^n),$$

c'est-à-dire qu'une itération de l'algorithme proximal pour J est équivalent à une itération de l'algorithme du gradient à pas fixe pour J_τ.

Exercice 3.2.9 Soit J_1 une fonction fortement convexe différentiable sur $\mathbb{R}^N$. Soit K un convexe fermé de $\mathbb{R}^N$ et $J_2 = I_K$, la fonction indicatrice de K définie comme à l'Exercice 3.2.7. Vérifier que l'algorithme explicite-implicite (3.60) n'est rien d'autre que l'algorithme de gradient projeté de la Sous-section 3.4.1.

Exercice 3.2.10 Soit A une matrice $N \times N$ symétrique réelle, définie positive, et $b \in \mathbb{R}^N$. On définit la fonction quadratique J sur $\mathbb{R}^N$ par

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

Déterminer l'opérateur proximal prox_J.

Nous ne démontrons pas ici la convergence de l'algorithme explicite-implicite (3.60) (voir par exemple [11]) et nous nous contentons de démontrer la convergence de l'algorithme proximal (3.57). Par souci de simplicité nous n'allons même prouver cette convergence que dans le cas particulier où la fonction J_2 est différentiable, ce qui n'est pas le cas d'application le plus intéressant. Néanmoins, la preuve de la convergence s'étend aux fonctions convexes, non-différentiables, grâce à la notion de sous-gradient mais la preuve est plus technique.

Lemme 3.2.39 *Soit $J(x)$ une fonction convexe différentiable de $\mathbb{R}^N$ dans $\mathbb{R}$ qui admet un unique point de minimum x^*. Pour $\tau > 0$ et $x^0 \in \mathbb{R}^N$ on définit l'algorithme proximal par la suite, indexée par $n \geq 0$,*

$$x^{n+1} = \text{prox}_{\tau J}(x^n) = \arg\min_{x\in\mathbb{R}^N} \left\{ \frac{1}{2}\|x-x^n\|^2 + \tau J(x) \right\}.$$

Alors la suite x^n converge vers le point de minimum x^*. De plus, si J est α-convexe, alors

$$\|x^n - x^*\|^2 \leq \|x^0 - x^*\|^2 \frac{1}{(1+2\tau\alpha)^n}.$$

Remarque 3.2.40 Le Lemme 3.2.39 donne la convergence de l'algorithme proximal sans aucune condition de petitesse sur le pas τ. On retrouve là une caractéristique des algorithmes implicites par rapport aux algorithmes explicites (voir par exemple [1] pour les schémas de différences finies). •

Démonstration. Comme J est convexe, la fonction $I(x) = J(x) + \|x - x^n\|^2/(2\tau)$ est $(1/\tau)$-convexe. On écrit la forte convexité (2.23) de $I(x)$ à son point de minimum x^{n+1} (où sa dérivée s'annule)

$$J(x) + \frac{1}{2\tau}\|x - x^n\|^2 \geq J(x^{n+1}) + \frac{1}{2\tau}\|x^{n+1} - x^n\|^2 + \frac{1}{2\tau}\|x - x^{n+1}\|^2. \tag{3.61}$$

En prenant $x = x^n$ on trouve que $J(x^{n+1}) \leq J(x^n)$. Si on choisit $x = x^*$ et que l'on somme l'inégalité ci-dessus pour tous les indices de 0 à $(n-1)$, on obtient

$$\frac{1}{2\tau}\|x^* - x^0\|^2 \geq \sum_{i=1}^{n}(J(x^i) - J(x^*)) + \frac{1}{2\tau}\|x^* - x^n\|^2 \geq n(J(x^n) - J(x^*))$$

dont on déduit que $J(x^n)$ converge vers $J(x^*)$ et que la suite (x^n) est bornée dans $\mathbb{R}^N$. Par conséquent, pour une sous-suite, $x^{n'}$ converge vers une limite x^∞ et, comme J est continue et que son point de minimum est unique, il vient que $x^\infty = x^*$ et que toute la suite x^n converge vers x^*.

Si J est α-convexe, alors on peut améliorer (3.61) car $I(x) = J(x) + \|x - x^n\|^2/(2\tau)$ est $(\alpha + 1/\tau)$-convexe

$$J(x) + \frac{1}{2\tau}\|x - x^n\|^2 \geq J(x^{n+1}) + \frac{1}{2\tau}\|x^{n+1} - x^n\|^2 + \frac{1}{2}(\alpha + 1/\tau)\|x - x^{n+1}\|^2.$$

On choisit encore $x = x^*$ et on combine cette inégalité avec la forte convexité (2.23) de J à son point de minimum x^* (où sa dérivée s'annule)

$$J(x^{n+1}) \geq J(x^*) + \frac{\alpha}{2}\|x^{n+1} - x^*\|^2.$$

En sommant les deux dernières inégalités on obtient

$$\|x^* - x^n\|^2 \geq (1 + 2\tau\alpha)\|x^* - x^{n+1}\|^2$$

d'où le résultat. □

3.3 Méthode de Newton

Tous les algorithmes que nous avons étudiés jusqu'ici sont des algorithmes dits du premier ordre car ils n'utilisent que la dérivée première de la fonction à minimiser. Au contraire, la méthode de Netwon est un algorithme d'ordre deux car elle utilise l'information des deux premières dérivées de la fonction que l'on veut minimiser. On imagine ainsi qu'elle peut être plus efficace qu'un algorithme du premier ordre, puisqu'elle utilise, en plus du gradient, la Hessienne de la fonction, mais qu'elle risque d'être plus compliquée à mettre en oeuvre puisqu'il faut justement calculer ces dérivées secondes. Comme dans les sections précédentes on considère un problème d'optimisation sans contrainte, sauf dans la dernière sous-section.

3.3.1 Cas de la dimension finie

Pour simplifier la présentation, on se place tout d'abord 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, à partir de maintenant, nous remplaçons J' par une fonction F de $\mathbb{R}^N$ dans $\mathbb{R}^N$, que nous supposons être de classe C^2. Soit u un zéro régulier de F c'est-à-dire que

$$F(u) = 0 \quad \text{et} \quad F'(u) \text{ matrice inversible.}$$

Rappelons que la matrice Jacobienne $F'(u)$ est définie par $\left(\frac{\partial F_i(u)}{\partial u_j}\right)_{1\leq i,j\leq N}$. 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

$$u^{n+1} = u^n - \left(F'(u^n)\right)^{-1} F(u^n) \quad \text{pour} \quad n \geq 0. \tag{3.62}$$

Rappelons que l'on ne calcule pas l'inverse de la matrice $F'(u^n)$ dans (3.62) 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

$$J(w) = J(v) + J'(v)\cdot(w-v) + \frac{1}{2}J''(v)(w-v)\cdot(w-v) + \mathcal{O}\left(\|w-v\|^3\right), \tag{3.63}$$

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 (3.63) 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 (3.62).

L'avantage principal de la méthode de Newton est sa convergence bien plus rapide que les méthodes précédentes.

Proposition 3.3.1 *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 (3.62) converge, c'est-à-dire que la suite (u^n) converge vers u, au sens où*

$$\|u^{n+1} - u\| \leq C\|u^n - u\|^2 \quad \text{et} \quad \|u^n - u\| \leq C^{-1}(C\epsilon)^{2^n}. \tag{3.64}$$

Remarque 3.3.2 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 3.1.8). La conséquence de la première inégalité de (3.64) est que le nombre de chiffres significatifs dans l'approximation u^n de la solution u double à chaque itération. Néanmoins cette convergence rapide a un prix car, à chaque itération de la méthode de Newton (3.62), il faut résoudre un système linéaire, ce qui est coûteux en temps de calcul. De plus, la convergence rapide donnée par (3.64) 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. En pratique, même dans des cas très simples (y compris en dimension $N = 1$), la méthode de Newton peut diverger pour certaines données initiales u^0. Par ailleurs, la convergence quadratique (3.64) ne se produit qu'au voisinage d'un zéro régulier, c'est-à-dire sous l'hypothèse que $F'(u)$ est inversible, comme le montre le contre-exemple de l'Exercice 3.3.2. Enfin, 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 (3.64) 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 (3.62)

$$u^{n+1} - u = u^n - u - \left(F'(u^n)\right)^{-1} \left(F(u^n) - F(u)\right).$$

On utilise alors le développement de Taylor suivant, autour de u^n avec reste exact,

$$F(u) = F(u^n) + F'(u^n)(u - u^n) + H^n(u - u^n) \cdot (u - u^n)$$

avec

$$H^n = \int_0^1 F''(u^n + s(u - u^n))(1 - s)\, ds\,,$$

qui permet d'obtenir

$$u^{n+1} - u = \left(F'(u^n)\right)^{-1} H^n(u^n - u) \cdot (u^n - u).$$

On peut alors majorer pour en déduire

$$\|u^{n+1} - u\| \leq \frac{1}{2} C_1 C_2 \|u^n - u\|^2.$$

qui n'est rien d'autre que la première partie de (3.64) avec $C = C_1 C_2/2$. 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 3.3.3 Un inconvénient majeur de la méthode de Newton est la nécessité d'avoir une initialisation u^0 proche de la solution (que l'on ne connait pas !). Pour pallier à ce problème et rendre la méthode robuste et convergente quelque soit le vecteur initial, il est possible d'hybrider l'algorithme en introduisant un pas de descente $0 < \mu^n \leq 1$ et en modifiant l'algorithme (3.62) comme suit

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

où on adapte μ^n pour garantir la décroissance de la fonction $J(u^n)$. Si J est convexe, on vérifie facilement que $\left(J''(u^n) \right)^{-1} J'(u^n)$ est une direction de descente. L'idée est de démarrer avec μ^n petit, puis d'augmenter μ^n jusqu'à la valeur 1 qui redonne la méthode de Newton (voire [26] pour les détails pratiques). •

Remarque 3.3.4 Un autre 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 $\left(F'(u^n) \right)^{-1}$. On remplace alors la formule (3.62) 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 [7] et [13]. •

Exercice 3.3.1 Soit la fonction J de $\mathbb{R}^N$ dans $\mathbb{R}$ définie par

$$J(x) = \frac{1+\alpha\|x\|^2}{2}Ax\cdot x - b\cdot x,$$

où A est une matrice symétrique définie positive de taille N, $b\in\mathbb{R}^N$ et $\alpha\geq 0$. Appliquer la méthode de Newton à la minimisation de J et étudier sa convergence.

Exercice 3.3.2 Soit la fonction $F(x) = \|x\|^2 x$ de $\mathbb{R}^N$ dans $\mathbb{R}^N$. Vérifier que 0 est le seul point qui annule F et que $F'(0) = 0$. Montrer que la méthode de Newton appliqué à $F(x)$ converge pour toute initialisation $x^0\in\mathbb{R}^N$ mais que la convergence n'est que géométrique (comme pour un algorithme de type gradient), c'est-à-dire que la conclusion (3.64) de la Proposition 3.3.1 n'a pas lieu.

3.3.2 Méthode de Gauss-Newton

Il s'agit d'une variante de l'algorithme Newton pour les problèmes de moindres carrés non-linéaires, qui s'écrivent

$$\min_{u\in\mathbb{R}^N}\|F(u)\|^2 \quad \text{avec } F(u) = (F_1(u),\cdots,F_M(u)),$$

où la fonction $F:\mathbb{R}^N\to\mathbb{R}^M$ est supposée de classe C^2. Si $F(u) = Au - b$, on retrouve les moindres carrés linéaires. L'idée de la méthode de Gauss-Newton est, à partir d'une initialisation $u^0\in\mathbb{R}^N$, de construire une suite de solutions approchées u^n, pour $n\geq 0$. Comme pour l'algorithme de Newton, à chaque itération n, $F(u)$ est approchée par son application linéaire tangente en u^n, à savoir

$$F(u)\approx F(u^n) + F'(u^n)(u-u^n) = A^n u - b^n,$$

où $A^n = F'(u^n)$ est une matrice $M\times N$ et $b^n = F'(u^n)u^n - F(u^n)$ est un vecteur de $\mathbb{R}^M$. On calcule alors la nouvelle itérée u^{n+1} comme un point de minimum de

$$\min_{u\in\mathbb{R}^N}\|A^n u - b^n\|^2,$$

c'est-à-dire que u^{n+1} est une solution de l'équation normale, $F'(u^n)^*F'(u^n)u^n = F'(u^n)^*b^n$ où, comme d'habitude, M^* désigne la matrice transposée de M. Si l'on suppose que $\operatorname{Ker}F'(u^n) = \{0\}$, alors la matrice $N\times N$ $F'(u^n)^*F'(u^n)$ est inversible et on vérifie aisément que u^{n+1} est donnée par

$$u^{n+1} = u^n - \Big(F'(u^n)^*F'(u^n)\Big)^{-1}F'(u^n)^*F(u^n)$$

S'il existe un zéro régulier u de F et si l'initialisation u^0 en est proche, on peut démontrer la convergence quadratique de l'algorithme de Gauss-Newton vers u.

Remarque 3.3.5 Si $N = M$ et $\operatorname{Ker}F'(u^n) = \{0\}$, alors $F'(u^n)$ est une matrice inversible et la formule ci-dessus se simplifie en

$$u^{n+1} = u^n - \Big(F'(u^n)\Big)^{-1}F(u^n)$$

et on retrouve exactement l'algorithme de Newton pour trouver les solutions de $F(u) = 0$. •

3.3.3 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 (3.63) 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 $V \times V \ni (w_1, w_2) \mapsto J''(u^n)(w_1, w_2)$ 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.

3.3.4 Méthode de Newton avec contraintes d'égalité

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

$$\min_{v \in \mathbb{R}^N \ , \ G(v)=0} J(v) \ . \tag{3.67}$$

Si les vecteurs $(G_1'(u), ..., G_M'(u))$ sont linéairement indépendants, la condition nécessaire d'optimalité du Théorème 2.5.6 est

$$J'(u) + \sum_{i=1}^{M} \lambda_i G_i'(u) = 0 \ , \quad G_i(u) = 0 \quad 1 \leq i \leq M \ , \tag{3.68}$$

où les $\lambda_1, \ldots, \lambda_M \in \mathbb{R}$ sont les multiplicateurs de Lagrange. On peut alors résoudre le système (3.68) 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 (3.62) 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 2.5.14. 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

$$(J''(u) + \lambda \cdot G''(u))\,(w,w) > 0 \quad \forall\, w \in \mathrm{Ker}G'(u), w \neq 0\,, \tag{3.69}$$

la matrice $F'(u,\lambda)$ est inversible. On remarque que (3.69) est l'inégalité stricte dans la condition d'optimalité d'ordre 2 de la Proposition 2.5.14. Il est donc naturel de faire l'hypothèse (3.69) qui permet d'utiliser l'algorithme de Newton. On peut ainsi démontrer la convergence de cette méthode (voir [7]). 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

$$\min_{\substack{w\in\mathbb{R}^N \\ G(u^n)+G'(u^n)\cdot(w-u^n)=0}} Q^n(w)\,, \tag{3.70}$$

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 (3.70). On remarque que dans (3.70) 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 3.3.6 Dans (3.70) 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

$$\min_{\substack{w\in\mathbb{R}^N \\ G(v)+G'(v)\cdot(w-v)=0}} \left(J(v) + J'(v)\cdot(w-v) + \frac{1}{2}J''(v)(w-v)\cdot(w-v)\right). \tag{3.71}$$

Malheureusement la méthode basée sur (3.71) peut ne pas converger ! En particulier, il n'est pas évident que le Hessien $J''(v)$ soit défini 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 2.5.14). •

3.4 Algorithmes de type gradient (avec contraintes)

On s'intéresse maintenant à la résolution numérique de problèmes d'optimisation avec contraintes

$$\inf_{v\in K} J(v) \ , \tag{3.72}$$

où J est une fonction α-convexe différentiable définie sur K, sous-ensemble fermé non vide de l'espace de Hilbert réel V. Nous commencerons par étudier le cas où K est convexe mais nous nous en éloignerons progressivement.

3.4.1 Algorithme de gradient à pas fixe avec projection

Nous nous limitons dans cette sous-section au cas où K est un convexe fermé non vide de V. Sous cette hypothèse, le Théorème 2.3.9 assure l'existence et l'unicité de la solution u de (3.72), caractérisée d'après le Théorème 2.5.1 par la condition d'optimalité

$$\langle J'(u), v-u\rangle \geq 0 \quad \forall\, v \in K \ . \tag{3.73}$$

Pour tout réel $\mu > 0$ (qui jouera le rôle d'un pas de descente), (3.73) s'écrit

$$\langle u - \big(u - \mu J'(u)\big), v-u\rangle \geq 0 \quad \forall\, v \in K \ . \tag{3.74}$$

Notons P_K l'opérateur de projection orthogonale sur l'ensemble convexe K, défini à la Remarque 8.1.4. D'après le Théorème 8.1.3 de projection sur un convexe, (3.74) n'est rien d'autre que la caractérisation de u comme la projection orthogonale de $u - \mu J'(u)$ sur K, c'est-à-dire,

$$u = P_K\big(u - \mu J'(u)\big) \ . \tag{3.75}$$

Plus précisément, (3.75) est en fait équivalent à (3.73) et caractérise donc la solution u de (3.72). L'algorithme de **gradient à pas fixe avec projection** (ou plus simplement de gradient projeté) est alors défini par l'itération

$$u^{n+1} = P_K\big(u^n - \mu J'(u^n)\big) \ , \tag{3.76}$$

où $\mu > 0$ est un pas de descente fixe. Autrement dit, (3.76) est une simple adaptation de l'algorithme de gradient à pas fixe auquel on a ajouté une étape de projection pour ne pas quitter l'ensemble admissible K.

Théorème 3.4.1 *On suppose que J est α-convexe différentiable et que J' est Lipschitzien sur V (de constante L, voir (3.9)). Alors, si $0 < \mu < 2\alpha/L^2$, l'algorithme de gradient à pas fixe avec projection converge : quel que soit $u^0 \in K$, la suite (u^n) définie par (3.76) converge vers la solution u de (3.72).*

Démonstration. La démonstration reprend celle du Théorème 3.1.6 et plus précisément la preuve de (3.10) qui montre que l'application $v \mapsto v - \mu J'(v)$ est strictement contractante lorsque $0 < \mu < 2\alpha/L^2$, c'est-à-dire qu'il existe $\gamma \in]0,1[$ tel que

$$\forall v, w \in V, \quad \big\| \big(v - \mu J'(v)\big) - \big(w - \mu J'(w)\big) \big\| \leq \gamma \|v - w\| \,.$$

En effet, cette inégalité est une conséquence de (3.10) où on a remplacé u^n par v et u par 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, donc elle admet un unique point fixe (qui n'est autre que u, solution de (3.75)) vers lequel converge la suite (u^n) définie par (3.76). □

Exercice 3.4.1 Soit $V = \mathbb{R}^N$ et $K = \{x \in \mathbb{R}^N \text{ tel que } \sum_{i=1}^N x_i = 1\}$. Expliciter l'opérateur de projection orthogonale P_K et interpréter dans ce cas la formule (3.75) en terme de multiplicateur de Lagrange.

Le Théorème 3.4.1 montre que l'algorithme de gradient à pas fixe avec projection est applicable à une large classe de problèmes d'optimisation convexe avec contraintes, dès lors que l'on sait calculer l'opérateur de projection orthogonale P_K. Mais malheureusement, il est très difficile, non seulement de connaitre explicitement l'opérateur de projection P_K, mais même de construire un algorithme simple et efficace pour l'évaluer, dans le cas d'un convexe fermé K quelconque dans V. Malgré sa simplicité apparente l'algorithme du gradient projeté est donc souvent un leurre du point de vue pratique par défaut d'une connaissance explicite de cet opérateur de projection P_K.

Donnons néanmoins deux exceptions importantes pour lesquelles on sait calculer explicitement P_K. Par souci de simplicité dans la présentation, on se limite au cas de la dimension finie $V = \mathbb{R}^N$, mais ces deux exemples se généralisent sans difficulté à des espaces de Hilbert de dimension infinie. On considère tout d'abord le cas d'un sous-espace affine

$$K = \{x \in \mathbb{R}^N \ Bx = c\},$$

avec B matrice $M \times N$ de rang maximal, $\mathrm{rang}(B) = M \leq N$, et $c \in \mathbb{R}^M$. On trouve alors facilement que

$$P_K(x) = x + B^*(BB^*)^{-1}(c - Bx).$$

En effet, la condition d'optimalité du Théorème 2.5.1 (inéquation d'Euler) est en fait une égalité qui dit que $(x - P_K(x))$ est orthogonal à $\mathrm{Ker}B$ donc appartient à $\mathrm{Im}B^*$, c'est-à-dire $x - P_K(x) = B^*y$. Enfin, la condition $BP_K(x) = c$ permet de calculer $y = (BB^*)^{-1}(Bx - c)$.

Un second cas particulier important concerne les sous-ensembles K de la forme

$$K = \prod_{i=1}^N [a_i, b_i] \tag{3.77}$$

(avec éventuellement $a_i = -\infty$ ou $b_i = +\infty$ pour certains indices i). En effet, il est alors facile de voir que, si $x = (x_1, x_2, \ldots, x_N) \in \mathbb{R}^N$, $y = P_K(x)$ a pour composantes

$$y_i = \min\big(\max(a_i, x_i), b_i\big) \quad \text{pour} \quad 1 \leq i \leq N \,, \tag{3.78}$$

autrement dit, il suffit juste de "tronquer" les composantes de x.

Remarque 3.4.2 Ce qui rend très attractif, mais très difficile à mettre en oeuvre, l'algorithme du gradient projeté est qu'il s'agit d'un **algorithme faisable**, c'est-à-dire qu'à chaque itération la solution approchée u^n appartient bien à l'ensemble K qui définit les contraintes. A l'opposé de cette approche il existe une classe d'**algorithmes infaisables** où la suite de solutions approchées u^n n'appartient pas à K mais, en cas de convergence, leur limite u appartient bien à K. C'est évidemment beaucoup plus facile à mettre en oeuvre mais, en général, avant la convergence on n'a pas une solution, même de qualité médiocre, qui satisfait les contraintes. La section suivante donne des exemples de tels algorithmes infaisables. •

3.4.2 Algorithme d'Uzawa

L'idée de l'algorithme d'Uzawa vient de la théorie de la dualité, entrevue dans la Section 2.6, qui permet de "transformer" un problème avec contraintes d'inégalités générales en un problème "dual" avec comme seule contrainte d'appartenir au convexe $(\mathbb{R}_+)^M$, qui est du type (3.77). Grâce à la formule explicite (3.78) de projection sur ce convexe particulier, ce problème dual peut être résolu par la méthode du gradient à pas fixe avec projection. La solution du problème primal pourrait ensuite être obtenue, par exemple, en résolvant un problème de minimisation **sans contrainte** pour le Lagrangien. En fait, l'algorithme d'Uzawa est aussi une méthode de recherche de point-selle pour le Lagrangien et il calcule à la fois une solution du problème dual et une du problème primal.

Considérons donc le problème de minimisation convexe

$$\inf_{F(v)\leq 0} J(v) \ , \tag{3.79}$$

où J est une fonctionnelle convexe définie sur V et F une fonction convexe de V dans $\mathbb{R}^M$. Sous les hypothèses du Théorème de Kuhn et Tucker 2.6.4, la résolution de (3.79) revient à trouver un point-selle (u, p) du Lagrangien

$$\mathcal{L}(v, q) = J(v) + q \cdot F(v) \ , \tag{3.80}$$

sur $V \times (\mathbb{R}_+)^M$. A partir de la Définition 2.6.1 du point-selle

$$\forall\, q \in (\mathbb{R}_+)^M \quad \mathcal{L}(u, q) \leq \mathcal{L}(u, p) \leq \mathcal{L}(v, p) \quad \forall\, v \in V \ , \tag{3.81}$$

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

$$p = P_{\mathbb{R}_+^M}(p + \mu F(u)) \quad \forall\, \mu > 0 \ , \tag{3.82}$$

$P_{\mathbb{R}_+^M}$ désignant la projection de $\mathbb{R}^M$ sur $(\mathbb{R}_+)^M$.

Au vu de cette propriété et de la seconde inégalité dans (3.81), 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

$$\begin{aligned} &\mathcal{L}(u^n, p^n) = \inf_{v \in V} \mathcal{L}(v, p^n) \ , \\ &p^{n+1} = P_{\mathbb{R}_+^M}(p^n + \mu F(u^n)) \ , \end{aligned} \tag{3.83}$$

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 3.4.3 *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 L telle que*

$$\|F(v) - F(w)\| \leq L\|v - w\| \quad \forall\, v, w \in V \ , \tag{3.84}$$

et que l'ensemble $K = \{v \in V \text{ tel que } F(v) \leq 0\}$ est non vide. On suppose que les contraintes sont qualifiées pour la solution u de (3.79). Si $0 < \mu < 2\alpha/L^2$, l'algorithme d'Uzawa converge : quel que soit l'élément initial p^0, la suite (u^n) définie par (3.83) converge vers la solution u du problème (3.79).

Démonstration. Tout d'abord, l'existence et l'unicité d'une solution u de (3.79) découle du Théorème 2.3.9 puisque l'on minimise une fonction fortement convexe sur un convexe fermé non vide K. Comme les contraintes sont qualifiées en u, le Théorème de Kuhn et Tucker 2.6.4 affirme alors qu'il existe un multiplicateur de Lagrange $p \in (\mathbb{R}_+)^M$ tel que (u, p) est un point-selle du Lagrangien (3.80) sur $V \times (\mathbb{R}_+)^M$. De même, p^n étant fixé, le problème de minimisation dans (3.83) 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

$$\langle J'(u), v - u\rangle + p \cdot \big(F(v) - F(u)\big) \geq 0 \quad \forall\, v \in V \ , \tag{3.85}$$

$$\langle J'(u^n), v - u^n\rangle + p^n \cdot \big(F(v) - F(u^n)\big) \geq 0 \quad \forall\, v \in V \ . \tag{3.86}$$

Prenant successivement $v = u^n$ dans (3.85) et $v = u$ dans (3.86) 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 $r^n = p^n - p$

$$r^n \cdot \big(F(u^n) - F(u)\big) \leq -\alpha\|u^n - u\|^2 \ . \tag{3.87}$$

D'autre part, la projection $P_{\mathbb{R}^M_+}$ étant faiblement contractante d'après (8.2), en soustrayant (3.82) à (3.83) 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 (3.84) et (3.87), il vient

$$\|r^{n+1}\|^2 \leq \|r^n\|^2 + \big(L^2\mu^2 - 2\mu\alpha\big)\|u^n - u\|^2 \,.$$

Si $0 < \mu < 2\alpha/L^2$, on peut trouver $\beta > 0$ tel que $L^2\mu^2 - 2\mu\alpha < -\beta$, d'où

$$\beta\|u^n - u\|^2 \leq \|r^n\|^2 - \|r^{n+1}\|^2 \,. \tag{3.88}$$

Ceci montre alors que la suite $\|r^n\|^2$ est décroissante : le membre de droite de (3.88) tend donc vers 0, ce qui entraîne que u^n tend vers u. □

Remarque 3.4.4 Il est possible d'appliquer l'algorithme d'Uzawa à un problème de minimisation avec contraintes d'égalité, $F(u) = 0$. Dans ce cas, le multiplicateur de Lagrange n'a pas de signe prescrit et l'algorithme s'écrit, pour $p^0 \in \mathbb{R}^M$,

$$\mathcal{L}(u^n, p^n) = \inf_{v \in V} \mathcal{L}(v, p^n) \,,$$
$$p^{n+1} = p^n + \mu F(u^n)$$

avec un pas $\mu > 0$. Néanmoins, le Théorème 3.4.3 de convergence ne s'applique que si les contraintes $F(u)$ sont affines car alors F et $-F$ sont à la fois convexes.

Ainsi, l'algorithme d'Uzawa permet d'approcher la solution de (3.79) en remplaçant ce problème avec contraintes par une suite de problèmes de minimisation sans contraintes (3.83) (autres que la positivité du multiplicateur de Lagrange). A chaque itération, la détermination de p^n est élémentaire, puisque d'après (3.78) l'opérateur de projection $P_{\mathbb{R}^M_+}$ est une simple troncature à zéro des composantes négatives. Il faut aussi noter que le Théorème 3.4.3 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 2.6.14 et l'Exercice 3.4.2 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 (3.79) s'écrit

$$\sup_{q \geq 0} \mathcal{G}(q) \,, \tag{3.89}$$

où, par définition

$$\mathcal{G}(q) = \inf_{v \in V} \mathcal{L}(v, q) \,, \tag{3.90}$$

et que le multiplicateur de Lagrange p est une solution du problème dual (3.89). 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 (3.90). 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 (3.83) 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 (3.76) vient du fait que le problème dual (3.89) 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 3.4.2 On reprend l'exemple de l'Exercice 2.5.10, à savoir un problème de minimisation quadratique avec contraintes affines d'égalité

$$\min_{v \in \mathbb{R}^N,\ Bv - c = 0} \left\{ J(v) = \frac{1}{2} Av \cdot v - b \cdot v \right\}, \tag{3.91}$$

où A est une matrice $N \times N$ symétrique, définie positive et B est une matrice $M \times N$ de rang $M \leq N$. Appliquer l'algorithme d'Uzawa à (3.91) (cf. la Remarque 3.4.4) et démontrer que la suite p^n converge vers p, l'unique multiplicateur de Lagrange de (3.91).

Exercice 3.4.3 On reprend l'exemple (3.91) de l'Exercice 3.4.2. Montrer que si le rang de B est M et A est définie positive, alors les valeurs propres de la matrice $BA^{-1}B^*$ de taille $M \times M$ sont strictement positives

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_M.$$

Montrer que la vitesse de convergence optimale de l'algorithme d'Uzawa est obtenue pour $\mu = 2/(\lambda_1 + \lambda_M)$ et que l'erreur vérifie

$$\|u^n - u\| + \|p^n - p\| \leq C \left(\frac{1 - \lambda_1/\lambda_N}{1 + \lambda_1/\lambda_N} \right)^n,$$

où (u, p) désigne le couple solution et multiplicateur de Lagrange.

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 (3.83), l'algorithme d'Arrow-Hurwicz effectue un seul pas d'une méthode de gradient. 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

$$\begin{aligned} u^{n+1} &= u^n - \mu^n \left(J'(u^n) + p^n \cdot F'(u^n) \right), \\ p^{n+1} &= P_{\mathbb{R}_+^M}(p^n + \mu^n F(u^n)), \end{aligned} \tag{3.92}$$

où $\mu^n > 0$ est un pas de descente positif. Autrement dit, (3.92) recherche un point selle en alternant un pas de minimisation en v et un pas de maximisation en q.

Théorème 3.4.5 *On suppose que J est α-convexe différentiable, que F est convexe différentiable de dérivée bornée par $C_F > 0$. On suppose aussi que, pour tout $p \in (\mathbb{R}_+)^M$, la fonction $v \mapsto J'(v) + p \cdot F'(v)$ est L-Lipschitzienne sur V. Alors il existe un point-selle (u, p) du Lagrangien (3.80) sur $V \times (\mathbb{R}_+)^M$ et u est la solution unique du problème (3.79). Si le pas de descente vérifie, pour $C^2 = \max(2L^2 + C_F^2, 2C_F^2)$,*

$$\mu^n = \frac{\alpha\gamma}{C^2} \frac{\|u^n - u\|^2}{\|u^n - u\|^2 + \|p^n - p\|^2} \qquad \textit{avec} \;\; 0 < \gamma < 1,$$

alors, l'algorithme d'Arrow-Hurwicz converge : quel que soit les éléments initiaux p^0, u^0, la suite (u^n) définie par (3.92) converge vers la solution u du problème (3.79).

Remarque 3.4.6 Bien que le Théorème 3.4.5 prescrive une valeur variable de μ^n, en pratique on utilise une valeur fixe $0 < \mu < \alpha/C^2$ car on ne connait pas le point selle (u, p) et on ne peut donc pas calculer μ^n.

Démonstration. Les hypothèses impliquent qu'il existe un point-selle (u, p) du Lagrangien (3.80) sur $V \times (\mathbb{R}_+)^M$. Rappelons que d'après (3.82) on a $p = P_{\mathbb{R}_+^M}(p + \mu F(u))$ pour tout réel $\mu > 0$. Dans ce qui suit on écrit μ au lieu de μ^n par souci de simplification. Comme la projection P_K est faiblement contractante d'après (8.2), on déduit de la deuxième équation de (3.92) que

$$\begin{aligned} \|p^{n+1} - p\|^2 &\leq \|p^n + \mu F(u^n) - (p + \mu F(u))\|^2 \\ &\leq \|p^n - p\|^2 + 2\mu(p^n - p) \cdot (F(u^n) - F(u)) + \mu^2 \|F(u^n) - F(u)\|^2 \\ &\leq \|p^n - p\|^2 + 2\mu(p^n - p) \cdot F(u^n) + \mu^2 C_F^2 \|u^n - u\|^2 \end{aligned} \tag{3.93}$$

car $p^n \cdot F(u) \leq p \cdot F(u) = 0$ et la dérivée de F est supposée bornée par C_F. La première équation de (3.92) conduit à

$$\|u^{n+1} - u\|^2 = \|u^n - u\|^2 - 2\mu \langle \frac{\partial \mathcal{L}}{\partial v}(u^n, p^n), (u^n - u) \rangle + \mu^2 \|\frac{\partial \mathcal{L}}{\partial v}(u^n, p^n)\|^2. \tag{3.94}$$

Or, comme $J'(u) + p \cdot F'(u) = 0$, on a

$$\frac{\partial \mathcal{L}}{\partial v}(u^n, p^n) = J'(u^n) + p \cdot F'(u^n) - (J'(u) + p \cdot F'(u)) + (p^n - p) \cdot F'(u^n),$$

donc

$$\begin{aligned} \|\frac{\partial \mathcal{L}}{\partial v}(u^n, p^n)\|^2 &\leq 2 \left(\|J'(u^n) + p \cdot F'(u^n) - (J'(u) + p \cdot F'(u))\|^2 + \|(p^n - p) \cdot F'(u^n)\|^2 \right) \\ &\leq 2 \left(L^2 \|u^n - u\|^2 + C_F^2 \|p^n - p\|^2 \right), \end{aligned}$$

car $J'(v)+p\cdot F'(v)$ est L-Lipschitzienne et F' est borné. Par conséquent, en sommant (3.93) et (3.94) et en appliquant le Lemme 3.4.7 pour $v = u^n$ et $q = p^n$, on trouve

$$\begin{aligned}\|u^{n+1}-u\|^2+\|p^{n+1}-p\|^2 &\leq \|u^n-u\|^2+\|p^n-p\|^2-\mu\alpha\|u-u^n\|^2\\ &\quad+\mu^2C^2\left(\|u^n-u\|^2+\|p^n-p\|^2\right)\\ &\leq \|u^n-u\|^2+\|p^n-p\|^2-\frac{\gamma(1-\gamma)\alpha^2\|u^n-u\|^4}{C^2(\|u^n-u\|^2+\|p^n-p\|^2)}\end{aligned}$$

avec $C^2=\max(2L^2+C_F^2,2C_F^2)$ et en ayant remplacé μ par sa valeur μ^n donnée dans l'énoncé. On en déduit que la suite $\|u^n-u\|^2+\|p^n-p\|^2$ est décroissante positive donc convergente, et par ailleurs que le rapport $\|u^n-u\|^4/\left(\|u^n-u\|^2+\|p^n-p\|^2\right)$ tend vers zéro. Si la suite $\|u^n-u\|^2+\|p^n-p\|^2$ tend vers zéro, alors la suite (u^n,p^n) converge vers le point selle (u,p). Sinon, alors c'est la suite $\|u^n-u\|^4$ qui tend vers zéro, c'est-à-dire que la suite u^n converge vers u (mais on ne peut rien dire pour la suite p^n). □

Lemme 3.4.7 *Soit le Lagrangien $\mathcal{L}(v,q)=J(v)+q\cdot F(v)$, défini sur $V\times(\mathbb{R}_+)^M$, avec J et F convexes différentiables et J α-convexe. On suppose que (u,p) est un point selle du Lagrangien. Alors, pour tout $(v,q)\in V\times(\mathbb{R}_+)^M$,*

$$\frac{\alpha}{2}\|u-v\|^2\leq\langle\frac{\partial\mathcal{L}}{\partial v}(v,q),(v-u)\rangle-\frac{\partial\mathcal{L}}{\partial q}(v,q)\cdot(q-p)\,.$$

Démonstration. Pour $q\geq 0$ le Lagrangien est α-convexe

$$\mathcal{L}(u,q)\geq\mathcal{L}(v,q)+\langle\frac{\partial\mathcal{L}}{\partial v}(v,q),(u-v)\rangle+\frac{\alpha}{2}\|u-v\|^2\,.$$

Comme (u,p) est un point selle, $\mathcal{L}(u,q)\leq\mathcal{L}(u,p)$ et on en déduit

$$\frac{\alpha}{2}\|u-v\|^2\leq\langle\frac{\partial\mathcal{L}}{\partial v}(v,q),(v-u)\rangle+\mathcal{L}(u,p)-\mathcal{L}(v,q)\,. \tag{3.95}$$

Or $\mathcal{L}(v,q)=\mathcal{L}(v,p)+(q-p)\cdot F(v)$ et $\frac{\partial\mathcal{L}}{\partial q}(v,q)=F(v)$. La deuxième inégalité du point selle, $\mathcal{L}(u,p)\leq\mathcal{L}(v,p)$ permet enfin de conclure à partir de (3.95). □

3.4.3 Pénalisation des contraintes

Une autre méthode pour approcher un problème de minimisation avec contraintes par une suite de problèmes de minimisation sans contraintes est la procédure de **pénalisation** des contraintes. Il ne s'agit pas d'un algorithme précis mais plutôt d'une stratégie de résolution qui doit, pour sa partie numérique, utiliser l'un des algorithmes de la Sous-section 3.1 pour l'optimisation sans contrainte. Cette résolution numérique peut d'ailleurs soulever des difficultés, car le problème "pénalisé" est souvent "mal conditionné". Néanmoins l'avantage de cette méthode de pénalisation est sa facilité de mise en oeuvre, même si elle est parfois décevante pour la qualité de ses résultats numériques.

Nous nous plaçons pour simplifier dans le cas où $V = \mathbb{R}^N$, et nous considérons de nouveau le problème de minimisation convexe

$$\inf_{F(v)\leq 0} J(v) \ , \tag{3.96}$$

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 $\varepsilon > 0$, on introduit le problème sans contraintes

$$\inf_{v\in\mathbb{R}^N} \left(J(v) + \frac{1}{\varepsilon} \sum_{i=1}^{M} \left[\max\left(F_i(v), 0\right)\right]^2 \right) \ , \tag{3.97}$$

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 (3.97) "approche bien" le problème (3.96).

Proposition 3.4.8 *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 (3.96) et, pour $\varepsilon > 0$, u_ε l'unique solution de (3.97), 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 $G(v) = \sum_{i=1}^{M} \left[\max\left(F_i(v), 0\right)\right]^2$ 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

$$J_\varepsilon(u_\varepsilon) = J(u_\varepsilon) + \frac{G(u_\varepsilon)}{\varepsilon} \leq J_\varepsilon(u) = J(u) \ . \tag{3.98}$$

Ceci montre que

$$J(u_\varepsilon) \leq J_\varepsilon(u_\varepsilon) \leq J(u) \ , \tag{3.99}$$

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 (3.98). Passant à la limite, on obtient $G(u_*) = 0$, qui montre que $u_* \in K$. Comme (3.99) 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 3.4.4 En plus des hypothèses de la Proposition 3.4.8, 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 2.5.16. 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 $i\in\{1,\ldots,M\}$

$$\lim_{\varepsilon\to 0}\left[\frac{2}{\varepsilon}\max\left(F_i(u_\varepsilon),0\right)\right]=\lambda_i\ .$$

En pratique, même si la Proposition 3.4.8 garantit que la solution u_ε est une bonne approximation de la solution exacte u pour ε suffisamment petit, on ne choisit pas des paramètres ε trop petits à cause d'un problème de conditionnement numérique. En effet, lorsqu'on évalue la fonction J_ε ou ses dérivées on additionne des termes d'ordre 1 par rapport à ε (qui viennent de J) et des termes d'ordre $1/\varepsilon$ (qui viennent de F), sans parler de la division de $\max(F_i(x),0)$ (qui peut être petit à convergence) par ε. Les erreurs d'arrondi peuvent être donc très marquées dans de tels calculs. Pour cette raison on a souvent recours à une **méthode de continuation** qui consiste à diminuer progressivement la valeur de ϵ. Ce problème de conditionnement numérique est en quelque sorte la face cachée de la facilité de mise en oeuvre pratique.

Remarque 3.4.9 L'approche proposée ci-dessus est une méthode de **pénalisation extérieure**, au sens où la solution pénalisée u_ε ne vérifie en général pas les contraintes. C'est uniquement sa limite qui les vérifie. A l'opposé de cette approche, il existe une technique de **pénalisation intérieure**, pour des contraintes d'inégalité, qui garantit à chaque itération que la solution pénalisée u_ε vérifie bien les contraintes, pour peu que l'initialisation les vérifie. Expliquons en le principe pour l'exemple du problème (3.96). Pour $\varepsilon>0$, nous introduisons le problème sans contraintes

$$\inf_{v\in\mathbb{R}^N, F(v)<0}\left(J(v)-\varepsilon\sum_{i=1}^M\frac{1}{F_i(v)}\right) \tag{3.100}$$

où le deuxième terme est appelée une fonction "barrière" dont le choix n'est pas unique. De manière générale, une fonction barrière est une fonction définie sur $\mathbb{R}_-$ qui tend vers $+\infty$ quand son argument tend vers zéro. Le but de cette barrière est d'empêcher une suite de solutions approchées de (3.100) de s'approcher de la limite $F_i(v)=0$. Ainsi les contraintes $F(v)<0$ ne sont jamais actives et peuvent être ignorées en pratique. Ainsi, si on utilise un algorithme de gradient pour le problème sans contrainte (3.100) avec une initialisation qui vérifie strictement les contraintes et un pas de descente suffisamment petit, alors la suite des itérées va rester dans le domaine strictement faisable $F(v)<0$. Pour que l'effet de la barrière soit de plus en plus négligeable, on fait progressivement tendre le paramètre $\varepsilon>0$ vers zéro (notez la différence fondamentale avec le coefficient $1/\varepsilon$ en pénalisation extérieure). Nous étudierons un exemple de pénalisation intérieure à la Sous-section 4.2.3. •

Exercice 3.4.5 Soit des fonctions J et $F_1,\ldots,F_M$ convexes. Vérifiez que la fonction minimisée dans (3.100) est bien convexe sur le domaine défini par $F(v)<0$.

3.4.4 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

$$\inf_{F(v)=0} J(v) \ , \tag{3.101}$$

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. On remarque que le problème (3.101) est équivalent à

$$\inf_{F(v)=0} \left\{ J(v) + \frac{\mu}{2}\|F(v)\|^2 \right\},$$

pour tout paramètre $\mu \in \mathbb{R}$, puisque $F(v) = 0$. Cela suggère d'introduire, pour $v \in \mathbb{R}^N$, $\lambda \in \mathbb{R}^M$ et $\mu > 0$, le **Lagrangien augmenté**

$$\mathcal{L}_{aug}(v, \lambda, \mu) = J(v) + \lambda \cdot F(v) + \frac{\mu}{2}\|F(v)\|^2, \tag{3.102}$$

où $\|F(v)\|$ est la norme Euclidienne dans $\mathbb{R}^M$ du vecteur $F(v)$. Autrement dit, on additionne une fonction quadratique de la contrainte au Lagrangien habituel (3.80). Dans (3.102), λ est bien sûr un multiplicateur de Lagrange pour la contrainte, tandis que μ ressemble à 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. En réalité, si l'interprétation du Lagrangien augmenté comme une méthode de pénalisation est commode, elle n'est pas tout à fait juste et il s'agit plutôt d'une méthode de régularisation du type Moreau-Yosida (voir l'Exercice 3.2.8) pour la fonction duale [10].

Le principe de l'algorithme est de trouver un point selle de (3.102). 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 $v \mapsto \mathcal{L}_{aug}(v, \lambda^n, \mu)$. Ecrivons la condition d'optimalité pour v^n

$$J'(v^n) + \lambda^n \cdot F'(v^n) + \mu F(v^n) \cdot F'(v^n) = 0. \tag{3.103}$$

Si le même vecteur v^n était solution de (3.101), on aurait la condition d'optimalité du Théorème 2.5.6

$$J'(v^n) + \lambda^* \cdot F'(v^n) = 0, \tag{3.104}$$

où λ^* est un multiplicateur de Lagrange optimal. En comparant (3.103) et (3.104) on obtient (en supposant que les contraintes sont qualifiées en v^n)

$$\lambda^* = \lambda^n + \mu F(v^n). \tag{3.105}$$

De (3.105) on tire deux informations. D'une part, si la suite λ^n tend 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, (3.105) 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 (3.92), 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 [26]), 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 3.4.10 *On suppose que les fonctions J et F sont de classe C^2. Soit v^* un point de minimum local de* (3.101) *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* (3.101) *est vérifiée. On suppose que la condition suffisante d'optimalité d'ordre 2 est satisfaite en v^*, à savoir*

$$\left(J''(v^*) + \lambda^* \cdot F''(v^*)\right)(w, w) > 0 \quad \forall\, w \in K(v^*) = \mathrm{Ker} F'(v^*), w \neq 0. \tag{3.106}$$

Alors, il existe $\mu_0 > 0$ tel que, pour tout $\mu \geq \mu_0$, v^* est un point de minimum local de $v \mapsto \mathcal{L}_{aug}(v, \lambda^*, \mu)$ (sans contrainte).

Remarque 3.4.11 L'intérêt du Lemme 3.4.10 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 (3.101). 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 (3.101) 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 $\mathcal{L}'_{aug}(v^*, \lambda^*, \mu) = 0$ (ici et dans toute la suite de cette preuve le signe $'$ indique une dérivée en v). Calculons la dérivée seconde en v du Lagrangien augmenté

$$\mathcal{L}''_{aug}(v^*, \lambda^*, \mu)(w, w) = \mathcal{L}''(v^*, \lambda^*)(w, w) + \mu \|F'(v^*)w\|^2, \tag{3.107}$$

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 (3.106) montre que le premier terme à droite dans (3.107) 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 finie $\mathcal{L}''_{aug}(v^*, \lambda^*, \mu)(w, w)$ 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. □

Exercice 3.4.6 Appliquer l'algorithme du Lagrangien augmenté au problème (3.91) (fonctionnelle quadratique et contraintes affines d'égalité). Pour étudier la vitesse de convergence (comme dans l'Exercice 3.4.3) on introduit les valeur propres $\nu_1 \leq \nu_2 \leq \cdots \leq \nu_N$ et les vecteurs propres $w_i \in \mathbb{R}^N$ de

$$B^*Bw_i = \nu_i Aw_i.$$

On suppose que A est symétrique définie positive de taille $N \times N$ et que B, de taille $M \times N$, est de rang $M < N$. Montrer que $\nu_1 = \cdots = \nu_{N-M} = 0$ et que $\nu_{N-M+1} > 0$. En utilisant le fait que $\mathbb{R}^N = \mathrm{Ker}B \oplus \mathrm{Im}B^*$, montrer que la vitesse de convergence de l'algorithme du Lagrangien augmenté est

$$\|u^n - u\| + \|p^n - p\| \leq \frac{C}{(1 + \mu\nu_{N-M+1})^n}.$$

Comparer avec le résultat de l'Exercice 3.4.3.

3.5 Méthodes d'approximations successives

Considérons un problème général d'optimisation sous contraintes d'égalité

$$\inf_{F(v)=0} J(v), \tag{3.108}$$

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 proposés dans ce chapitre est trop compliquée ou coûteuse pour (3.108), une stratégie courante consiste à remplacer ce dernier par une succession de problèmes approchés, obtenus, par exemple, par développement de Taylor des fonctions J et F autour de la solution approchée

précédente. L'idée sous-jacente est qu'il est plus facile de résoudre le problème approché que le problème exact. Ces approximations n'ayant en général 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.

Une méthode d'approximation successive peut donc se présenter comme suit. Etant donnée une initialisation $v^0 \in V = \mathbb{R}^N$, on construit une suite de solutions approchées $v^n \in V$, $n \geq 1$, où v^n est le point de minimum du problème

$$\inf_{F_{n-1}(v)=0} J_{n-1}(v), \tag{3.109}$$

où F_{n-1} et J_{n-1} sont des approximations de F et J, respectivement, au voisinage de la solution précédente v^{n-1}. Evidemment, il faut choisir F_{n-1} et J_{n-1}, non seulement pour que (3.109) admette une solution unique v^n, mais aussi pour que le calcul de v^n soit facile.

Expliquons le principe de cette méthode dans un cas très simple, en l'absence de contraintes d'égalité, c'est-à-dire $F \equiv 0$. On choisit $F_{n-1} \equiv 0$ et

$$J_{n-1}(v) = J(v^{n-1}) + J'(v^{n-1}) \cdot (v - v^{n-1}) + \frac{1}{2\mu}\|v - v^{n-1}\|^2, \tag{3.110}$$

avec un paramètre $\mu > 0$. Notons que les fonctions J et J_{n-1} sont tangentes en v^{n-1} et donc que J_{n-1} est bien une approximation de J en ce point.

Lemme 3.5.1 *Soit J une fonction différentiable de V dans $\mathbb{R}$ et $F \equiv 0$. Alors la méthode des approximations successives avec le choix* (3.110) *n'est rien d'autre que l'algorithme du gradient à pas fixe μ.*

Démonstration. Tout d'abord, la formule (3.110) est une approximation de Taylor au premier ordre de la fonction objectif J au point v^{n-1} à laquelle on a ajouté un terme quadratique de manière à ce que J_{n-1} soit une fonction fortement convexe qui admet donc un unique point de minimum v^n. On écrit alors la condition d'optimalité nécessaire et suffisante, $J'_{n-1}(v^n) = 0$, qui donne

$$J'(v^{n-1}) + \frac{1}{\mu}(v^n - v^{n-1}) = 0,$$

ce qui est précisément la formule (3.8) de l'algorithme du gradient à pas fixe. □

Evidemment le Lemme 3.5.1 n'apporte rien de concret à l'algorithme du gradient à pas fixe mais il donne un point de vue nouveau qui permet des généralisations dont nous présentons deux exemples dans ce qui suit.

3.5.1 Programmation linéaire séquentielle

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 $v^0 \in \mathbb{R}^N$ (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

$$\inf_{F(v^{n-1})+F'(v^{n-1})\cdot(v-v^{n-1})=0} \Big\{ J(v^{n-1}) + J'(v^{n-1}) \cdot (v - v^{n-1}) \Big\}, \tag{3.111}$$

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 (3.111) est que sa valeur minimum peut être $-\infty$ et qu'il n'y a 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 (3.111) 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

$$\|v - v^{n-1}\| \leq \delta, \tag{3.112}$$

où $\delta > 0$ est un paramètre qui définit la taille du voisinage de v^{n-1} dans lequel (3.111) est une bonne approximation de (3.108). La norme dans (3.112) 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.

Remarque 3.5.2 Une variante de la programmation linéaire séquentielle est l'algorithme de Frank-Wolfe qui s'applique au problème suivant

$$\inf_{v\in K} J(v),$$

où K est un ensemble convexe. Etant donné une initialisation $v^0 \in K$, on calcule une suite de solutions approchées v^n, $n \geq 1$, en deux étapes : tout d'abord, on calcule la solution $\tilde{v}^n$ de

$$\inf_{v\in K} \Big\{ J(v^{n-1}) + J'(v^{n-1}) \cdot (v - v^{n-1}) \Big\},$$

ce qui est facile si, par exemple, K est un polyèdre. Puis, pour un pas $\mu^n \in (0, 1)$, on définit v^n par

$$v^n = v^{n-1} - \mu^{n-1}(v^{n-1} - \tilde{v}^n).$$

Par convexité on vérifie que $v^n \in K$ et, d'autre part, que $(v^{n-1} - \tilde{v}^n)$ est bien une direction de descente. Pour déterminer le pas, on peut chercher le pas optimal ou bien suivre la règle d'Armijo (voir la Sous-section 3.1.3) ou encore choisir la valeur $\mu^n = 2/(n+2)$ dont on peut montrer qu'elle garantit la convergence de l'algorithme sous des hypothèses adéquates. •

Remarque 3.5.3 L'idée de linéariser le problème d'optimisation peut se généraliser en choisissant de linéariser par rapport aux variables inverses. Evidemment, l'intérêt d'une telle approche dépend du contexte mais elle a trouvé un certain succès en optimisation de forme pour les structures mécaniques. Nous décrivons ici une variante

simple de l'algorithme MMA (méthode des asymptotes mobiles, due à Svanberg, qui généralise l'algorithme CONLIN de Fleury) qui s'applique aux problèmes du type

$$\inf_{\substack{F(v)\leq 0\\ v^{\min}\leq v\leq v^{\max}}} J(v),$$

où les contraintes d'inégalité s'appliquent composantes par composante, $v_i^{\min} \leq v_i \leq v_i^{\max}$ pour tout $1 \leq i \leq N$. Etant donné une initialisation v^0, qui vérifie $v^{\min} \leq v^0 \leq v^{\max}$, on calcule une suite de solutions approchées v^n, $n \geq 1$, définies comme solutions de

$$\inf_{\substack{F_n(v)\leq 0\\ v^{\min}\leq v\leq v^{\max}}} \left\{ J_n(v) = J(v^{n-1}) + \sum_{i=1}^{N} \frac{M_i^{n-1}}{v_i^{\max} - v_i} + \frac{m_i^{n-1}}{v_i - v_i^{\min}} \right\}, \tag{3.113}$$

avec

$$\begin{cases} M_i^{n-1} &= \dfrac{\partial J}{\partial v_i}(v^{n-1})(v_i^{\max} - v_i^{n-1})^2 \qquad \text{si } \frac{\partial J}{\partial v_i}(v^{n-1}) > 0, \\ m_i^{n-1} &= -\dfrac{\partial J}{\partial v_i}(v^{n-1})(v_i^{n-1} - v_i^{\min})^2 \quad \text{si } \frac{\partial J}{\partial v_i}(v^{n-1}) < 0. \end{cases}$$

On vérifie sans peine que $J_n(v)$ est convexe et tangente à $J(v)$ au point v^{n-1}. La contrainte $F(v)$ est "linéarisée" de la même manière en $F_n(v)$. On peut aussi faire dépendre les valeurs des "asymptotes" $v^{\min}$ et $v^{\max}$ du numéro d'itération n (d'où l'appellation d'asymptotes mobiles). La résolution de (3.113) est facile car il s'agit d'un problème convexe, même si J et F ne l'étaient pas à l'origine. •

3.5.2 Programmation quadratique séquentielle

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 $v^0 \in \mathbb{R}^N$ (ne vérifiant pas nécessairement la contrainte $F(v) = 0$) et $\lambda^0 \in \mathbb{R}^M$, on calcule une suite de solutions approchées v^n, $n \geq 1$, définies comme les solutions de

$$\inf_{F(v^{n-1})+F'(v^{n-1})\cdot(v-v^{n-1})=0} \left\{ J(v^{n-1}) + J'(v^{n-1})\cdot(v-v^{n-1}) + \frac{1}{2} Q^{n-1}(v-v^{n-1})\cdot(v-v^{n-1}) \right\}, \tag{3.114}$$

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 (3.114) (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 $F(v) = 0$.

Donnons une intuition de la raison de présence de la Hessienne du Lagrangien Q^{n-1} au lieu de $J''(v^{n-1})$ dans (3.114). Soit v^* un point de minimum de (3.108) qui vérifie les conditions d'optimalité $F(v^*) = 0$ et $J'(v^*) + \lambda^* \cdot F'(v^*) = 0$, avec un multiplicateur de Lagrange λ^*. Un développement de Taylor au voisinage de v^* conduit à

$$J(v) \approx J(v^*) + J'(v^*) \cdot (v - v^*) + \frac{1}{2} J''(v^*)(v - v^*) \cdot (v - v^*) \tag{3.115}$$

et

$$F(v) \approx F(v^*) + F'(v^*) \cdot (v - v^*) + \frac{1}{2} F''(v^*)(v - v^*) \cdot (v - v^*). \tag{3.116}$$

On se restreint aux vecteurs v qui vérifient la contrainte $F(v) = 0$, on multiplie (3.116) par le multiplicateur de Lagrange λ^*, puis on somme le résultat à (3.115) pour obtenir, en tenant compte de la condition d'optimalité,

$$J(v) \approx J(v^*) + \frac{1}{2}\Big(J''(v^*) + \lambda^* \cdot F''(v^*)\Big)(v - v^*) \cdot (v - v^*),$$

qui donne bien le même comportement quadratique que la fonction objectif de (3.114). D'ailleurs, la Proposition 2.5.14 (condition d'optimalité du 2ème ordre) montre que c'est la matrice Hessienne du Lagrangien qui est positive au point de minimum de (3.108), et pas la Hessienne de J. Ainsi, lorsque v^{n-1} est proche de v^* et λ^{n-1} de λ^*, on peut espérer que la matrice Q^{n-1} soit positive et donc que (3.114) admette au moins une solution optimale v^n. Néanmoins, si Q^{n-1} n'est pas positive, il peut être nécessaire de recourir à nouveau à une contrainte de région de confiance du type de (3.112). Pour plus de détails nous renvoyons à [26].

Remarque 3.5.4 Une autre manière d'arriver à (3.114), qui caractérise la méthode SQP, est d'appliquer l'algorithme de Newton aux conditions d'optimalité de (3.108)

$$G(v^*, \lambda^*) = \begin{pmatrix} J'(v^*) + \lambda^* \cdot F'(v^*) \\ F(v^*) \end{pmatrix} = 0,$$

dont la matrice Jacobienne est

$$G'(v^*, \lambda^*) = \begin{pmatrix} J''(v^*) + \lambda^* \cdot F''(v^*) & F'(v^*) \\ F'(v^*) & 0 \end{pmatrix}, \tag{3.117}$$

qui fait apparaitre explicitement la Hessienne du Lagrangien. L'algorithme de Newton s'écrit alors, pour $n \geq 1$ et avec une initialisation $(v^0, \lambda^0) \in \mathbb{R}^N \times \mathbb{R}^M$,

$$\begin{pmatrix} v^n \\ \lambda^n \end{pmatrix} = \begin{pmatrix} v^{n-1} \\ \lambda^{n-1} \end{pmatrix} - \big(G'(v^{n-1}, \lambda^{n-1})\big)^{-1} G(v^{n-1}, \lambda^{n-1}).$$

Si on multiplie cette équation par la matrice G' on obtient

$$\begin{cases} Q^{n-1}(v^n - v^{n-1}) + J'(v^{n-1}) + \lambda^n \cdot F'(v^{n-1}) & = 0, \\ F'(v^{n-1}) \cdot (v^n - v^{n-1}) + F(v^{n-1}) & = 0, \end{cases}$$

qui est exactement la condition d'optimalité pour que v^n soit le point de minimum de (3.114) avec le multiplicateur de Lagrange λ^n. •

Exercice 3.5.1 Montrer que la matrice $G'(v^*, \lambda^*)$, définie par (3.117) et de taille $(N+M) \times (N+M)$, est inversible si la matrice $F'(v^*)$, de taille $M \times N$, est de rang M et $J''(v^*) + \lambda^* \cdot F''(v^*)$ est une matrice $N \times N$ définie positive.

3.6 Modélisation, structures et algorithmes spécifiques

Nous terminons ce chapitre en évoquant quelques cas particuliers où se mêlent des questions de modélisation et d'algorithmique. Nous n'avons presque rien dit de la modélisation et pourtant il s'agit d'un aspect pratique très important en optimisation. Un problème d'optimisation n'est pas toujours présenté sous une forme mathématique claire et le praticien doit lui-même mettre en forme la question d'optimisation à laquelle il doit répondre. Comme toujours en modélisation, il s'agit de faire un compromis entre un modèle très précis et réaliste, mais intractable du point de vue algorithmique, et un modèle simplifié mais facilement optimisable en pratique. Dans cette étape de modélisation il est essentiel de faire émerger une structure du problème qui permette une mise en oeuvre efficace des algorithmes d'optimisation. Nous avons déjà vu des exemples de structures particulières qui autorisent des algorithmes efficaces (par exemple, un problème linéaire se traite avec l'algorithme du simplexe, voir le Chapitre 4) et nous allons voir ici trois cas particuliers où l'on peut concevoir des méthodes qui tirent parti de la structure du problème pour résoudre efficacement leur optimisation. Bien sûr, nous ne sommes pas exhaustifs et nous ne voulons pas non plus donner l'impression de faire une liste de "recettes" : notre but est juste de montrer qu'il faut parfois un peu de réflexion avant de choisir tel ou tel algorithme plutôt qu'un autre. Dans le même ordre d'idées nous favorisons l'explication des idées à leur traitement mathématique rigoureux et complet.

3.6.1 Fonctions composées et rétro-propagation du gradient

On considère un problème d'optimisation sans contraintes où la fonction objectif J est la composée de m fonctions J_i

$$\inf_{v \in \mathbb{R}^N} J(v) = J_m \circ J_{m-1} \circ \cdots \circ J_1(v), \tag{3.118}$$

où chaque fonction $J_i(v_i)$, $1 \leq i \leq m$, est définie de $\mathbb{R}^{N_i}$ dans $\mathbb{R}^{N_{i+1}}$, avec $N_1 = N$ et $N_{m+1} = 1$ de sorte que J est une fonction de $\mathbb{R}^N$ dans $\mathbb{R}$. On suppose que chacune des fonctions J_i est différentiable. On se propose d'appliquer l'algorithme de gradient à pas fixe pour minimiser (3.118). Il faut donc calculer le gradient de la fonction composée J de façon efficace d'un point de vue numérique.

Cette structure particulière de fonction composée apparait dans au moins deux classes de problèmes importants en pratique. En premier lieu c'est une situation qu'on rencontre en apprentissage machine ou en intelligence artificielle lorsqu'on utilise des réseaux de neurones profonds. Chaque couche du réseau de neurones correspond à une fonction J_i. Dans la phase d'entrainement du réseau on résout

le problème (3.118) où l'on optimise les coefficients des fonctions de transfert. Par souci de simplicité, on a supposé dans (3.118) que seuls les coefficients de la première couche sont optimisables. Bien entendu, on peut facilement généraliser à l'optimisation des coefficients de toutes les couches du réseau mais les notations deviennent plus lourdes et cachent la simplicité de l'idée que nous allons présenter. En second lieu, la fonction composée $J(v)$ est un modèle d'un programme informatique qui permet de calculer une quantité d'intérêt à l'aide de m sous-programmes qui s'enchainent comme la composition des m fonctions J_i. La différentiation automatique est une discipline, à cheval entre mathématiques et informatique, qui se donne pour but d'analyser le programme correspondant à $J(v)$ et de produire un autre programme qui fournit automatiquement le gradient $J'(v)$. Evidemment l'intérêt de cette approche est de pouvoir ensuite appliquer des algorithmes d'optimisation à la fonction J dont on n'a pas une expression analytique mais seulement un programme de calcul.

Rappelons le théorème de dérivation composée dans ce cas précis.

Lemme 3.6.1 *On note $v_i = J_{i-1} \circ \cdots \circ J_1(v)$ pour $1 \leq i \leq m$ (avec la convention $v_1 = v$). La fonction $J(v) = J_m \circ J_{m-1} \circ \cdots \circ J_1(v)$ est différentiable au sens de Fréchet en tout $v \in \mathbb{R}^N$ et vérifie pour tout $h \in \mathbb{R}^N$*

$$J(v+h) = J(v) + J'_m(v_m)J'_{m-1}(v_{m-1}) \cdots J'_1(v_1)h + o(h), \tag{3.119}$$

où, pour $1 \leq i \leq m$, $J'_i(v_i)$ est une matrice $N_{i+1} \times N_i$.

Remarque 3.6.2 On rappelle que, si $\mathcal{J}(v)$ est une fonction dérivable de $\mathbb{R}^N$ dans $\mathbb{R}^M$, sa différentielle est la matrice

$$\mathcal{J}'(v) = \left(\frac{\partial \mathcal{J}_i}{\partial v_j}(v)\right)_{1 \leq i \leq M,\, 1 \leq j \leq N},$$

où $\mathcal{J}_i$ sont les composantes du vecteur $\mathcal{J}$ et v_j celles de v. La formule (3.119) donne donc l'expression suivante (une suite de produits de matrices) pour la différentielle de J

$$J'(v)(h) = J'_m(v_m)J'_{m-1}(v_{m-1}) \cdots J'_1(v_1)h \quad \forall h \in \mathbb{R}^N.$$

Cette expression est déroutante et peu pratique car, comme expliqué dans la Remarque 2.4.2, on souhaite identifier la différentielle $J'(v)$ à un vecteur de $\mathbb{R}^N$. Autrement dit, on voudrait écrire (3.119) sous la forme plus usuelle

$$J(v+h) = J(v) + J'(v) \cdot h + o(h) \quad \forall h \in \mathbb{R}^N.$$

Mais la formule (3.119) conduit à un vecteur ligne $J'(v) = J'_m(v_m)J'_{m-1}(v_{m-1}) \cdots J'_1(v_1)$ (de taille $1 \times N$) et surtout son évaluation pratique est très peu économique comme nous allons le voir plus loin. •

Démonstration. Commençons par prouver le résultat dans le cas $m = 2$. Pour $v \in \mathbb{R}^N$ et $v_2 \in \mathbb{R}^{N_2}$, la différentiabilité de J_1 et J_2 s'écrit

$$\begin{cases} J_1(v+h) = J_1(v) + J'_1(v)h + o(h) & \text{pour tout } h \in \mathbb{R}^N, \\ J_2(v_2+h_2) = J_2(v_2) + J'_2(v_2)h_2 + o(h_2) & \text{pour tout } h_2 \in \mathbb{R}^{N_2}, \end{cases}$$

où $J_1'(v)$ est une matrice $N_2 \times N$ et $J_2'(v_2)$ est un vecteur ligne de $\mathbb{R}^{N_2}$ puisque J_2 est une fonction de $\mathbb{R}^{N_2}$ dans $\mathbb{R}$. En posant $v_2 = J_1(v)$, $h_2 = J_1'(v)h + o(h)$ et en composant les deux dérivées on obtient

$$J_2 \circ J_1(v+h) = J_2 \circ J_1(v) + J_2' \circ J_1(v) J_1'(v)h + o(h) \quad \text{pour tout } h \in \mathbb{R}^N.$$

Une simple récurrence permet d'obtenir le résultat pour tout m. □

Analysons d'un point de vue calculatoire la formule (3.119) et sa conclusion

$$J'(v) = J_m'(v_m) J_{m-1}'(v_{m-1}) \cdots J_1'(v_1). \tag{3.120}$$

Cette formule permet de calculer la dérivée dans le même ordre que J : on commence par calculer $J_1'(v_1)$ que l'on multiplie à $J_2'(v_2)$, puis à $J_3'(v_3)$ et ainsi de suite jusqu'à $J_m'(v_m)$. Calculons combien d'opérations cela nécessite (on ne compte que les multiplications pour faire simple). A l'étape i on multiplie la matrice $J_i'(v_i)$, de taille $N_{i+1} \times N_i$, au résultat des multiplications précédentes $J_{i-1}'(v_{i-1}) \cdots J_1'(v_1)$ qui est une matrice de taille $N_i \times N$. Au total on effectue donc

$$\sum_{i=2}^{m} N_{i+1} N_i N$$

opérations pour obtenir le vecteur ligne $J'(v) \in \mathbb{R}^N$. En fait, cet algorithme de calcul de $J'(v)$ est très inefficace car il repose sur des produits de matrices qui sont coûteux et conduisent à un facteur N très pénalisant dans le compte d'opérations ci-dessus. Pour faire mieux l'idée est plutôt de vouloir identifier $J'(v)$ à un vecteur colonne (comme expliqué dans la Remarque 3.6.2) en transposant la formule (3.120), c'est-à-dire en écrivant

$$J'(v) = J_1'(v_1)^* J_2'(v_3)^* \cdots J_{m-1}'(v_{m-1})^* J_m'(v_m)^*, \tag{3.121}$$

où l'on rappelle que l'exposant $*$ désigne la matrice adjointe ou transposée. La formule (3.121) est appelée **rétro-propagation du gradient** car les opérations se font en sens inverse de celles pour calculer J : on part du dernier indice m et par multiplication successive dans l'ordre décroissant des indices on arrive au premier indice. Le fait que cela soit les matrices adjointes des gradients qui jouent un rôle lui donne aussi le nom de formule adjointe (au moins dans le domaine de la différentiation automatique). L'avantage essentiel de (3.121) est que cette formule est beaucoup plus efficace car elle repose sur des produits matrices-vecteurs, au lieu de produits de matrices comme pour (3.120). Le point essentiel est que, comme $N_{m+1} = 1$, $J_m'(v_m)^*$ est un vecteur colonne dans $\mathbb{R}^{N_m}$. Ainsi, à chaque étape i, on multiplie à la matrice $J_i'(v_i)^*$, de taille $N_i \times N_{i+1}$ un vecteur de taille N_{i+1} qui est le résultat des multiplications précédentes $J_{i-1}'(v_{i-1})^* \cdots J_m'(v_m)^*$. Au total on effectue donc

$$\sum_{i=1}^{m-1} N_{i+1} N_i$$

opérations, ce qui est approximativement N fois moins que le compte d'opérations pour la formule (3.120). Notons tout de même un léger inconvénient pour la formule (3.121) en terme de mémoire informatique car, puisqu'en général on calcule en même temps la fonction J_i et sa dérivée J_i', il faut garder en mémoire toutes les matrices dérivées pour effectuer la descente d'indice dans (3.121). Néanmoins, en pratique cette formule de rétro-propagation du gradient s'est imposée par son faible coût de calcul.

Exercice 3.6.1 Montrer que pour calculer une seule dérivée directionnelle $J'(v)(h)$ la formule (3.119) nécessite $\sum_{i=1}^{m} N_i N_{i+1}$ multiplications, c'est-à-dire autant que pour calculer le gradient **complet** $J'(v)$ avec la formule (3.121).

3.6.2 Optimisation sous contrainte de modèle et état adjoint

On considère maintenant une classe de problèmes d'optimisation où la variable d'optimisation est une combinaison de deux variables : $v \in \mathbb{R}^N$ la variable des paramètres de conception ou de contrôle d'un système et $y \in V$ la variable d'état décrivant ce système, avec un espace de Hilbert V. En fait, v et y sont reliés par une contrainte d'égalité, $F(v, y) = 0$, qui s'interprète en disant que l'état y du système est déterminé par les paramètres v. Cette contrainte d'égalité est typiquement un modèle qui permet de calculer y en fonction de v. Par exemple, y est la solution d'une équation différentielle où v est un paramètre (un coefficient, un terme source, etc.). Nous rencontrerons cette situation en théorie du contrôle optimal. En tout état de cause, il faut comprendre que la résolution du modèle pour trouver y en fonction de v est une opération non explicite et coûteuse.

Etant donné une fonction objectif $J(v, y)$, différentiable de $\mathbb{R}^N \times V$ dans $\mathbb{R}$, et une fonction contrainte $F(v, y)$, différentiable de $\mathbb{R}^N \times V$ dans V, on considère le problème d'optimisation

$$\inf_{v \in \mathbb{R}^N,\, y \in V \text{ tel que } F(v,y)=0} J(v, y). \tag{3.122}$$

Bien sûr, il est possible d'avoir d'autres contraintes mais nous nous concentrons ici sur la contrainte de modèle qui relie y à v. Nous faisons l'hypothèse suivante sur ce modèle.

Hypothèse (H). *Pour tout $v \in \mathbb{R}^N$, il existe une unique solution $y(v) \in V$ de la contrainte $F(v, y) = 0$, cette solution $v \mapsto y(v)$ est différentiable de $\mathbb{R}^N$ dans V et l'opérateur linéarisé $z \mapsto \frac{\partial F}{\partial y}(v, y(v))(z)$ est continu, inversible de V dans V et d'inverse continu de V dans V.*

Sous cette hypothèse, le problème (3.122) est équivalent au problème sans contrainte

$$\inf_{v \in \mathbb{R}^N} \widetilde{J}(v) = J(v, y(v)). \tag{3.123}$$

Il est conceptuellement facile de mettre en oeuvre un algorithme de gradient pour résoudre (3.123) mais nous allons voir que c'est une approche très inefficace dès que la dimension N est grande.

Lemme 3.6.3 *Sous l'hypothèse (H) la fonction $\widetilde{J}$ est différentiable sur $\mathbb{R}^N$ et, pour tout $h \in \mathbb{R}^N$, on a*

$$\widetilde{J}'(v) \cdot h = \frac{\partial J}{\partial v}(v, y(v)) \cdot h + \left\langle \frac{\partial J}{\partial y}(v, y(v)) \, , \, y'(v) \cdot h \right\rangle ,$$

où $\langle , \rangle$ désigne le produit scalaire de V et $y'(v) \cdot h$ est la dérivée de $v \mapsto y(v)$ dans la direction h qui vérifie

$$\frac{\partial F}{\partial y}(v, y(v))(y'(v) \cdot h) = -\frac{\partial F}{\partial v}(v, y(v)) \cdot h. \tag{3.124}$$

Démonstration. C'est une simple application du théorème de dérivation composée où l'on doit juste noter que $y'(v) \cdot h$ et $\frac{\partial F}{\partial v}(v, y(v)) \cdot h$ appartiennent à V. Le système linéarisé (3.124) admet alors une solution unique $y'(v) \cdot h$ grâce à l'hypothèse (H). □

Remarque 3.6.4 Pour bien comprendre la portée du Lemme 3.6.3 et de l'hypothèse (H), on considère l'exemple suivant : $V = \mathbb{R}^M$ et $F(v, y) = Ay - b(v)$ où A est une matrice inversible $M \times M$ et $b(v)$ est une fonction régulière de $\mathbb{R}^N$ dans $\mathbb{R}^M$. Dans ce cas $\frac{\partial F}{\partial y}(v, y(v)) = A$ et l'hypothèse (H) est vérifiée car A est inversible et (3.124) est un simple système linéaire. Cet exemple est tout à fait représentatif car il correspond, par exemple, à la discrétisation d'une équation différentielle linéaire dont la solution y dépend d'une variable de contrôle v. La difficulté avec le Lemme 3.6.3 est qu'il ne donne pas une formule pour le gradient $\widetilde{J}'(v)$ mais seulement pour la dérivée directionnelle dans la direction h. Pour obtenir le gradient il faut calculer chacune de ses composantes en résolvant (3.124) pour h successivement égal à chacun des vecteurs e_i de la base canonique de $\mathbb{R}^N$, ce qui revient à résoudre N fois ce système linéaire, opération trop coûteuse dans de nombreux cas pratiques où M et N sont grands. •

Pour dépasser la difficulté mentionnée ci-dessus et trouver une formule explicite, et économique, du gradient $\widetilde{J}'(v)$ nous allons introduire la notion **d'état adjoint** qui est en fait un multiplicateur de Lagrange pour la contrainte $F(v, y) = 0$. Pour tout $(v, y, p) \in \mathbb{R}^N \times V \times V$ on définit le Lagrangien

$$\mathcal{L}(v, y, p) = J(v, y) + \langle p, F(v, y) \rangle.$$

Le Lagrangien est affine en p et sa dérivée partielle par rapport à p n'est rien d'autre que la contrainte $F(v, y)$. Cela motive l'étude d'une autre dérivée partielle du Lagrangien, celle par rapport à y, qui vaut, pour tout $z \in V$,

$$\langle \frac{\partial \mathcal{L}}{\partial y}(v, y, p), z \rangle = \langle \frac{\partial J}{\partial y}(v, y), z \rangle + \langle p, \frac{\partial F}{\partial y}(v, y)(z) \rangle. \tag{3.125}$$

Lemme 3.6.5 *Le* **problème adjoint** *est défini par : trouver $p \in V$ solution de*

$$\left(\frac{\partial F}{\partial y}(v, y(v)) \right)^* p = -\frac{\partial J}{\partial y}(v, y(v)). \tag{3.126}$$

Sous l'hypothèse (H), ce problème admet une unique solution $p \equiv p(v) \in V$, appelé **état adjoint**. *Autrement dit, l'état adjoint est l'unique $p(v) \in V$ tel que*

$$\langle \frac{\partial \mathcal{L}}{\partial y}(v, y(v), p(v)), z\rangle = 0 \quad \text{pour tout } z \in V.$$

Démonstration. Comme $\frac{\partial F}{\partial y}(v, y(v))$ est un opérateur linéaire continu de V dans V, il admet un adjoint, noté $\left(\frac{\partial F}{\partial y}(v, y(v))\right)^*$. D'après l'hypothèse (H) il est de plus inversible d'inverse continu, et c'est vrai aussi pour son adjoint. Puisque J est différentiable, le second membre de (3.126) appartient à V et on en déduit donc l'existence et l'unicité de l'état adjoint $p(v)$ solution de (3.126). On remplace alors y par $y(v)$ et p par $p(v)$ dans (3.125) pour obtenir que la dérivée partielle du Lagrangien par rapport à y s'annule au point $(v, y(y), p(v))$. □

Proposition 3.6.6 *Sous l'hypothèse (H) la fonction $\widetilde{J}$ est différentiable sur $\mathbb{R}^N$ et on a*

$$\widetilde{J}'(v) = \frac{\partial J}{\partial v}(v, y(v)) + \langle p(v), \frac{\partial F}{\partial v}(v, y(v))\rangle, \tag{3.127}$$

où $p(v) \in V$ est **l'état adjoint**.

Remarque 3.6.7 La Proposition 3.6.6 est une amélioration considérable du Lemme 3.6.3 dès que le nombre de variables N est grand. En effet, la formule (3.127) donne le gradient complet de $\widetilde{J}$ au prix d'un seul calcul supplémentaire, celui de l'état adjoint $p(v)$. Au contraire, le Lemme 3.6.3 ne donnait qu'une dérivée directionnelle et il fallait résoudre N problèmes linéaires (3.124) (de même complexité que le problème adjoint) pour obtenir le gradient complet. Pour continuer sur l'exemple de la Remarque 3.6.4, où $V = \mathbb{R}^M$ et $F(v, y) = Ay - b(v)$ avec A une matrice inversible $M \times M$, le problème adjoint est un simple système linéaire à résoudre avec la matrice adjointe A^*. •

Démonstration. La contrainte étant satisfaite en $y(v)$, pour tout $p \in V$, on a

$$\widetilde{J}(v) = \mathcal{L}(v, y(v), p).$$

On dérive cette égalité par rapport à v pour obtenir

$$\widetilde{J}'(v) \cdot h = \frac{\partial \mathcal{L}}{\partial v}(v, y(v), p) \cdot h + \langle \frac{\partial \mathcal{L}}{\partial y}(v, y(v), p), (y'(v) \cdot h)\rangle.$$

On remplace alors p par l'état adjoint $p(v)$ dont la définition est précisément que $\langle \frac{\partial \mathcal{L}}{\partial y}(v, y(v), p(v)), z\rangle = 0$ pour tout $z \in V$. Par conséquent, on en déduit

$$\widetilde{J}'(v) \cdot h = \frac{\partial \mathcal{L}}{\partial v}(v, y(v), p(v)) \cdot h,$$

ce qui donne la formule (3.127). □

3.6.3 Décomposition-coordination

On considère un problème d'optimisation sous contraintes d'égalité

$$\inf_{v\in\mathbb{R}^N,\ F(v)=F_0} J(v), \tag{3.128}$$

où J est une fonction de $\mathbb{R}^N$ dans $\mathbb{R}$, F est une fonction de $\mathbb{R}^N$ dans $\mathbb{R}^M$ et $F_0 \in \mathbb{R}^M$ est un niveau de contrainte donné. On suppose que le problème est **décomposable**, c'est-à-dire qu'il existe une partition de la variable $v = (v_1, ..., v_n)$, avec $v_i \in \mathbb{R}^{n_i}$ et $\sum_{i=1}^n n_i = N$, et des fonctions objectifs et contraintes

$$J(v) = \sum_{i=1}^n J_i(v_i) \quad \text{et} \quad F(v) = \sum_{i=1}^n F_i(v_i),$$

de telle façon que le problème (3.128) est en fait équivalent à

$$\inf_{v_i\in\mathbb{R}^{n_i}} \sum_{i=1}^n J_i(v_i) \quad \text{sous la contrainte } \sum_{i=1}^N F_i(v_i) = F_0. \tag{3.129}$$

De nombreux problèmes pratiques peuvent se mettre sous cette forme et, pour fixer les idées, nous allons considérer le cas de l'optimisation du fonctionnement d'une entreprise composée de n divisions ou unités qui ont chacune leur propre variable de décision v_i et qui doivent minimiser leur coût de fonctionnement $J_i(v_i)$ avec leur contrainte de production $F_i(v_i)$, indépendants de celui des autres unités $j \neq i$, mais en partageant le niveau global de contraintes imposées, comme un budget, des ressources humaines ou des niveaux de production qui sont agrégées au niveau de l'entreprise entière.

Si on note p le multiplicateur de Lagrange associé à la contrainte $F(v) = F_0$, pour $p \in \mathbb{R}^M$ et $v \in \mathbb{R}^N$, on introduit le Lagrangien

$$\mathcal{L}(v, p) = J(v) + p \cdot (F(v) - F_0) = \sum_{i=1}^n (J_i(y_i) + p \cdot F_i(v)) - p \cdot F_0.$$

L'algorithme d'Uzawa (3.83) appliqué au problème (3.128) consiste, pour une initialisation $p_0 \in \mathbb{R}^M$, à construire les suites (u^k) et (p^k) déterminées par les itérations, pour $k \geq 0$,

$$\begin{aligned} \mathcal{L}(u^k, p^k) &= \inf_{v\in\mathbb{R}^N} \mathcal{L}(v, p^k)\,, \\ p^{k+1} &= p^k + \mu(F(u^k) - F_0)\,, \end{aligned} \tag{3.130}$$

où $\mu > 0$ est un pas positif fixé. L'hypothèse de décomposition des fonctions J et F rend en fait la minimisation du Lagrangien, à p fixé, particulièrement facile car il suffit de construire $u^k = (u_1^k, ..., u_n^k)$, où chaque composante u_i^k est la solution d'un problème de minimisation

$$J_i(u_i^k) + p^k \cdot F_i(u_i^k) = \inf_{v_i\in\mathbb{R}^{n_i}} \left(J_i(v_i) + p^k \cdot F_i(v_i)\right). \tag{3.131}$$

Chacun des problèmes (3.131) est de plus petite taille que (3.128) et indépendant des autres composantes $j \neq i$ du problème. Dans ce contexte, l'algorithme d'Uzawa est appelé algorithme de **décomposition par les prix**.

Une interprétation usuelle de cet algorithme est la suivante. Supposons, pour simplifier, que les fonctions J_i et F_i sont positives : elles correspondent au coût de production et au niveau de production, respectivement, de l'unité i. A chaque itération k, un organe central de l'entreprise fixe un prix $-p^k$ pour la valeur de la production (le vecteur p^k a toutes ses composantes négatives). Du coup, chaque unité veut maximiser son gain, c'est-à-dire minimiser son coût moins ses revenus correspondant au prix $-p^k$ de ce qu'elle produit. Autrement dit, chaque unité résout le problème (3.131) sans se soucier des autres unités. A la fin de l'iteration k, la direction de l'entreprise regarde si elle a atteint le niveau de production souhaité F_0. Si c'est le cas, la procédure a convergé. Sinon, on corrige les prix en les augmentant si le niveau de production est trop bas, $F(u^k) < F_0$, ou en les diminuant si la production est trop forte, $F(u^k) > F_0$ (on rappelle que le prix est $-p^k$).

Remarque 3.6.8 Nous ne disons rien de la convergence de cet algorithme de décomposition par les prix et nous renvoyons à des ouvrages plus spécialisés, comme [10]. Néanmoins, si les contraintes $F_i(v_i)$ sont affines et si les fonctions coûts $J_i(v_i)$ sont fortement convexes, alors le Théorème 3.4.3 affirme que l'algorithme d'Uzawa converge, c'est-à-dire que l'algorithme de décomposition par les prix converge dans ce cas particulier. On suppose ici que $F(v)$ est affine, ce qui permet de remplacer la contrainte d'égalité par deux contraintes (opposées) d'inégalité, qui sont toutes les deux convexes. •

Une autre approche de décomposition applicable au problème (3.129) est la **décomposition par les quantités**. Expliquons-en le principe en reprenant l'exemple de l'entreprise composée de n unités. Il s'agit toujours d'un algorithme itératif et à chaque itération k la direction de l'entreprise confie à chaque unité i le soin de produire une quantité $F_i^k \in \mathbb{R}^M$, en ayant pris soin de choisir ses quantités qui vérifient la contrainte globale

$$\sum_{i=1}^{n} F_i^k = F_0.$$

Chaque unité doit donc résoudre un problème d'optimisation sous contrainte

$$\inf_{v_i \in \mathbb{R}^{n_i},\ F_i(v_i)=F_i^k} J_i(v_i), \tag{3.132}$$

dont on suppose qu'il admet une solution unique $u_i^k \in \mathbb{R}^{n_i}$. On suppose aussi qu'il existe un unique multiplicateur de Lagrange $p_i^k \in \mathbb{R}^M$ pour la contrainte dans (3.132). On note $G_i(F_i^k)$ la valeur du minimum dans (3.132), qui dépend du niveau de contraintes $F_i^k \in \mathbb{R}^M$. On sait par le Lemme 2.5.13 que, si la fonction G_i est différentiable, alors on a

$$p_i^k = -\nabla G_i(F_i^k).$$

Autrement dit, la sensibilité de la valeur du minimum par rapport au niveau de contraintes est égale à l'opposé du multiplicateur de Lagrange (qui s'interprète classiquement comme l'opposé d'un prix marginal). Ainsi, si deux unités $i \neq j$ ont

des multiplicateurs de Lagrange différents $p_i^k \neq p_j^k$, alors la direction de l'entreprise peut changer l'allocation de production à ces deux unités, en $F_i^k - \mu(p_j^k - p_i^k)$ et $F_j^k + \mu(p_j^k - p_i^k)$, pour $\mu > 0$ petit, ce qui conduit à une diminution de la somme des minima dans (3.132) pour i et j qui vaut au premier ordre $-\mu|p_j^k - p_i^k|^2 < 0$. On en conclut que tant que les multiplicateurs de Lagrange des unités sont différents, l'entreprise peut améliorer son minimum en changeant l'allocation des quantités à produire. Concrètement, pour passer à l'itération $(k+1)$, la direction de l'entreprise forme le prix moyen $p^{k+1} = \sum_{i=1}^n p_i^k/n$ et propose la nouvelle allocation

$$F_i^{k+1} = F_i^k + \mu(p_i^k - p^{k+1}).$$

On vérifie aisément que la contrainte globale $\sum_{i=1}^n F_i^{k+1} = F_0$ est toujours vérifiée et que, pour $\mu > 0$ petit, le gain dans le minimum de (3.129) est égal au premier ordre à

$$-\mu \sum_{i=1}^n |p_i^k - p^{k+1}|^2.$$

On améliore donc systématiquement le minimum avec cette méthode de décomposition par les quantités qui, par ailleurs, vérifie toujours la contrainte $F(u^k) = F_0$ au cours des itérations, contrairement à la méthode de décomposition par les prix (voir [10] pour plus de détails).

Chapitre 4

PROGRAMMATION LINÉAIRE

4.1 Introduction

Ce chapitre est consacré à la **programmation linéaire** qui permet de résoudre efficacement les problèmes d'optimisation 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 le domaine de la **recherche opérationnelle** (ou plus simplement RO). La RO est l'ensemble des méthodes scientifiques qui permettent de modéliser et analyser des situations complexes afin de prendre des décisions, sinon optimales, du moins les plus efficaces possibles. C'est un domaine au carrefour des mathématiques, de l'informatique et de l'ingénierie (au sens très large) où le mot "opérationnel" fait référence à la planification des opérations militaires car la RO est née pendant la seconde guerre mondiale. Depuis elle s'est très largement civilisée et elle est employée dans de très nombreuses applications pour l'industrie et les services, en économie et sciences de la gestion

La RO ne se limite pas du tout à la programmation linéaire et, bien au contraire, utilise des outils très divers, issus de plusieurs champs scientifiques : optimisation (continue et combinatoire), probabilités (algorithmes stochastiques), théorie des jeux, mathématiques discrètes, théorie des graphes, théorie de la complexité (informatique), et programmation par contraintes. Nous n'étudierons pas tous ces aspects de la RO, pas plus que nous ne toucherons aux questions de modélisation ou à la conception d'"heuristiques", lorsque des méthodes rigoureuses font défaut, ce qui représente souvent une part importante de la pratique en RO... 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. Pour une introduction à la partie mathématisée de la RO nous renvoyons le lecteur vers l'ouvrage [6], issu d'un cours de troisième année à l'Ecole Polytechnique, ou bien vers [27].

La Section 4.2 est une présentation classique de la programmation linéaire. Au delà des (nombreux) cas où le modèle d'optimisation (fonction objectif et contraintes) est linéaire, rappelons que la programmation linéaire est aussi une brique de base dans de nombreux algorithmes d'optimisation qui linéarisent les problèmes à résoudre. La Section 4.3 est plus spécifique à la RO car elle étudie quelques exemples où l'on recherche des solutions à valeurs entières, et non pas seulement réelles, de

programmes linéaires. Lorsqu'on impose cette contrainte additionnelle (et très forte) de solutions entières, on parle alors de programmation linéaire en nombres entiers, voire d'optimisation combinatoire car il serait possible en théorie, mais illusoire en pratique, d'énumérer tous les candidats possibles pour trouver la solution optimale. Là aussi nous ne ferons qu'effleurer cette vaste question et nous renvoyons à [6] pour plus de détails.

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

$$\inf_{(x^+,x^-,\lambda)\in\mathbb{R}^{(2n+m)} \text{tel que } Ax^+-Ax^-=b+\lambda,\ x^+\geq 0, x^-\geq 0, \lambda\geq 0} c\cdot(x^+-x^-). \tag{4.4}$$

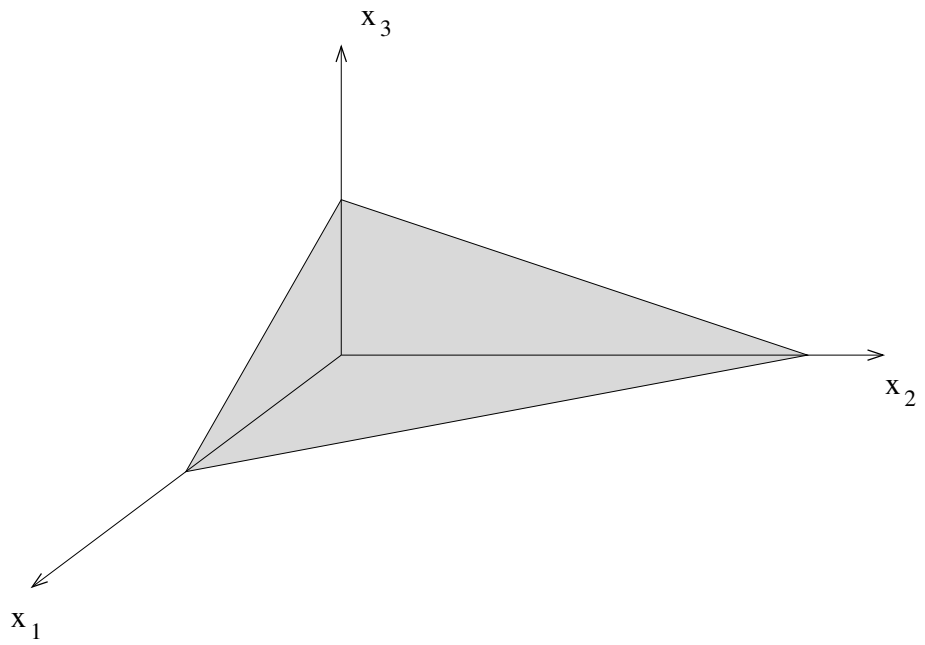

FIGURE 4.1 – Ensemble admissible pour l'exemple (4.5).

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

$$\min_{\substack{x_1\geq 0, x_2\geq 0, x_3\geq 0\\ 2x_1+x_2+3x_3=6}} x_1 + 4x_2 + 2x_3 \ . \tag{4.5}$$

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 $x_1 + 4x_2 + 2x_3 = v$ paramétrée par v. En augmentant la valeur de v 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 $(3, 0, 0)$ 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 combinaison 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, d'une part convexe, d'autre part 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*

$$-\infty < \inf_{x \in \mathbb{R}^n \text{tel que } Ax=b,\ x \geq 0} c \cdot x < +\infty.$$

De plus, une solution optimale $\overline{x}$ est caractérisée par la condition d'optimalité

$$c - A^* p \geq 0, \quad \overline{x} \geq 0, \quad \overline{x} \cdot (c - A^* p) = 0, \quad b - A\overline{x} = 0. \tag{4.6}$$

où $p \in \mathbb{R}^m$ est le multiplicateur de Lagrange pour la contrainte $b - Ax = 0$.

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 2.5.20 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 $\overline{x}$. Par ailleurs, toutes les contraintes étant affines, elles sont automatiquement qualifiées et on peut appliquer le Théorème 2.5.18 qui donne la condition d'optimalité (4.6), où l'on n'a pas écrit le multiplicateur de Lagrange $\mu \in \mathbb{R}^n_+$ pour la contrainte $-x \leq 0$. Notons que la fonction objectif et les contraintes étant convexes (puisqu'affines), on peut aussi appliquer le Théorème 2.6.4 de Kuhn et Tucker qui affirme que la condition d'optimalité (4.6) est non seulement nécessaire mais aussi suffisante. □

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 $\overline{x}$

$$c \cdot x - c \cdot \overline{x} = c_B \cdot B^{-1}(b - Nx_N) + c_N \cdot x_N - c_B \cdot B^{-1}b = (c_N - N^*(B^{-1})^* c_B) \cdot x_N. \quad (4.7)$$

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*

$$\tilde{c}_N = c_N - N^*(B^{-1})^* c_B \geq 0. \quad (4.8)$$

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 (4.7) 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 (4.8) 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 (4.8) est nécessaire. □

Remarque 4.2.9 Dans le cadre de la Proposition 4.2.8, si la solution basique considérée est dégénérée, la condition (4.8) 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, $-\mathrm{Id}_m$ 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

$$\inf_{\substack{x\geq 0,\ y\geq 0\\ Ax+y=b}} k \cdot y \,, \tag{4.9}$$

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 (4.9) 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 (4.9), 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 (4.9) 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

$$B^{k+1} = B^k E^k \quad \text{avec } E^k = \begin{pmatrix} 1 & & & l_1 & & & \\ & \ddots & & \vdots & & 0 & \\ & & 1 & \vdots & & & \\ & & & l_j & & & \\ & & & \vdots & 1 & & \\ & 0 & & \vdots & & \ddots & \\ & & & l_n & & & 1 \end{pmatrix},$$

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

$$\begin{cases} -3x_1 + 2x_2 + x_3 = 2, \\ -x_1 + 2x_2 + x_4 = 4, \\ x_1 + x_2 + x_5 = 5. \end{cases}$$

Indication : on choisira une initialisation astucieuse.

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

$$\inf_{x\in\mathbb{R}^n \text{tel que } Ax=b,\ x\geq 0} c\cdot x. \tag{4.10}$$

On définit un potentiel logarithmique pour $x > 0$

$$\pi(x) = -\sum_{i=1}^{n} \log x_i. \tag{4.11}$$

Pour un paramètre de pénalisation $\mu > 0$, on introduit le problème strictement convexe

$$\min_{x\in\mathbb{R}^n \text{tel que } Ax=b,\ x>0} \mu\pi(x) + c\cdot x. \tag{4.12}$$

Remarquons qu'en pratique la contrainte $x > 0$ n'en est pas une car elle n'est jamais active : quand on minimise (4.12) on ne peut pas "s'approcher" du bord de $x > 0$ sous peine de faire "exploser" le potentiel $\pi(x)$ vers $+\infty$. Il s'agit donc d'un exemple de méthode de **pénalisation intérieure**.

Le principe de l'algorithme de trajectoire centrale est de minimiser (4.12) 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 (4.12) tend vers le programme linéaire (4.10).

Lemme 4.2.11 *On définit l'ensemble admissible de (4.12) par*

$$X_{ad}^0 = \{x \in \mathbb{R}^n \text{ tel que } Ax = b,\ x > 0\}.$$

On suppose que X_{ad}^0 est borné non vide. Alors le problème pénalisé (4.12) admet une unique solution optimale x^μ. De plus, si (4.10) admet une unique solution optimale x^0, alors x^μ converge vers x^0 lorsque μ tend vers zéro.

Remarque 4.2.12 L'hypothèse que X_{ad}^0 est non vide est nécessaire pour que (4.12) ait au moins une solution admissible. Il est facile de construire des exemples d'ensemble admissible X_{ad} non vide mais tel que X_{ad}^0 est vide (voir l'Exercice 4.2.6). L'hypothèse que X_{ad}^0 est borné est équivalente à celle que $X_{ad} = \overline{X_{ad}^0}$ est borné.

Cette hypothèse est nécessaire car on peut construire des exemples où, X_{ad} n'étant pas borné, le programme linéaire (4.10) admet une solution optimale x^0 mais le problème pénalisé (4.12) n'admet pas de solution optimale (voir l'Exercice 4.2.6 pour un tel contre-exemple). •

Démonstration. La fonction $\mu\pi(x) + c \cdot x$ est strictement convexe sur $\mathbb{R}^n_+$ donc, s'il existe un point de minimum à (4.12), il est unique. Par contre, cette fonction n'est pas infinie à l'infini donc on a besoin de l'hypothèse que X^0_{ad} est borné. Par ailleurs, X^0_{ad} est convexe mais pas fermé, donc on ne peut pas utiliser directement le Théorème 2.2.1 bien que J soit continue et X^0_{ad} borné. Soit $x^n \in X^0_{ad}$ une suite minimisante de (4.12). Comme $\overline{X^0_{ad}} = X_{ad}$ est borné, on peut en extraire une sous-suite, toujours notée x^n, qui converge vers une limite x^μ dans X_{ad}. En fait, toutes les composantes de x^μ sont strictement positives et $x^\mu \in X^0_{ad}$ car sinon

$$\lim_{n\to+\infty} \pi(x^n) = +\infty,$$

ce qui serait une contradiction avec le fait que la valeur minimale de (4.12) est bornée supérieurement puisque X^0_{ad} n'est pas vide. Comme la fonction $\mu\pi(x) + c \cdot x$ est continue, on a donc obtenu

$$\mu\pi(x^\mu) + c \cdot x^\mu = \lim_{n\to+\infty} \mu\pi(x^n) + c \cdot x^n = \min_{x\in X^0_{ad}} \mu\pi(x) + c \cdot x.$$

Autrement dit, x^μ est la solution optimale de (4.12). On écrit les conditions d'optimalité (comme toujours dans ce chapitre, on suppose que la matrice A, de taille $m \times n$, est de rang $m \leq n$, donc les contraintes sont qualifiées) pour x^μ : il existe un multiplicateur de Lagrange $p^\mu \in \mathbb{R}^m$ tel que

$$Ax^\mu = b, \quad c - \mu\frac{1}{x^\mu} + A^* p^\mu = 0, \tag{4.13}$$

où $1/x$ désigne le vecteur de $\mathbb{R}^n$ de composantes $1/x_i$. On veut passer à la limite, quand $\mu \to 0$, dans (4.13). Comme l'ensemble X_{ad} est borné, la suite x^μ est bornée et on peut en extraire une sous-suite qui converge vers une limite $x^0 \in X_{ad}$ car X_{ad} est fermé. Par contre, en général la suite p^μ n'est pas bornée, quand $\mu \to 0$, et nous allons simplement montrer que la suite A^*p^μ est bornée dans $\mathbb{R}^n$. On déduit de (4.13) que

$$A^* p^\mu + c \geq 0. \tag{4.14}$$

Au vu de (4.14) les composantes de A^*p^μ sont bornées inférieurement. Montrons qu'elles sont aussi bornées supérieurement. Soit $y \in X^0_{ad}$ qui est supposé non vide. On multiplie la contrainte $Ay = b$ par p^μ pour obtenir

$$y \cdot A^* p^\mu = b \cdot p^\mu.$$

Par ailleurs, si on multiplie la deuxième égalité de (4.13) par x^μ on obtient

$$c \cdot x^\mu - n\mu + A^* p^\mu \cdot x^\mu = c \cdot x^\mu - n\mu + p^\mu \cdot b = 0, \tag{4.15}$$

donc

$$y \cdot A^* p^\mu = n\mu - c \cdot x^\mu \leq C,$$

car la suite x^μ est bornée. Comme $y > 0$, les composantes de $A^* p^\mu$ sont bornées supérieurement. Autrement dit, $A^* p^\mu$ est une suite bornée dans $\mathbb{R}^n$. On peut donc en extraire une sous-suite qui converge. Par ailleurs, elle appartient à $\mathrm{Im} A^*$ qui est un sous-espace vectoriel fermé, donc il existe $p^0 \in \mathbb{R}^m$ tel que la sous-suite $A^* p^\mu$ converge vers $A^* p^0$.

On passe alors à la limite dans la première égalité de (4.13), dans (4.14) et (4.15)

$$Ax^0 = b, \quad A^* p^0 + c \geq 0, \quad c \cdot x^0 + A^* p^0 \cdot x^0 = 0.$$

Il s'agit des conditions d'optimalité de (4.10), donc en vertu du théorème de Kuhn et Tucker (voir le Théorème 4.2.13 plus bas) x^0 est la solution optimale de (4.10) et p^0 est un multiplicateur de Lagrange pour ce problème. Si on suppose que (4.10) admet une solution unique, alors toute la suite x^μ converge vers x^0 lorsque μ tend vers zéro. □

Exercice 4.2.6 Donner un exemple de programme linéaire (4.10) qui admet une solution optimale x^0 mais pour lequel l'ensemble admissible X_{ad} n'est pas borné et le problème pénalisé (4.12) n'admet pas de solution optimale. Donner un autre exemple d'ensemble admissible X_{ad} non vide mais tel que X^0_{ad} est vide.

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)

$$\inf_{x \in \mathbb{R}^n \text{tel que } Ax=b,\ x \geq 0} c \cdot x, \tag{4.16}$$

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 (4.16)

$$L(x, p) = c \cdot x + p \cdot (b - Ax), \tag{4.17}$$

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

$$G(p) = \begin{cases} p \cdot b & \text{si } A^* p - c \leq 0 \\ -\infty & \text{sinon.} \end{cases} \tag{4.18}$$

Le problème dual de (4.16) est donc

$$\sup_{p \in \mathbb{R}^m \text{tel que } A^* p - c \leq 0} p \cdot b. \tag{4.19}$$

L'espace de solutions admissibles du problème dual (4.19) 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 (4.16) est

$$X_{ad} = \{x \in \mathbb{R}^n \text{ tel que } Ax = b,\ x \geq 0\}.$$

Les programmes linéaires (4.16) et (4.19) sont dits en **dualité**. L'intérêt de cette notion vient du résultat suivant qui est un cas particulier du Théorème de dualité 2.6.13.

Théorème 4.2.13 *Si (4.16) ou (4.19) a une valeur optimale finie, alors il existe $\overline{x} \in X_{ad}$ solution optimale de (4.16) et $\overline{p} \in P_{ad}$ solution optimale de (4.19) qui vérifient*

$$\left(\min_{x\in\mathbb{R}^n \text{tel que } Ax=b,\ x\geq 0} c\cdot x\right) = c\cdot\overline{x} = \overline{p}\cdot b = \left(\max_{p\in\mathbb{R}^m \text{tel que } A^*p-c\leq 0} p\cdot b\right) \tag{4.20}$$

De plus, $\overline{x}$ et $\overline{p}$ sont solutions optimales de (4.16) et (4.19) si et seulement si elles vérifient les conditions d'optimalité de Kuhn et Tucker

$$A\overline{x} = b,\ \overline{x} \geq 0,\ A^*\overline{p} - c \leq 0,\ \overline{x}\cdot(c - A^*\overline{p}) = 0. \tag{4.21}$$

Si (4.16) ou (4.19) a une valeur optimale infinie, alors l'ensemble des solutions admissibles de l'autre problème est vide.

Remarque 4.2.14 Une conséquence immédiate du Théorème 4.2.13 de dualité est que, si $x \in X_{ad}$ et $p \in P_{ad}$ sont deux solutions admissibles de (4.16) et (4.19), 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 (4.16) et $\overline{p}$ de (4.19). Ces deux propriétés permettent de trouver facilement des bornes pour les valeurs optimales de (4.16) et (4.19), 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 (4.20) et la condition d'optimalité (4.21) sont alors une conséquence du Théorème de Kuhn et Tucker 2.6.4.

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 (4.19). 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 2.5.20. 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 2.5.20 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 (4.16) $-\infty$. 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 (4.16) est infinie, alors X_{ad} est vide. □

L'intérêt de la dualité pour résoudre le programme linéaire (4.16) est multiple. D'une part, selon l'algorithme choisi, il peut être plus facile de résoudre le problème dual (4.19) (qui a m variables et n contraintes d'inégalités) que le problème primal (4.16) (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 (4.16) 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 (4.16) 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 (4.19) 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 (4.19) est à nouveau (4.16).

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

$$\inf_{\substack{v \geq 0 \\ Av \leq b}} c \cdot v \,. \tag{4.22}$$

Montrer que le problème dual peut se mettre sous la forme suivante, avec $q \in \mathbb{R}^m$

$$\sup_{\substack{q \geq 0 \\ A^* q + c \geq 0}} -b \cdot q \,. \tag{4.23}$$

Soient v et q des solutions admissibles de (4.22) et (4.23), respectivement. Montrer que v et q sont des solutions optimales si, et seulement si,

$$(c + A^* q) \cdot v = 0 \quad \text{et} \quad (b - Av) \cdot q = 0. \tag{4.24}$$

Les deux égalités de (4.24) sont appelées **conditions des écarts complémentaires** (primales et duales, respectivement).

4.3 Vers la programmation linéaire en nombres entiers

Jusqu'ici nous avons considéré des problèmes d'**optimisation continue**, c'est-à-dire que la variable d'optimisation variait continûment dans un ensemble admissible inclus dans $\mathbb{R}^n$. Dans le cas de la programmation linéaire, cet ensemble admissible X_{ad} est un polyèdre. La programmation linéaire **en nombres entiers** consiste à abandonner ce point de vue continu et à rajouter la contrainte que la solution doit appartenir à $\mathbb{Z}^n$. Dans le cas d'un programme linéaire sous forme standard (4.1) on considère donc

$$\inf_{x \in \mathbb{Z}^n \text{ tel que } Ax = b,\ x \geq 0} c \cdot x, \tag{4.25}$$

où A est une matrice de taille $m \times n$ (avec $m \leq n$ et $\mathrm{rg}(A) = m$), $b \in \mathbb{R}^m$ et $c \in \mathbb{R}^n$. On dit que (4.25) est un programme linéaire en nombres entiers, ou bien en variables discrètes, ou encore un problème combinatoire puisqu'on peut, en théorie, énumérer toutes les solutions possibles afin de trouver la solution optimale. Evidemment, le nombre de solutions entières possibles est exponentiel dans la taille n, m des données et il n'est en général pas possible de faire cette énumération en pratique.

En toute généralité le problème (4.25) est extrêmement difficile à résoudre, en particulier puisque, l'espace des solutions admissibles étant discret, on ne peut pas utiliser la notion de gradient pour tester l'optimalité d'une solution ou pour passer d'une solution à une autre meilleure. Nous ne disons rien des méthodes générales pour résoudre (4.25) et nous nous contentons de traiter deux cas particuliers, importants du point de vue des applications, pour lesquels la programmation linéaire continue, vue à la section précédente, est utile et efficace. Ces deux cas particuliers correspondent à un petit miracle que nous expliquons tout de suite. En général,

puisque $\mathbb{Z}^n$ est un sous-ensemble (très petit !) de $\mathbb{R}^n$, la valeur minimale de (4.25) est plus grande que la valeur minimale de (4.1), qui est le même problème où $\mathbb{Z}^n$ est remplacé par $\mathbb{R}^n$ (parfois on appelle (4.1) problème relaxé, ou relâché, de (4.25) car on abandonne la contrainte que x doit être à valeurs entières). Le miracle est que, justement pour ces deux cas, la valeur minimale est la même ! Plus précisément, on verra qu'en utilisant l'algorithme du simplexe pour (4.1) on obtient une solution de (4.25), qui est donc à valeurs entières. Bien entendu, cela n'est pas vrai sans des hypothèses de structure très fortes sur les données. Ces deux cas, qui peuvent sembler exceptionnels, apparaissent en fait naturellement dans un certain nombre de problèmes combinatoires concrets : affectations, plus courts chemins et plus généralement problèmes de flots à coût minimum.

4.3.1 Matrices totalement unimodulaires

Commençons par étudier à quelles conditions sur les données entières A et b les solutions admissibles $x \in X_{ad}$ ou plus généralement les solutions du système linéaire $Ax = b$ sont, elles aussi entières. Etudions d'abord le cas des matrices carrées dans le lemme suivant.

Lemme 4.3.1 *Pour une matrice inversible $A \in \mathbb{Z}^{n\times n}$ les deux propriétés 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. Au vu des formules de Cramer pour la résolution du système linéaire $Ax = b$, il est clair que si $\det A = \pm 1$, alors $A^{-1}b \in \mathbb{Z}^n$ pour tout second membre $b \in \mathbb{Z}^n$. Pour prouver la réciproque montrons tout d'abord que $A^{-1} \in \mathbb{Z}^{n\times n}$. En effet, si on choisit b comme le i-ème vecteur de la base canonique de $\mathbb{R}^n$, on en déduit que la i-ème colonne de A^{-1}, qui coïncide avec $A^{-1}b$, est à coefficients entiers. En variant $1 \leq i \leq n$, on obtient donc $A^{-1} \in \mathbb{Z}^{n\times n}$. En particulier, cela implique que $\det A^{-1} \in \mathbb{Z}$ et, comme $1 = \det A \det A^{-1}$, cela montre que $\det A$ divise 1, c'est-à-dire que $\det A = \pm 1$. □

Ce lemme motive la définition suivante qui est centrale dans cette section.

Définition 4.3.2 *On dit qu'une matrice carrée $A \in \mathbb{Z}^{n\times n}$ est* **unimodulaire** *quand $\det A = \pm 1$, et qu'une matrice générale $A \in \mathbb{Z}^{m\times n}$ est* **totalement unimodulaire** *quand toute sous-matrice carrée extraite de A 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.

Proposition 4.3.3 *Soit le programme linéaire en nombres entiers (4.25). On suppose que les données $A \in \mathbb{Z}^{m\times n}$ et $b \in \mathbb{Z}^m$ sont entières et que la matrice A est totalement unimodulaire. Alors toute solution basique x de (4.25) est entière, c'est-à-dire que $x \in \mathbb{Z}^n$.*

Remarque 4.3.4 La Proposition 4.3.3 est très importante en pratique car elle implique que si l'on résout (4.25) par l'algorithme du simplexe "usuel", alors la solution que l'on obtiendra (ainsi que toutes les solutions intermédiaires) seront entières car le simplexe énumère des solutions basiques. Il n'y a rien à changer à l'algorithme du simplexe pour obtenir des solutions entières, du moins sous les hypothèses de la Proposition 4.3.3.

La Proposition 4.3.3 ne dit surtout pas que toutes les solutions optimales de (4.25) sont entières. D'ailleurs, il ne peut en être ainsi, à moins que la solution optimale ne soit unique, car, par linéarité du coût $c \cdot x$, tout barycentre de solutions optimales d'un programme linéaire est aussi solution optimale.

Il est intéressant de remarquer que l'on ne fait aucune hypothèse sur le vecteur c qui apparait dans le coût. Celui-ci peut ne pas être à valeurs entières mais, dans tous les cas, les solutions basiques (et donc une solution optimale au moins) seront entières. •

Démonstration. Rappelons qu'une solution basique correspond à un réarrangement des colonnes de A tel que $A = (B, N)$, où B est une matrice inversible $m \times m$ et N est une matrice $m \times (n-m)$, et que dans cette base on a $x = (x_B, x_N)$ avec $x_N = 0$, $x_B \geq 0$ et $Bx_B = b$. Comme A est totalement unimodulaire, B est unimodulaire et $x_B \in \mathbb{Z}^m$. Comme $x_N = 0$, on a bien $x \in \mathbb{Z}^n$. □

Bien sûr, pour que la Proposition 4.3.3 tienne toutes ses promesses, il faut vérifier qu'il existe "suffisamment" de matrices totalement unimodulaires. Il n'existe pas de caractérisation des matrices totalement unimodulaires mais on connait certaines conditions suffisantes. On donne ici une telle condition, très utile en pratique, due à Poincaré.

Lemme 4.3.5 *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 l'hypothèse sur les valeurs des coefficients de A est aussi vraie pour toutes les sous-matrices de A, il suffit de considérer le cas où la matrice A est carrée et de vérifier que $\det A$ vaut ± 1 ou 0. Si A possède une colonne nulle, alors $\det A = 0$. Si A possède une colonne avec seulement un coefficient non-nul, on développe son déterminant par rapport à cette colonne, et l'on conclut par récurrence sur la dimension de A que $\det A \in \{\pm 1, 0\}$. Il reste donc à considérer le cas où chaque colonne de A a exactement un coefficient 1 et un coefficient -1 (tous les autres étant nuls). Ainsi, la somme des coefficients de chaque colonne est nulle. Autrement dit, le vecteur $(1, \cdots, 1)$ appartient au noyau de la transposée de A et donc $\det A = 0$. □

Nous allons voir dans les sous-sections suivantes que l'on rencontre souvent en pratique des matrices qui vérifient les hypothèses du Lemme 4.3.5 de Poincaré.

Exercice 4.3.1 Soit une matrice $A \in \mathbb{Z}^{m \times n}$ telle que les premiers éléments de chacune de ses colonnes sont des 1 alors que tous les autres éléments sont des 0. Autrement dit, pour tout $1 \leq j \leq n$, il existe $n_j \in \mathbb{N}$ tel que $a_{ij} = 1$ pour $1 \leq i \leq n_j$ et $a_{ij} = 0$

pour $n_j + 1 \leq i \leq m$. Montrer qu'une telle matrice (appelée matrice d'intervalles) est totalement unimodulaire. Indication : on vérifiera qu'il suffit de démontrer le résultat pour des matrices carrées et on fera une récurrence sur la taille de la matrice.

4.3.2 Introduction aux problèmes de flots

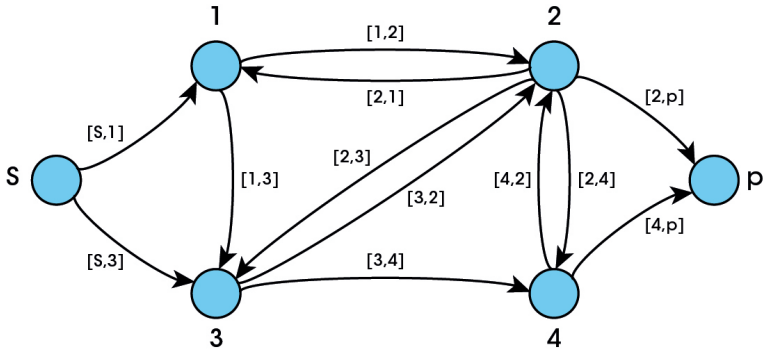

FIGURE 4.2 – Graphe orienté.

Nous faisons une brève présentation des problèmes de flots qui ne servent ici que de prétexte à illustrer la résolution de programmes linéaires en nombres entiers et, accessoirement, à donner des exemples de matrices vérifiant les hypothèses du Lemme 4.3.5 de Poincaré. Pour plus de détails sur les problèmes de flots et notamment pour d'autres algorithmes permettant de les résoudre, nous renvoyons à [6], [18]. Les problèmes de flots sont des problèmes d'optimisation sur des graphes, qui permettent de modéliser des réseaux (de transport, de communication, ou sociaux). Pour illustrer notre propos, imaginons qu'il s'agit d'un réseau de transport.

On introduit donc un graphe orienté $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ où $\mathcal{N}$ est l'ensemble des **nœuds**, numérotés de 1 à m, et $\mathcal{A} \subset \mathcal{N} \times \mathcal{N}$ est l'ensemble des **arcs** ou arêtes qui relient deux nœuds distincts. Les nœuds sont par exemple des villes et les arcs des lignes de transport qui les relient. On note (i, j) l'arc qui relier le nœud i au nœud j. Le graphe est dit orienté car l'arc (i, j) n'est pas le même que l'arc (j, i) (voir la Figure 4.2) : cela permet de distinguer le sens du trajet, ce qui est utile dans de nombreuses applications. Tous les nœuds ne sont pas forcément reliés entre eux par un arc et on note n le nombre total d'arcs. Chaque arc $(i, j) \in \mathcal{A}$ possède une **capacité** maximale $u_{ij} \in \mathbb{R}_+ \cup \{+\infty\}$ (la quantité maximale de biens ou de personnes pouvant circuler sur cette ligne de transport) ainsi qu'un **coût** $c_{ij} \in \mathbb{R}$ (le prix unitaire du voyage sur cette ligne). Remarquez que, si la capacité maximale est bien positive, on laisse la possibilité d'un coût négatif en cas de subventions de certaines lignes... En chaque nœud i du graphe il est aussi possible d'avoir la donnée d'un flot extérieur $b_i \in \mathbb{R}$, c'est-à-dire une quantité de biens ou de personnes qui arrivent en i par d'autres moyens de transport que ceux modélisés par le graphe. Si $b_i > 0$, ce flot exogène est dit entrant, tandis que si $b_i < 0$, le flot exogène est sortant.

On appelle **flot** un vecteur $x \in \mathbb{R}^n$, défini sur sur $\mathcal{A}$ par $(i, j) \mapsto x_{ij}$, vérifiant

la **loi des nœuds de Kirchoff**

$$b_i + \sum_{j\in\mathcal{N},(j,i)\in\mathcal{A}} x_{ji} = \sum_{j\in\mathcal{N},(i,j)\in\mathcal{A}} x_{ij}, \qquad \forall i \in \mathcal{N}, \tag{4.26}$$

ainsi que la contrainte de positivité

$$0 \leq x_{ij}, \qquad \forall (i,j) \in \mathcal{A} \ . \tag{4.27}$$

La loi des nœuds de Kirchoff (4.26) exprime la conservation de la quantité de biens ou de personnes au nœud i : tout ce qui arrive en i (terme de gauche) égale ce qui en repart (terme de droite). En sommant en i les lois des nœuds (4.26), il apparait une condition nécessaire pour l'existence d'un flot, à savoir que la somme des flots extérieurs ou exogènes doit être nulle

$$\sum_{i\in\mathcal{N}} b_i = 0 \ . \tag{4.28}$$

Dans la suite, nous supposerons systématiquement que cette condition (4.28) est vérifiée. Un flot est dit **admissible** s'il satisfait en plus les contraintes de capacité

$$x_{ij} \leq u_{ij}, \qquad \forall (i,j) \in \mathcal{A} \ . \tag{4.29}$$

On appelle **problème de flot à coût minimum** le problème d'optimisation

$$\min_{x\in\mathbb{R}^n} \sum_{(i,j)\in\mathcal{A}} c_{ij}x_{ij} \quad \text{sous les contraintes} \quad (4.26), (4.27), (4.29). \tag{4.30}$$

Il s'agit d'un programme linéaire qui admet toujours une solution si les coûts sont positifs, $c_{ij} \geq 0$, ou bien si les capacités sont finies, $u_{ij} < +\infty$. Pour donner une forme plus explicite de (4.30), sous la forme d'un programme linéaire, nous réécrivons la loi des nœuds de Kirchoff (4.26) comme $Ax = b$, où la matrice $A \in \mathbb{R}^{m\times n}$, appelée **matrice d'incidence nœuds-arcs** de $\mathcal{G}$, est définie par

$$A_{i,(j,k)} = \begin{cases} -1 & \text{si } i = k, \\ 1 & \text{si } i = j, \\ 0 & \text{sinon.} \end{cases} \tag{4.31}$$

La i-ème ligne de A est la loi de Kirchoff au nœud i, tandis que les élément non nuls de la (j,k)-ème colonne de A indiquent à quels nœuds (entrant ou sortant selon le signe) est relié l'arc (j,k). Le problème de flot (4.30) est donc équivalent à

$$\min_{x\in\mathbb{R}^n, Ax=b,\ 0\leq x\leq u} \sum_{(i,j)\in\mathcal{A}} c_{ij}x_{ij} \, . \tag{4.32}$$

Ce problème (4.32) n'est pas sous la forme standard (4.1) à cause de la contrainte de capacité maximale $x \leq u$. Néanmoins, grâce à l'introduction d'une variable d'écart

$y \in \mathbb{R}^n$, $y \geq 0$, la contrainte $x \leq u$ est équivalente à $x + y = u$. Par conséquent, (4.30) et (4.32) sont équivalents au programme linéaire sous forme standard

$$\min_{z \in \mathbb{R}^{2n}, \tilde{A}z = \tilde{b},\ 0 \leq z} \sum_{(i,j) \in \mathcal{A}} c_{ij} x_{ij} \,, \tag{4.33}$$

avec les notations $\tilde{b} \in \mathbb{R}^{m+n}$, $\tilde{A} \in \mathbb{R}^{(m+n) \times 2n}$ et

$$z = \begin{pmatrix} x \\ y \end{pmatrix}, \tilde{b} = \begin{pmatrix} b \\ u \end{pmatrix}, \tilde{A} = \begin{pmatrix} A & 0 \\ I_n & I_n \end{pmatrix} , \tag{4.34}$$

où I_n est la matrice identité dans $\mathbb{R}^{n \times n}$.

Remarque 4.3.6 Le problème de transport de l'Exemple 1.2.1 est en fait un exemple de problème de flot à coût minimum. Définissons le graphe dont les nœuds $\mathcal{N}$ sont les entrepôts, indicés par $1 \leq i \leq M$, et les clients, indicés par $1' \leq j' \leq N'$ (on rajoute un exposant $'$ pour distinguer les clients des entrepôts dans la liste des nœuds). Les arcs $\mathcal{A}$ de ce graphe orienté sont les arcs qui relient un entrepôt i à un client j', munis du coût c_{ij}. Il n'y a pas d'arcs entre clients, ni entre entrepôts, ni d'un client vers un entrepôt. Si les stocks sont égaux à la demande, $\sum_{i=1}^{M} s_i = \sum_{j=1}^{N} r_j$, alors les contraintes sur les quantités livrées $v_{ij} \geq 0$ sont la loi des nœuds de Kirchoff appliquée aux entrepôts i

$$\sum_{j=1}^{N} v_{ij} = s_i, \quad \text{pour } 1 \leq i \leq M,$$

et aux clients j

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

On choisit donc les flots externes comme $b_i = s_i$ pour les entrepôts et $b_{j'} = -r_{j'}$ pour les clients (notez le changement de signe de r_j pour prendre en compte le fait que les clients reçoivent alors que les entrepôts expédient). Avec le choix des capacités maximales infinies, $u_{ij} = +\infty$, le problème de transport de l'Exemple 1.2.1 est donc un problème de flot à coût minimum. Si les stocks sont supérieurs à la demande, $\sum_{i=1}^{M} s_i > \sum_{j=1}^{N} r_j$, alors il faut rajouter un client imaginaire qui reçoit tout le stock non consommé par les autres clients avec un coût de transport nul. Le problème est à nouveau un problème de flot à coût minimum. •

Il arrive souvent en pratique que l'on cherche 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 une fraction de colis peut ne pas avoir de sens. Il est donc naturel de se demander si un problème de flot à coût minimum, dont les données sont entières, a des solutions optimales qui sont aussi entières. Une réponse positive à cette question est apportée par le résultat suivant.

Théorème 4.3.7 *On suppose que les flots extérieurs ou exogènes sont entiers, $b_i \in \mathbb{Z}$, ainsi que les capacités, $u_{ij} \in \mathbb{N} \cup \{+\infty\}$. Alors les solutions basiques de* (4.33) *sont entières. En particulier, le problème* (4.30) *de flot à coût minimal admet une solution optimale entière $x \in \mathbb{N}^n$.*

Remarque 4.3.8 Une conséquence pratique importante du Théorème 4.3.7 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 (ou solution basique) à un autre du polyèdre des solutions admissibles, les solutions trouvées par le simplexe sont entières. •

Démonstration. Si les capacités sont toutes infinies, $u_{ij} = +\infty$, alors il n'y a pas de contraintes $x \leq u$ et le programme linéaire (4.32) est déjà sous forme standard, sans nécessité d'introduire les variables d'écart y. Or le Lemme 4.3.9 ci-dessous montre que la matrice A est totalement unimodulaire. On conclut alors à l'aide de la Proposition 4.3.3.

Si certaines capacités sont finies, alors il faut raisonner sur le programme linéaire (4.33) où il n'est pas évident que la matrice $\tilde{A}$ est totalement unimodulaire. Néanmoins, si une solution basique z vérifie $\tilde{A}z = \tilde{b}$, alors, au vu de la structure (4.34) de $\tilde{A}$, on a $Ax = b$ et $y = u - x$. Comme A est totalement unimodulaire, x est entier, donc y et z aussi, ce qui conclut la preuve. □

Lemme 4.3.9 *La matrice d'incidence nœuds-arcs A d'un graphe, définie par* (4.31)*, 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. On conclut grâce au Lemme 4.3.5 de Poincaré. □

Terminons cette section en donnant un exemple, très important en pratique, de problème de flot à coût minimum : le problème du **flot maximal**, qui s'intéresse seulement aux capacités et pas aux coûts. Dans le graphe $\mathcal{G}$ on distingue deux nœuds, notés s et p, appelés respectivement source et puits (voir la Figure 4.2), et on suppose que s n'a pas de prédécesseur, autrement dit $\{i \in \mathcal{N} \mid (i,s) \in \mathcal{A}\} = \emptyset$, et p n'a pas de successeur, c'est-à-dire $\{i \in \mathcal{N} \mid (p,i) \in \mathcal{A}\} = \emptyset$. Soit $v \in \mathbb{R}_+$. On appelle flot admissible de s à p de valeur v une solution x de $Ax = b$, $0 \leq x \leq u$ avec

$$b_i = \begin{cases} v & \text{si } i = s, \\ -v & \text{si } i = p, \\ 0 & \text{sinon.} \end{cases} \tag{4.35}$$

Le problème du flot maximal consiste à trouver un flot admissible de s à p de valeur v maximale. En d'autres termes, on cherche à faire passer de la source vers le puits le trafic maximal à travers le graphe $\mathcal{G}$.

Exercice 4.3.2 Montrer que le problème du flot maximal est bien un exemple de problème de flot à coût minimal. Pour cela, on rajoutera un arc reliant p à s au graphe

du problème du flot maximal, on annulera les deux flots externes b_s et b_p et on prendra les coûts $c_{ij} = 0$ sauf $c_{ps} = -1$.

Exercice 4.3.3 On reprend le graphe orienté $\mathcal{G}$ avec les deux nœuds, source s et puits p, tel que s n'a pas de prédécesseur et p n'a pas de successeur. On munit chaque arc (i, j) d'un coût $c_{ij} \geq 0$ qu'on interprète comme une distance entre les nœuds i et j. Le problème du **chemin de coût minimum** est de trouver une suite de nœuds consécutifs sur le graphe qui relient s à p et tels que la somme des coûts le long de ce chemin d'arcs est minimale. Montrer que ce problème peut se modéliser comme un problème de flot à coût minimum avec le choix (4.35) de flots externes pour $v = 1$ et des capacités maximales infinies.

4.3.3 Problème d'affectation

Rappelons le problème d'affectation, ou bien de mariage, tel que présenté dans l'Exemple 1.2.2. On a N femmes, indicées par $1 \leq i \leq N$, et N hommes, indicés par $1 \leq j \leq N$, avec une variable d'accord $a_{ij} \in \mathbb{R}^+$ qui indique l'attirance de i pour j. On cherche une permutation σ dans l'ensemble des permutations $\mathcal{S}_N$ de $\{1, ..., N\}$ qui réalise le maximum de

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

Bien sûr, comme le cardinal de l'ensemble $\mathcal{S}_N$ est $N!$, il est illusoire de vouloir résoudre ce problème d'affectation en énumérant simplement toutes les possibilités. C'est là une caractéristique des problèmes combinatoires où, même si l'espace de recherche est fini, sa grande taille rend impossible toute énumération exhaustive en pratique.

Si on introduit les variables de décision v_{ij}, qui vaut 1 s'il y a mariage entre la femme i et l'homme j et 0 sinon, on peut réécrire ce problème comme

$$\max_{(v_{ij})} \left(\sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} v_{ij} \right) \tag{4.36}$$

sous les contraintes

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

Les deux dernières contraintes égalité indiquent que chaque femme trouvera un mari et chaque homme une épouse (comme on maximise la réussite des mariages, on aurait pu écrire ces contraintes comme des inégalités aussi). Les matrices $v = (v_{ij})_{1 \leq i,j \leq N}$ qui vérifient ces contraintes sont précisément les matrices correspondantes aux permutations $\sigma \in \mathcal{S}_N$.

Si maintenant on relâche (ou relaxe) la contrainte $v_{ij} = 0$ ou 1, en $0 \leq v_{ij} \leq 1$, les matrices qui vérifient les contraintes relaxées

$$0 \leq v_{ij} \leq 1, \ \sum_{j=1}^{N} v_{ij} = 1, \ \sum_{i=1}^{N} v_{ij} = 1 \ \text{ pour } 1 \leq i, j \leq N \tag{4.37}$$

sont appelées matrices **doublement stochastiques**. Le problème (4.36) sous les contraintes (4.37) est un simple programme linéaire qu'on peut résoudre par l'algorithme du simplexe. Il est même posé sous forme standard car la contrainte $v_{ij} \leq 1$ est inutile au vu des contraintes égalité. Modulo le changement du max en min (il suffit de changer le signe des coûts a_{ij}), le problème (4.36) sous les contraintes (4.37) est un exemple de problème de transport qui, comme discuté à la Remarque 4.3.6, est un cas particulier de problème de flot à coût minimum. Pour montrer que ses solutions sont en fait entières, nous écrivons les contraintes égalité dans (4.37) sous la forme

$$\sum_{j=1}^{N} v_{ij} = 1 \text{ pour } 1 \leq i \leq N, \ -\sum_{i=1}^{N} v_{ij} = -1 \ \text{ pour } 1 \leq j \leq N,$$

où on a effectué la même astuce que dans la Remarque 4.3.6 pour que la contrainte de mariage pour chaque homme j ressemble à la loi des nœuds de Kirchoff. On peut alors résumer ces contraintes d'égalité par $Av = b$ avec $b \in \mathbb{R}^{2N}$ le vecteur dont les N premières composantes sont égales à 1 et les N dernières égales à -1 et $A \in \mathbb{Z}^{2N \times N^2}$ est une matrice dont chacune des colonnes a exactement un élément égal à 1, un autre égal à -1 et tous les autres nuls. Le programme linéaire relaxé est donc

$$\max_{v \geq 0, \ Av = b} \left(\sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} v_{ij} \right). \tag{4.38}$$

En vertu du Lemme 4.3.5 de Poincaré, la matrice A est totalement unimodulaire. Par conséquent, la Proposition 4.3.3 garantit que les solutions optimales de (4.38) sont entières. Autrement dit, il suffit d'appliquer l'algorithme du simplexe au programme linéaire relaxé (4.38), sans aucune précaution particulière, pour résoudre le problème d'affectation qui est de nature combinatoire !

Remarque 4.3.10 Il existe une autre preuve du caractère entier des solutions optimales de (4.38) (voir [6]). L'ensemble des matrices doublement stochastiques, c'est-à-dire qui vérifient (4.37), est un ensemble convexe compact dont on peut montrer (théorème de Birkhoff - von Neumann) que les points extrêmes sont précisément les matrices de permutations. On peut aussi montrer (théorème de Minkowski) que l'ensemble des matrices doublement stochastiques est l'enveloppe convexe de ses points extrêmes. Par linéarité de la fonction coût, on en déduit qu'il existe des solutions qui sont des matrices de permutation, autrement dit des solutions optimales entières. •

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, du terme source 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 temps final fixé, appelé aussi horizon temporel. 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 linéaire

$$\begin{cases} \dot{x}(t) = A(t)x(t) + B(t)u(t) + f(t), \quad \forall t \in [0,T], \\ x(0) = x_0, \end{cases} \tag{5.1}$$

avec $A \in L^1([0,T];\mathbb{R}^{d\times d})$, $B \in L^1([0,T];\mathbb{R}^{d\times k})$, où $d \geq 1$ et $k \geq 1$, un terme source $f \in L^1([0,T];\mathbb{R}^d)$, appelé **terme de dérive**, et une fonction temporelle, appelée **contrôle**,

$$u : [0,T] \to \mathbb{R}^k$$

qui permet d'agir sur le système afin d'en modifier l'état. 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

$$\dot{x}_u(t) = A(t)x_u(t) + B(t)u(t) + f(t), \quad \forall t \in [0,T], \qquad x_u(0) = x_0 \in \mathbb{R}^d. \tag{5.2}$$

Par la suite, nous supposerons que

$$u \in L^1([0,T];\mathbb{R}^k),$$

c'est-à-dire que le contrôle est une fonction mesurable et intégrable en temps. Notez que nous ne supposons pas que le contrôle est continu car il est souvent intéressant en pratique d'avoir des contrôles discontinus, comme les contrôles "bang-bang" que nous verrons plus loin (ainsi, un moteur qu'on allume ou qu'on éteint). 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

$$\|u\|_{L^1([0,T];\mathbb{R}^k)} = \int_0^T |u(s)|_{\mathbb{R}^k}\, ds,$$

où $|\cdot|_{\mathbb{R}^k}$ désigne la norme euclidienne sur $\mathbb{R}^k$. (On peut remplacer la norme euclidienne par toute autre norme sur $\mathbb{R}^k$.)

Définition 5.1.1 (Systèmes de contrôle linéaires) *On dit que* (5.2) *est un* ***système de contrôle linéaire****. Lorsque les matrices $A(t)$ et $B(t)$ dépendent du temps, on dit que le système de contrôle linéaire est* ***instationnaire****. Au contraire, si les matrices $A \in \mathbb{R}^{d\times d}$, $B \in \mathbb{R}^{d\times k}$ ne dépendent pas du temps, on dit que le système est* ***autonome*** *(ou* ***stationnaire****).*

Pour commencer, nous considérerons le système de contrôle linéaire autonome

$$\dot{x}(t) = Ax(t) + Bu(t) + f(t), \quad \forall t \in [0,T], \qquad x(0) = x_0 \in \mathbb{R}^d. \tag{5.3}$$

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 (5.3). 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 $G : [0,T] \to \mathbb{R}^d$ est absolument continue sur $[0,T]$ et on écrit $G \in AC([0,T];\mathbb{R}^d)$ s'il existe $g \in L^1([0,T];\mathbb{R}^d)$ telle que*

$$G(t) - G(0) = \int_0^t g(s)\, ds, \quad \forall t \in [0,T].$$

Il est facile de vérifier qu'une fonction G, absolument continue sur $[0,T]$, est aussi continue sur $[0,T]$ et dérivable presque partout, de dérivée égale à g.

Lemme 5.1.3 (Formule de Duhamel) *Pour tout contrôle $u \in L^1([0,T];\mathbb{R}^k)$ et toute dérive $f \in L^1([0,T];\mathbb{R}^d)$, il existe une unique trajectoire $x_u \in AC([0,T];\mathbb{R}^d)$ solution de* (5.3) *au sens où cette trajectoire vérifie la condition initiale $x_u(0) = x_0$ et le système différentiel $\dot{x}_u(t) = Ax_u(t) + Bu(t) + f(t)$ presque partout (p.p.) sur $[0,T]$. Cette trajectoire est donnée par la formule de Duhamel*

$$x_u(t) = e^{tA}x_0 + \int_0^t e^{(t-s)A}(Bu(s) + f(s))\,ds, \quad \forall t \in [0,T].$$

On notera que cette expression a bien un sens pour $(Bu + f) \in L^1([0,T];\mathbb{R}^d)$ car la fonction $s \mapsto e^{(t-s)A}$ est bornée sur $[0,T]$.

Démonstration. On rappelle que, pour une matrice carrée A, l'exponentielle de matrice est définie par

$$e^A = \sum_{n\geq 0} \frac{1}{n!}A^n.$$

On vérifie alors que $\frac{d}{dt}e^{tA} = Ae^{tA} = e^{tA}A$, donc que $x_u(t)$, défini par la formule de Duhamel, est bien une solution dans $AC([0,T];\mathbb{R}^d)$ de (5.3). L'unicité est une conséquence du Théorème 8.3.2 de Cauchy–Lipschitz (dans le cas mesurable en temps). □

5.1.2 Cas sans contraintes : critère de Kalman

Définition 5.1.4 (Contrôlabilité) *On dit que le système autonome* (5.3) *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 5.1.5 Dans la Définition 5.1.4 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 $x_2 = x_1 - e^{TA}x_0 - \int_0^T e^{(T-s)A}f(s)\,ds$, la contrôlabilité en T à partir de x_0 équivaut à

$$\forall x_2 \in \mathbb{R}^d, \quad \exists u \in L^\infty([0,T];\mathbb{R}^k), \quad x_2 = \int_0^T e^{(T-s)A}Bu(s)\,ds,$$

i.e., à la **surjectivité** de l'application

$$\Phi : L^\infty([0,T];\mathbb{R}^k) \to \mathbb{R}^d, \quad \Phi(u) = \int_0^T e^{(T-s)A}Bu(s)\,ds. \tag{5.4}$$

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 5.1.6 (Critère de Kalman) *Le système linéaire autonome* (5.3) *est contrôlable pour tout $T > 0$, pour tout $f \in L^1([0,T];\mathbb{R}^d)$ 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, c'est-à-dire que

$$\operatorname{rang}(C) = d.$$

Remarque 5.1.7 La condition de Kalman, $\operatorname{rang}(C) = d$, est **indépendante** de l'horizon temporel $T > 0$, de la dérive $f(t)$ et de la donnée initiale $x_0 \in \mathbb{R}^d$. La contrôlabilité d'un système linéaire autonome est donc indépendante de ces trois 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 5.1.8 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. Dans la nouvelle base, le système linéaire autonome (5.3) s'écrit

$$\dot{y}(t) = \widetilde{A}y(t) + \widetilde{B}u(t) + \widetilde{f}(t),$$

avec $y(t) = P^{-1}x(t)$, $\widetilde{A} = P^{-1}AP$, $\widetilde{B} = P^{-1}B$, $\widetilde{f}(t) = P^{-1}f(t)$, 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, $\operatorname{rang}(C) = \operatorname{rang}(\widetilde{C})$. •

Démonstration. (1) Supposons d'abord que $\operatorname{rang}(C) < d$. Par conséquent les lignes de C sont liées et il existe un vecteur $\Psi \in \mathbb{R}^d$, $\Psi \neq 0$, tel que $\Psi^* C = 0$. Rappelons que Ψ^* est le vecteur transposé de Ψ, c'est-à-dire le vecteur ligne correspondant au vecteur colonne Ψ, et que $\Psi^* C$ désigne la multiplication à gauche de la matrice C par ce vecteur ligne. Par définition de C on a donc les égalités suivantes entre vecteurs de $\mathbb{R}^k$

$$\Psi^* B = \Psi^* AB = \cdots = \Psi^* A^{d-1}B = 0\,.$$

D'après le théorème de Cayley-Hamilton, il existe des réels $s_0, \cdots, s_{d-1}$ tels que

$$A^d = s_0\,\mathrm{Id} + \cdots + s_{d-1}A^{d-1},$$

où Id est la matrice identité dans $\mathbb{R}^{d\times d}$. On en déduit par récurrence que $\Psi^* A^k B = 0$ pour tout $k \in \mathbb{N}$, puis que $\Psi^* e^{tA} B = 0$ pour tout $t \in [0,T]$. Par conséquent, au vu de la définition (5.4) de l'application Φ, on a $\Psi^*\Phi(u) = 0$ pour tout contrôle u, i.e., l'application Φ ne peut être surjective dans $\mathbb{R}^d$.

(2) Réciproquement, si l'application Φ n'est pas surjective, il existe un vecteur $\Psi \in \mathbb{R}^d$, $\Psi \neq 0$, tel que

$$\Psi^* \int_0^T e^{(T-s)A} B u(s)\, ds = 0, \quad \forall u \in L^\infty([0,T];\mathbb{R}^k).$$

En choisissant le contrôle $u(s) = B^* e^{(T-s)A^*}\Psi$, qui est bien dans $L^\infty([0,T];\mathbb{R}^k)$, et comme $\xi^*\xi = \xi \cdot \xi = |\xi|^2$ pour tout vecteur $\xi \in \mathbb{R}^k$, on en déduit que

$$\int_0^T \left|\Psi^* e^{(T-s)A} B\right|^2 ds = 0,$$

c'est-à-dire que $\Psi^* e^{tA} B = 0$ pour tout $t \in [0,T]$. 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 5.1.1 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

$$C = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \operatorname{rang}(C) = 2.$$

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 5.1.2 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

$$C = \begin{pmatrix} 0 & \frac{1}{L} \\ \frac{1}{L} & \frac{-R}{L^2} \end{pmatrix}, \qquad \operatorname{rang}(C) = 2,$$

ce qui montre que le circuit RLC est contrôlable. •

Exercice 5.1.1 On considère un point matériel se déplaçant le long d'une droite et dont la position est repérée à l'instant $t \geq 0$ par la variable $x(t) \in \mathbb{R}$, régie par l'équation différentielle en temps d'ordre $n \geq 1$ suivante :

$$\frac{d^n x_u}{dt^n}(t) + a_1 \frac{d^{n-1} x_u}{dt^{n-1}}(t) + \ldots + a_{n-1} \frac{dx_u}{dt}(t) + a_n x_u(t) = u(t),$$

où les coefficients $a_1, \ldots, a_n$ sont des réels donnés et la fonction $u(t) \in \mathbb{R}$ est le contrôle. Ecrire cette équation sous la forme d'un système de contrôle linéaire, différentiel du premier ordre. Montrer que ce système est contrôlable.

En particulier, l'Exercice 5.1.1 montre que le système masse-ressort ($m > 0$ et $k > 0$) pour un point matériel $x(t) \in \mathbb{R}$ (appelé aussi oscillateur harmonique)

$$m\ddot{x}(t) + kx(t) = u(t),$$

avec un contrôle $u(t) \in \mathbb{R}$ qui est la force appliquée, est contrôlable.

Exercice 5.1.2 Soit $\alpha \in \mathbb{R}$ un paramètre. Pour $t \geq 0$ on considère le système

$$\begin{pmatrix} \dot{x}(t) \\ \dot{y}(t) \end{pmatrix} = \begin{pmatrix} 1 & e^\alpha \\ 0 & 8 \end{pmatrix} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} + \begin{pmatrix} 1 & 1 \\ \alpha^2 - \alpha & 0 \end{pmatrix} \begin{pmatrix} u(t) \\ v(t) \end{pmatrix},$$

où $\begin{pmatrix} x \\ y \end{pmatrix}$ est l'état et $\begin{pmatrix} u \\ v \end{pmatrix}$ est le contrôle. Pour quelles valeurs de α le système est-il contrôlable ?

Il est intéressant pour la suite de considérer une reformulation du critère de Kalman. On introduit la matrice $G_T \in \mathbb{R}^{d\times d}$ telle que

$$G_T = \int_0^T e^{(T-s)A} B B^* e^{(T-s)A^*}\, ds. \tag{5.5}$$

Il est clair que la matrice G_T est symétrique, et on vérifie facilement qu'elle est semi-définie positive car $y^* G_T y = \int_0^T |B^* e^{(T-s)A^*} y|^2_{\mathbb{R}^k}\, ds \geq 0$ pour tout vecteur $y \in \mathbb{R}^d$.

Lemme 5.1.9 (Reformulation du critère de Kalman) *Le système linéaire autonome* (5.3) *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* (5.5)*, est inversible.*

Démonstration. (1) Soit $x_1 \in \mathbb{R}^d$. Supposons la matrice G_T inversible et posons

$$\overline{u}(t) = B^* e^{(T-t)A^*} y \quad \text{où} \quad y = G_T^{-1} \left(x_1 - e^{TA} x_0 - \int_0^T e^{(T-s)A} f(s)\, ds \right).$$

Par la formule de Duhamel, on voit que

$$x_{\overline{u}}(T) = e^{TA}x_0 + \int_0^T e^{(T-s)A}(B\overline{u}(s)+f(s))\,ds = e^{TA}x_0 + \int_0^T e^{(T-s)A}f(s)\,ds + G_T y = x_1.$$

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

$$0 = \Psi^* G_T \Psi = \int_0^T |B^* e^{(T-s)A^*}\Psi|^2_{\mathbb{R}^k}\,ds,$$

si bien que $\Psi^* e^{(T-s)A}B = 0$ pour tout $s \in [0,T]$. Par la formule de Duhamel, on obtient $\Psi^*(x_u(T) - e^{TA}x_0 - \int_0^T e^{(T-s)A}f(s)\,ds) = 0$, ce qui montre que $x_u(T)$ est dans un hyperplan affine. Par conséquent, le système n'est pas contrôlable. □

Remarque 5.1.10 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 se généralise au cas de la contrôlabilité des systèmes linéaires instationnaires •

Revenons maintenant au cas général d'un système linéaire instationnaire de contrôle (5.1), qui est

$$\dot{x}_u(t) = A(t)x_u(t) + B(t)u(t) + f(t), \quad \forall t \in [0,T], \qquad x_u(0) = x_0,$$

avec $A \in L^1([0,T];\mathbb{R}^{d\times d})$ et $B \in L^1([0,T];\mathbb{R}^{d\times k})$. Pour un tel système, 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

$$\dot{R}(t) = A(t)R(t), \quad R(0) = \mathrm{Id},$$

où Id est la matrice identité de $\mathbb{R}^{d\times d}$. On vérifie facilement que, dans le cas autonome où $A(t) = A$, on a $R(t) = e^{tA}$.

Lemme 5.1.11 (Formule de la résolvante) *Pour tout contrôle $u \in L^\infty([0,T];\mathbb{R}^k)$ et toute dérive $f \in L^1([0,T];\mathbb{R}^d)$, il existe une unique trajectoire $x_u \in AC([0,T];\mathbb{R}^d)$ solution de* (5.1) *qui est donnée par la formule*

$$x_u(t) = R(t)x_0 + R(t)\int_0^t R(s)^{-1}B(s)u(s)\,ds, \quad \forall t \in [0,T].$$

Démonstration. Comme $A \in L^1([0,T];\mathbb{R}^{d\times d})$ implique que $R \in AC([0,T];\mathbb{R}^{d\times d})$, et comme $u \in L^\infty([0,T];\mathbb{R}^k)$, l'intégrale dans la formule ci-dessus est bien définie. On vérifie aisément que $x_u(t)$, donné par la formule ci-dessus, est une solution de (5.1) dans $AC([0,T];\mathbb{R}^d)$. L'unicité est toujours une conséquence du Théorème 8.3.2 de Cauchy–Lipschitz (dans le cas mesurable en temps). Le seul point à vérifier est que la matrice $R(t)$ est bien inversible à tout temps. Pour cela on rappelle que, si

$J(R) = \det R$ est défini sur l'ensemble des matrices $\mathbb{R}^{d\times d}$, alors sa différentielle est $\langle J'(R), S\rangle = \det R \operatorname{tr}(R^{-1}S)$ pour tout $S \in \mathbb{R}^{d\times d}$. Donc, avec $S = \dot{R}(t)$,

$$\frac{d}{dt}\det(R(t)) = \det(R(t))\operatorname{tr}\left(R(t)^{-1}\dot{R}(t)\right) = \operatorname{tr}(A(t))\det(R(t)),$$

Comme $\det(R(0)) = 1$, on en déduit par intégration

$$\det(R(t)) = e^{\int_0^t \operatorname{tr}(A(s))ds} > 0,$$

c'est-à-dire que $R(t)$ est inversible pour tout $t \in [0,T]$ (la quantité $\det(R(t))$ s'appelle le Wronskien au temps t). □

Proposition 5.1.12 (Critère de contrôlabilité, cas instationnaire) *Le système instationnaire* (5.1) *est contrôlable en temps T à partir de x_0* ***si et seulement si*** *la matrice de contrôlabilité*

$$K_T = \int_0^T R(s)^{-1}B(s)B(s)^*(R(s)^{-1})^*\,ds \in \mathbb{R}^{d\times d} \tag{5.6}$$

est inversible.

Démonstration. Identique au cas autonome du Lemme 5.1.9. □

Remarque 5.1.13 La condition (5.6) 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

$$K_T = e^{-TA}\left(\int_0^T e^{(T-s)A}BB^*e^{(T-s)A^*}\,ds\right)e^{-TA^*} = e^{-TA}G_Te^{-TA^*}$$

On retrouve donc le critère du Lemme 5.1.9 sur la matrice G_T. •

Contre-exemple 5.1.1 (Non-contrôlabilité) On considère le système de contrôle linéaire instationnaire ($d = 2$ et $k = 1$)

$$\dot{x}_u(t) = \begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}x_u(t) + \begin{pmatrix}\cos(t)\\ \sin(t)\end{pmatrix}u(t). \tag{5.7}$$

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 (5.7) 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. •

Exercice 5.1.3 Pour un état $x(t) \in \mathbb{R}^3$ et un contrôle $u(t) \in \mathbb{R}$ on considère le système $\dot{x}(t) = A(t)x(t) + B(t)u(t)$ avec

$$A(t) = \begin{pmatrix} t & 1 & 0 \\ 0 & t^3 & 0 \\ 0 & 0 & t^2 \end{pmatrix} \text{ et } B(t) \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}.$$

Montrer que ce système est contrôlable en temps final quelconque.

5.1.3 Observabilité

Dans cette section nous faisons une brève parenthèse pour évoquer la notion **d'observabilité** qui est, en quelque sorte, une notion duale de la contrôlabilité. Nous nous restreignons ici au cas des systèmes autonomes, pour simplifier. On considère donc

$$\begin{cases} \dot{x}_u(t) = Ax_u(t) + Bu(t) + f(t), \quad \forall t \in [0,T], \\ x_u(0) = x_0 \in \mathbb{R}^d, \end{cases} \tag{5.8}$$

avec des matrices constantes $A \in \mathbb{R}^{d\times d}$, $B \in \mathbb{R}^{d\times k}$, une dérive $f \in L^1([0,T];\mathbb{R}^d)$ et un contrôle $u \in L^1([0,T];\mathbb{R}^k)$. Alors que f et u sont des données connues, on suppose que la solution x_u n'est pas intégralement connue. Plus précisément, seules certaines combinaisons linéaires de la solution sont connues. On introduit une variable d'observation

$$y_u(t) = Dx_u(t), \tag{5.9}$$

où la matrice constante D appartient à $\mathbb{R}^{p\times d}$ avec $1 \leq p < d$, et on suppose que seule la fonction $t \mapsto y_u(t) \in \mathbb{R}^p$ est "observée", c'est-à-dire connue. C'est typiquement le cas si (5.8) modélise un phénomène physique dont on ne connait pas la donnée initiale x_0 et dont on mesure seulement un nombre p de sorties (et pas toutes les d composantes de x_u). La question de l'observabilité du système (5.8) est alors de savoir s'il existe un contrôle u qui permette de reconstruire la solution x_u à partir de l'observation y_u. Avant de donner une définition précise de l'observabilité posons nous la question de savoir quelle est l'obstruction possible à cette reconstruction. Tout d'abord, connaitre tout $x_u(t)$ est équivalent à connaitre simplement la donnée initiale x_0, à cause de la formule de Duhamel du Lemme 5.1.3. L'observabilité est donc la détermination de x_0 à partir de $y_u(t)$. Evidemment, on sait que l'observation $y_u(t)$ provient d'au moins une donnée initiale. La seule obstruction possible à la détermination univoque de x_0 est donc que plusieurs données initiales produisent la même observation $y_u(t)$, quelque soit le choix du contrôle u. L'observabilité se réduit ainsi à une question d'unicité.

Définition 5.1.14 (Observabilité) *On dit que le système autonome* (5.8) *est observable par* (5.9) *en temps T si pour tout $x_0, \tilde{x}_0 \in \mathbb{R}^d$, $x_0 \neq \tilde{x}_0$, il existe $u \in L^1([0,T];\mathbb{R}^k)$ tel que $y_u \neq \tilde{y}_u$ sur $[0,T]$, où $y_u = Dx_u$ et $\tilde{y}_u = D\tilde{x}_u$ désignent respectivement les observations pour les données initiales x_0 et $\tilde{x}_0$.*

Théorème 5.1.15 (Critère de Kalman) *Le système linéaire autonome* (5.8) *est observable par* (5.9) *pour tout $T > 0$, pour tout $f \in L^1([0,T];\mathbb{R}^d)$ et pour tout*

$x_0 \in \mathbb{R}^d$ ***si et seulement si*** *la matrice d'observabilité de Kalman* $\mathcal{O} \in \mathbb{R}^{pd\times d}$, *définie par*

$$\mathcal{O} = \begin{pmatrix} D \\ DA \\ DA^2 \\ \vdots \\ DA^{d-1} \end{pmatrix},$$

est de rang maximal, c'est-à-dire que $\text{rang}(\mathcal{O}) = d$.

Remarque 5.1.16 La condition $\text{rang}(\mathcal{O}) = d$ est **indépendante** de l'horizon temporel $T > 0$, de la dérive $f(t)$ et de la matrice B. En prenant la transposée (ou adjointe) de la matrice $\mathcal{O}$ et en comparant avec le critère de Kalman du Théorème 5.1.6, on voit que l'observabilité de (5.8) est équivalente à la contrôlabilité d'un autre système linéaire autonome $\dot{z}_u(t) = A^* z_u(t) + D^* u(t)$. Cette "dualité" entre les deux notions est exploitée dans de nombreuses situations (voir [33] pour plus de détails). •

Démonstration. Montrons tout d'abord que l'observabilité de (5.8) est équivalente à la condition suivante pour le système simplifié $\dot{x}(t) = Ax(t),\ x(0) = x_0,\ y(t) = Dx(t)$

$$x_0 \neq 0 \ \Rightarrow\ y(t) \neq 0 \text{ sur } [0,T]. \tag{5.10}$$

En effet, grâce à la formule de Duhamel la condition d'observabilité peut s'écrire

$$x_0 \neq \tilde{x}_0 \ \Rightarrow\ \exists t \in [0,T] \text{ tel que}$$
$$D\left(e^{tA}x_0 + \int_0^t e^{(t-s)A}(Bu(s) + f(s))\,ds\right) \neq D\left(e^{tA}\tilde{x}_0 + \int_0^t e^{(t-s)A}(Bu(s) + f(s))\,ds\right),$$

ou bien de manière équivalente

$$x_0 \neq \tilde{x}_0 \ \Rightarrow\ \exists t \in [0,T] \text{ tel que } De^{tA}x_0 \neq De^{tA}\tilde{x}_0,$$

qui n'est rien d'autre que la condition (5.10) en soustrayant les deux équations pour x_u et $\tilde{x}_u$ (ce qui donne le système simplifié) et en remplaçant $x_0 - \tilde{x}_0$ par x_0.

Montrons maintenant par contrapposée que (5.10) est équivalent à $\text{rang}(\mathcal{O}) = d$. (1) Si (5.10) n'est pas vrai, alors pour le système simplifié il existe $x_0 \neq 0$ tel que, pour tout $t \in [0,T]$,

$$y(t) = De^{tA}x_0 = 0.$$

En dérivant successivement $d-1$ fois par rapport à t et en prenant la valeur à $t = 0$, on obtient

$$Dx_0 = DAx_0 = DA^2x_0 = \cdots = DA^{d-1}x_0 = 0.$$

Autrement dit, $\mathcal{O}x_0 = 0$ et donc $\text{rang}(\mathcal{O}) < d$.
(2) Réciproquement, supposons que $\text{rang}(\mathcal{O}) < d$. Par conséquent, il existe un vecteur $x_0 \in \mathbb{R}^d$, $x_0 \neq 0$, tel que $\mathcal{O}x_0 = 0$, c'est-à-dire

$$Dx_0 = DAx_0 = DA^2x_0 = \cdots = DA^{d-1}x_0 = 0.$$

D'après le théorème de Cayley-Hamilton, il existe des réels $s_0, \cdots, s_{d-1}$ tels que

$$A^d = s_0 \,\mathrm{Id} + \cdots + s_{d-1} A^{d-1},$$

où Id est la matrice identité dans $\mathbb{R}^{d\times d}$. On en déduit par récurrence que $DA^k x_0 = 0$ pour tout $k \in \mathbb{N}$, puis que $De^{tA}x_0 = 0$ pour tout $t \in [0,T]$. Par conséquent, (5.10) n'est pas vérifié. □

La preuve du Théorème 5.1.15 n'est pas constructive et ne dit pas comment on pourrait en pratique calculer la solution x_u à partir de l'observation y_u. On va donc indiquer une procédure possible au cas où le contrôle et la dérive sont réguliers en temps.

Proposition 5.1.17 *On suppose que le système linéaire autonome* (5.8) *est observable par* (5.9) *et que la dérive $f(t)$ et le contrôle $u(t)$ sont $(d-2)$ fois dérivables à valeur dans $L^1([0,T];\mathbb{R}^d)$ et $L^1([0,T];\mathbb{R}^k)$, respectivement. On définit des observables $y^0(t) = y_u(t)$ et $y^j(t) = \dot{y}^{j-1}(t) - DA^{j-1}(Bu(t)+f(t))$ pour $1 \le j \le d-1$. Alors, la solution x_u de* (5.8) *est donnée par*

$$x_u(t) = \mathcal{P}\begin{pmatrix} y^0(t) \\ y^1(t) \\ y^2(t) \\ \vdots \\ y^{d-1}(t) \end{pmatrix},$$

où $\mathcal{P}$ est un inverse à gauche de $\mathcal{O}$, c'est-à-dire une matrice de $\mathbb{R}^{d\times pd}$ telle que $\mathcal{P}\mathcal{O} = \mathrm{Id}_{d\times d}$.

Démonstration. On dérive la relation $y(t) = Dx(t)$ et on remplace $\dot{x}$ grâce à (5.8), ce qui conduit à

$$DAx(t) = \dot{y}(t) - D(Bu(t)+f(t)) = y^1(t).$$

Autrement dit, la nouvelle observable y^1 (qui dépend des données f, u et de la dérivée en temps de l'observable initiale $y^0 = y$) donne une information sur $DAx(t)$. Si on dérive à nouveau cette formule, on obtient

$$DA^2x(t) = \dot{y}^1(t) - DA(Bu(t)+f(t)) = y^2(t).$$

Par récurrence, pour $1 \le j \le d-1$, on a donc

$$DA^jx(t) = \dot{y}^{j-1}(t) - DA^{j-1}(Bu(t)+f(t)) = y^j(t).$$

Grâce à l'hypothèse que $f(t)$ et $u(t)$ sont $(d-2)$ fois dérivables, toutes les observables y^j sont bien définies. Au final,

$$\mathcal{O}x_0 = \begin{pmatrix} y^0 \\ y^1 \\ y^2 \\ \vdots \\ y^{d-1} \end{pmatrix}.$$

Comme on a supposé le système (5.8) observable, $\text{rang}(\mathcal{O}) = d$ et donc $\mathcal{O}$ admet un inverse à gauche. Cette propriété classique peut se retrouver en notant $(\mathcal{O}_j)_{1\leq j\leq d}$ les vecteurs colonnes de $\mathcal{O}$ dans $\mathbb{R}^{pd}$, qui forment une famille libre. Soit $\mathcal{P}$ une matrice de $\mathbb{R}^{d\times pd}$ dont on note les vecteurs lignes $(\mathcal{P}_i^*)_{1\leq i\leq d}$ avec $\mathcal{P}_i \in \mathbb{R}^{pd}$. On choisit ces vecteurs sous la forme $\mathcal{P}_i = \sum_{j=1}^d p_{ij}\mathcal{O}_j$: la relation $\mathcal{P}\mathcal{O} = \text{Id}_{d\times d}$ est alors équivalente à $PO = \text{Id}_{d\times d}$, où P est la matrice $d \times d$ de coefficients p_{ij} et O est la matrice $d \times d$ de coefficients $o_{ij} = \mathcal{O}_i \cdot \mathcal{O}_j$ qui est inversible puisque la famille $(\mathcal{O}_j)_{1\leq j\leq d}$ est libre. On prend donc $P = O^{-1}$ qui définit un inverse à gauche (non unique) $\mathcal{P}$ de $\mathcal{O}$. □

Remarque 5.1.18 La Proposition 5.1.17 donne une formule **explicite** de la solution x_u en fonction de l'observable y_u et de ses dérivées. Par ailleurs, cette formule est valable pour n'importe quel contrôle u, y compris pour $u = 0$. En pratique la Proposition 5.1.17 n'est pas très commode à utiliser. En effet, s'il n'est pas difficile de calculer un inverse à gauche de la matrice $\mathcal{O}$, il est nettement plus délicat de dériver en temps les observables $y^j(t)$, qui sont en général issues de mesures, car la dérivation d'un signal temporel entraine des erreurs ou du bruit numérique très instable. •

Exercice 5.1.4 On considère un point matériel $x(t) \in \mathbb{R}$ dans un système masse-ressort ($m > 0$ et $k > 0$) où le contrôle $u(t) \in \mathbb{R}$ est la force appliquée, dont l'équation est $m\ddot{x}(t)+kx(t) = u(t)$. On observe ce système avec la variable $y(t) = \alpha x(t)+\beta\dot{x}(t) \in \mathbb{R}$. Donner l'ensemble des valeurs des paramètres $\alpha, \beta \in \mathbb{R}$ pour lesquelles ce système est observable.

5.1.4 Cas avec contraintes : ensemble atteignable

On considère toujours le système de contrôle linéaire autonome (5.3) mais désormais on suppose que le contrôle u est à valeurs dans un sous-ensemble **compact non-vide**

$$U \subset \mathbb{R}^k. \tag{5.11}$$

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, mais pour simplifier, nous ne traiterons pas ce cas plus général.

Définition 5.1.19 (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 :*

$$\mathcal{A}(t, x_0) = \{x_1 \in \mathbb{R}^d \mid \exists u \in L^\infty([0,t];U) \text{ tel que } x_u(t) = x_1\}, \tag{5.12}$$

où x_u est la trajectoire associée à u, solution de (5.3).

Proposition 5.1.20 (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*

$$\forall t_1, t_2 \in [0,T], \quad |t_1 - t_2| \leq \delta \implies d(\mathcal{A}(t_1, x_0), \mathcal{A}(t_2, x_0)) \leq \epsilon, \tag{5.13}$$

où d est la distance de Hausdorff entre deux sous-ensembles $\mathcal{A}_1$ et $\mathcal{A}_2$ de $\mathbb{R}^d$, définie comme suit (voir la Figure 5.1) :

$$d(\mathcal{A}_1,\mathcal{A}_2) := \max\left(\sup_{x_1\in\mathcal{A}_1} d(x_1,\mathcal{A}_2),\ \sup_{x_2\in\mathcal{A}_2} d(x_2,\mathcal{A}_1)\right)$$
$$= \max\left(\sup_{x_1\in\mathcal{A}_1}\inf_{y_2\in\mathcal{A}_2}|x_1-y_2|_{\mathbb{R}^d},\ \sup_{x_2\in\mathcal{A}_2}\inf_{y_1\in\mathcal{A}_1}|x_2-y_1|_{\mathbb{R}^d}\right). \tag{5.14}$$

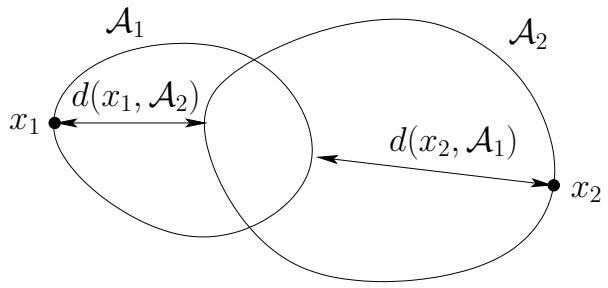

FIGURE 5.1 – Distance de Hausdorff entre deux sous-ensembles $\mathcal{A}_1$ et $\mathcal{A}_2$ de $\mathbb{R}^d$.

Remarque 5.1.21 Contrairement au cas de la Sous-section précédente où, en l'absence de restrictions sur le contrôle, et dans le cas d'un système contrôlable, l'ensemble atteignable est l'espace $\mathbb{R}^d$ tout entier, pour n'importe quel temps $t > 0$, ici le fait que U soit compact (donc borné) entraine immédiatement que l'ensemble atteignable $\mathcal{A}(t,x_0)$ est aussi compact (donc borné). Par conséquent, l'atteignabilité d'une cible prescrite n'est pas évidente et dépend bien sûr du temps alloué. On étudiera un peu plus loin la notion de temps minimal pour atteindre une cible. •

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

$$x_i = e^{tA}x_0 + \int_0^t e^{(t-s)A}(Bu_i(s)-f(s))\,ds, \tag{5.15}$$

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

$$\begin{aligned} x_u(t) &= e^{tA}x_0 + \int_0^t e^{(t-s)A}(Bu(s) - f(s))\,ds \\ &= e^{tA}x_0 + \theta\int_0^t e^{(t-s)A}(Bu_1(s) - f(s))\,ds + (1-\theta)\int_0^t e^{(t-s)A}(Bu_2(s) - f(s))\,ds \\ &= \theta x_1 + (1-\theta)x_2, \end{aligned}$$

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, l'argument est plus subtil car on mélange les deux contrôles u_1 et u_2 au cours du temps. On commence par invoquer le Lemme de Lyapunov 5.1.22 rappelé ci-dessous (pour la preuve, voir par exemple [19]). Soit $x_1, x_2 \in \mathcal{A}(t, x_0)$ et $\theta \in [0,1]$. Montrons 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 est donné par (5.15). Posons $y_i = x_i - e^{tA}x_0 - \int_0^t e^{(t-s)A}f(s)\,ds$ et considérons la fonction $g \in L^1([0,t]; \mathbb{R}^{2d})$ définie par

$$g(s) = \begin{pmatrix} e^{(t-s)A}Bu_1(s) \\ e^{(t-s)A}Bu_2(s) \end{pmatrix} \in \mathbb{R}^{2d}.$$

On a $\int_{\{0\}} g(s)\,ds = (0,0)^*$ et $\int_{[0,t]} g(s)\,ds = (y_1, y_2)^*$. En invoquant le lemme de Lyapunov, on en déduit qu'il existe un sous-ensemble mesurable $E \subset [0,t]$ tel que

$$\int_E g(s)\,ds = \begin{pmatrix} \theta y_1 \\ \theta y_2 \end{pmatrix}.$$

En notant E^c le complémentaire de E dans $[0,t]$, on a

$$\int_{E^c} g(s)\,ds = \int_{[0,t]} g(s)\,ds - \int_E g(s)\,ds = \begin{pmatrix} (1-\theta)y_1 \\ (1-\theta)y_2 \end{pmatrix}.$$

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

$$\begin{aligned} x_u(t) - e^{tA}x_0 - \int_0^t e^{(t-s)A}f(s)\,ds &= \int_{[0,t]} e^{(t-s)A}Bu(s)\,ds \\ = \int_E e^{(t-s)A}Bu_1(s)\,ds + \int_{E^c} e^{(t-s)A}Bu_2(s)\,ds &= \theta y_1 + (1-\theta)y_2, \end{aligned}$$

ce qui montre que $\theta x_1 + (1-\theta)x_2 = x_u(t) \in \mathcal{A}(t, x_0)$. □

Lemme 5.1.22 (Lyapunov) *Soit $t > 0$. Soit une fonction $g \in L^1([0,t];\mathbb{R}^n)$. Alors, le sous-ensemble*

$$\left\{ \int_E g(s)\,ds \mid E \subset [0,t] \text{ mesurable} \right\} \tag{5.16}$$

est un sous-ensemble ***convexe*** *de $\mathbb{R}^n$.*

Remarque 5.1.23 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 5.1.3 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. Les valeurs du temps t où le contrôle bang-bang est discontinu, c'est-à-dire saute d'une valeur ± 1 à son opposée, sont appelées les **temps de commutation**. Une illustration est présentée à la Figure 5.2. •

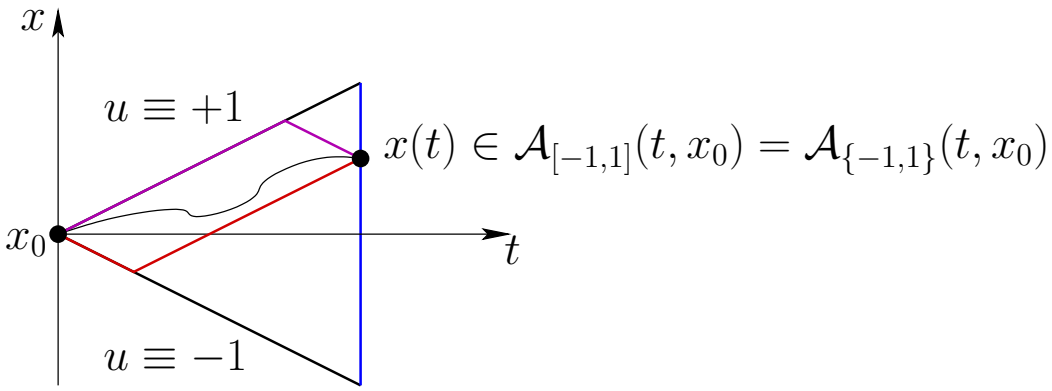

FIGURE 5.2 – Ensemble atteignable par un point matériel dont on contrôle la vitesse dans $U = [-1,1]$.

Exercice 5.1.5 On considère à nouveau un point matériel en mouvement rectiligne, mais on contrôle l'accélération de ce point par un contrôle à valeurs dans $U = [-1,1]$:

$$\ddot{x}(t) = u(t),\ \forall t \in [0,T], \quad x(0) = x_0,\ \dot{x}(0) = v_0, \quad u(t) \in U = [-1,1],$$

où $x_0 \in \mathbb{R}$ est la position initiale, et $v_0 \in \mathbb{R}$ la vitesse initiale, du point matériel. Déterminer explicitement l'ensemble atteignable $\mathcal{A}(t,x_0,v_0)$ au temps $t > 0$ et vérifier qu'il est

compact, convexe et varie continûment en t. Montrer qu'on obtient le même ensemble atteignable si on se restreint à des contrôles "bang-bang" à valeurs dans $\partial U = \{-1, 1\}$. Quel est l'ensemble atteignable si on prend désormais $U = [-1, -1/2] \cup [1/4, 1]$?

Exercice 5.1.6 On considère encore un point matériel $x(t) \in \mathbb{R}$ dans un système masse-ressort ($m > 0$ et $k > 0$) où le contrôle $u(t) \in U = [-1, 1]$ est la force appliquée, dont l'équation est

$$m\ddot{x}(t) + kx(t) = u(t) \; \forall t \in [0,T], \quad x(0) = x_0, \; \dot{x}(0) = v_0,$$

avec $x_0 \in \mathbb{R}$ et $v_0 \in \mathbb{R}$. Déterminer explicitement l'ensemble atteignable $\mathcal{A}(t, x_0, v_0)$ au temps $t > 0$ et vérifier qu'il est compact, convexe et varie continûment en t. Montrer que, si $x_0 = v_0 = 0$, l'ensemble $\mathcal{A}(t, 0, 0)$ est croissant en temps. Vérifier que ce n'est pas forcément le cas si (x_0, v_0) est grand.

Une fois défini l'ensemble atteignable il est naturel de se poser la question de l'atteignabilité en temps minimal d'un point x_1. Plus précisément, on considère le problème suivant

$$\inf_{\substack{u \in L^\infty([0,T];U) \\ \text{tel que } x_1 \in \mathcal{A}(t,x_0)}} t \tag{5.17}$$

autrement dit, étant donné un point initial x_0 et un point final x_1, il s'agit de trouver le contrôle u qui minimise le temps t où la trajectoire passe par la cible, $x_u(t) = x_1$.

Théorème 5.1.24 *Si la cible x_1 est atteignable à partir de x_0, alors il existe un contrôle qui minimise* (5.17)*, c'est-à-dire une trajectoire de temps minimal. Par ailleurs si t^* est le temps minimal, alors x_1 appartient au bord de l'ensemble atteignable, $x_1 \in \partial\mathcal{A}(t^*, x_0)$.*

Démonstration. Comme on suppose que x_1 est atteignable à partir de x_0, l'ensemble $I = \{t \geq 0 \text{ tel que } x_1 \in \mathcal{A}(t, x_0)\}$ est non vide. Montrons que I est fermé dans $\mathbb{R}$. Soit $t^n \in I, n \geq 0$, une suite qui converge vers $t \in \mathbb{R}$. D'après la Proposition 5.1.20 l'ensemble $\mathcal{A}(t, x_0)$ est compact (donc fermé et borné) et varie continûment en temps. Ainsi, la limite de $\mathcal{A}(t_n, x_0)$ n'est autre que $\mathcal{A}(t, x_0)$, donc x_1 qui appartient à chaque $\mathcal{A}(t_n, x_0)$ appartient aussi à $\mathcal{A}(t, x_0)$, ce qui veut dire que $t \in I$ et l'ensemble I est bien fermé. On en déduit que I a un minimum t^* qui appartient à I. Autrement dit, t^* est la valeur minimale de (5.17) qui vérifie $x_1 \in \mathcal{A}(t^*, x_0)$, donc le minimum est atteint dans (5.17).

D'autre part, si x_1 appartenait à l'intérieur de $\mathcal{A}(t^*, x_0)$, alors comme l'ensemble atteignable varie continûment en temps, il existerait un temps $t < t^*$ tel que $x_1 \in \mathcal{A}(t, x_0)$, ce qui contredit l'optimalité de t^*. Donc $x_1 \in \partial\mathcal{A}(t^*, x_0)$. □

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 introduisons la notion **d'application entrée-sortie**, qui à un contrôle associe l'état du système au temps final, dont nous calculons la différentielle. Certaines preuves sont un peu techniques (et font appel à des résultats rappelés dans l'Annexe 8) et il importe ici de comprendre l'esprit des résultats plus que la lettre des démonstrations.

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

$$\dot{x}_u(t) = f(t, x_u(t), u(t)), \quad \forall t \in [0,T], \qquad x_u(0) = x_0. \tag{5.18}$$

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 5.1.19).

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 :*

$$\mathcal{A}(t, x_0) = \{x_1 \in \mathbb{R}^d \mid \exists u \in L^\infty([0,t]; U) \text{ tel que } x_u(t) = x_1\},$$

où x_u est la trajectoire associée à u, solution de (5.18).

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 entre deux sous-ensembles, définie en (5.14).

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

$$x_2 = x_0 + \int_0^{t_2} f(s, x(s), u(s))\, ds.$$

Avec ce même contrôle, on pose

$$x_1 = x_0 + \int_0^{t_1} f(s, x(s), u(s))\, ds \in \mathcal{A}(t_1, x_0).$$

D'après les hypothèses sur f, x et u, on a

$$|x_2 - x_1|_{\mathbb{R}^d} \leq \int_{t_1}^{t_2} |f(s, x(s), u(s))|_{\mathbb{R}^d}\, ds \leq C|t_2 - t_1|.$$

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. En raison de l'hypothèse (iii), l'ensemble atteignable $\mathcal{A}(t, x_0)$ est borné. Il reste donc à montrer qu'il est fermé dans $\mathbb{R}^d$. La preuve est délicate car elle utilise des notions de topologie faible dans les espaces de Hilbert (voir la Sous-section 2.3.3 pour cette notion). L'espace de Hilbert en question est $V = L^2([0,T]; \mathbb{R}^d)$.
(1) Soit $(y_n)_{n \in \mathbb{N}}$ une suite d'éléments de $\mathcal{A}(T, x_0) \subset \mathbb{R}^d$. On va montrer qu'il existe une sous-suite de y_n qui converge vers une limite appartenant aussi à $\mathcal{A}(T, x_0)$. Soit

$(u_n)_{n\in\mathbb{N}}$ une suite de contrôles dans $L^\infty([0,T];U)$ et $(x_n = x_{u_n})_{n\in\mathbb{N}}$ la suite de trajectoires correspondantes dans $AC([0,T];\mathbb{R}^d)$ menant de x_0 à y_n. Posons $g_n(s) = f(s, x_n(s), u_n(s))$ pour tout $n \in \mathbb{N}$ et $s \in [0,T]$. On a

$$x_n(t) = x_0 + \int_0^t g_n(s)\,ds,\ \forall t \in [0,T] \quad \text{et} \quad y_n = x_n(T).$$

D'après les hypothèses, la suite $(g_n(s))_{n\in\mathbb{N}}$ est bornée dans $\mathbb{R}^d$, uniformément par rapport à $s \in [0,T]$. Donc la suite $(g_n)_{n\in\mathbb{N}}$ est bornée dans V. En invoquant le Lemme 2.3.14 sur la compacité faible dans les espaces de Hilbert, on en déduit qu'il existe une sous-suite n' telle que $(g_{n'})_{n'}$ 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

$$x(t) = x_0 + \int_0^t g(s)\,ds,\ \forall t \in [0,T].$$

Par convergence faible, on a

$$\lim_{n'\to+\infty} \left\{ \int_0^t g_{n'}(s)\,ds = \langle g_{n'}, 1_{[0,t]}\rangle_V \right\} = \langle g, 1_{[0,t]}\rangle_V = \int_0^t g(s)\,ds,$$

c'est-à-dire

$$\lim_{n'\to+\infty} x_{n'}(t) = x(t), \quad \forall t \in [0,T].$$

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 $\zeta_n(s) = f(s, x(s), u_n(s))$ et introduisons l'ensemble

$$\mathcal{K} = \{\zeta \in V \mid \zeta(s) \in K(s, x(s)),\ \forall s \in [0,T]\},$$

de sorte que $(\zeta_n)_{n\in\mathbb{N}}$ est une suite de $\mathcal{K}$. 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 $\mathcal{K}$ est un sous-ensemble convexe de V. De plus, $\mathcal{K}$ 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 $\mathbb{R}^d$ (puisque f est continue). Grâce au Lemme 2.3.16 sur la fermeture faible des convexes dans les espaces de Hilbert, on en déduit que $\mathcal{K}$ est faiblement fermé dans V. De plus, comme la suite $(\zeta_n)_{n\in\mathbb{N}}$ est bornée dans V, on déduit du Lemme 2.3.14 qu'elle converge faiblement, à une sous-suite près, vers une fonction $\zeta \in \mathcal{K}$. Il existe donc une fonction $u : [0,T] \to U$ telle que $\zeta(s) = f(s, x(s), u(s))$ 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

$$\begin{aligned} \int_0^T g_n(s)\varphi(s)\,ds &= \int_0^T \zeta_n(s)\varphi(s)\,ds \\ &\quad + \int_0^T (f(s, x_n(s), u_n(s)) - f(s, x(s), u_n(s))\varphi(s)\,ds. \end{aligned} \tag{5.19}$$

Comme $|f(s,x_n(s),u_n(s))-f(s,x(s),u_n(s))|_{\mathbb{R}^d} \leq C|x_n(s)-x(s)|_{\mathbb{R}^d}$ et $|x_n(s)-x(s)|_{\mathbb{R}^d}$ tend vers zéro p.p. dans $[0,T]$ (à une sous-suite près), le deuxième terme au membre de droite de (5.19) tend vers zéro (invoquer le théorème de convergence dominée de Lebesgue). En outre, par convergence faible, on a $\int_0^T g(s)\varphi(s)\,ds = \int_0^T \zeta(s)\varphi(s)\,ds$, i.e., $g(s) = \zeta(s)$ p.p. dans $[0,T]$. En conclusion, on a bien $g(s) = f(s,x(s),u(s))$ p.p. sur $[0,T]$. □

Remarque 5.2.6 Le Théorème 5.1.24 d'existence d'une trajectoire de temps minimal se généralise sans difficulté au cas non-linéaire sous les hypothèses des Lemmes 5.2.2 et 5.2.4 qui garantissent que l'ensemble atteignable $\mathcal{A}(t,x_0)$ est compact et varie continûment en temps. •

5.2.2 Contrôlabilité locale des systèmes non-linéaires

On considère toujours le système de contrôle non-linéaire (5.18). On suppose désormais que la fonction $f(t,x,u)$ est de classe C^1 en (x,u).

Définition 5.2.7 (Application entrée-sortie) *L'**application entrée-sortie** en temps T à partir de x_0 est l'application E_{T,x_0} définie par*

$$\begin{array}{rcl} \mathcal{U}_{T,x_0} & \to & \mathcal{A}(T,x_0) \\ u & \to & E_{T,x_0}(u) = x_u(T) \end{array}$$

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 $\mathcal{A}(T,x_0) \subset \mathbb{R}^d$ 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 5.2.8 (Contrôlabilité locale) *On dit que le système de contrôle non-linéaire* (5.18) *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 (5.18), 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 $U = \mathbb{R}^k$. 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 $L^\infty([0,T];\mathbb{R}^k)$. 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.,

$$\dot{\delta x}(t) = A_u(t)\delta x(t) + B_u(t)\delta u(t), \quad \forall t \in [0,T], \qquad \delta x(0) = 0, \tag{5.20}$$

où pour tout $t \in [0,T]$,

$$A_u(t) = \frac{\partial f}{\partial x}(t, x_u(t), u(t)) \in \mathbb{R}^{d\times d}, \qquad B_u(t) = \frac{\partial f}{\partial u}(t, x_u(t), u(t)) \in \mathbb{R}^{d\times k}.$$

Définition 5.2.9 *Soit $f(x,u)$ une application de classe C^1 de $\mathbb{R}^d \times \mathbb{R}^k$ dans $\mathbb{R}^d$. On appelle matrice Jacobienne de f le couple de matrices $\left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial u}\right) \in \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times k}$ où*

$$\frac{\partial f}{\partial x}(x,u) = \left(\frac{\partial f_i}{\partial x_j}(x,u)\right)_{1\le i,j\le d}, \quad \frac{\partial f}{\partial u}(x,u) = \left(\frac{\partial f_i}{\partial u_j}(x,u)\right)_{1\le i\le d, 1\le j\le k}.$$

Autrement dit, les gradients des composantes de f sont rangés en ligne (et pas en colonne). Si $f(x,u) = Ax + Bu$, on retrouve bien $\frac{\partial f}{\partial x}(x,u) = A$ et $\frac{\partial f}{\partial u}(x,u) = B$.

Lemme 5.2.10 *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 $E'_{T,x_0}(u)$, une application linéaire continue de $L^\infty([0,T];\mathbb{R}^k)$ dans $\mathbb{R}^d$, est égale à 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*

$$\langle E'_{T,x_0}(u), \delta u\rangle = \delta x(T), \tag{5.21}$$

où δx est solution du système différentiel linéarisé (5.20).

Remarque 5.2.11 Dans le Lemme 5.2.10 la notion de différentielle de Fréchet est celle pour une application définie sur un espace de Banach $\mathcal{B}$ et la notation $\langle L, \delta u\rangle$ désigne le produit de dualité entre une application linéaire continue L (qui appartient au dual $\mathcal{B}'$) et un élément $\delta u \in \mathcal{B}$. La Définition 2.4.1 ne considérait que des applications définies sur un espace de Hilbert mais, comme expliqué dans la Remarque 2.4.2, il suffit de remplacer le produit scalaire de l'espace de Hilbert par le produit de dualité entre l'espace de Banach $\mathcal{B}$ et son dual $\mathcal{B}'$. •

Démonstration. Par souci de simplification on se contente d'esquisser la preuve sans donner le détail des vérifications des termes en $o(\delta u)$. Soit $\delta u \in \mathcal{B} = L^\infty([0,T];\mathbb{R}^k)$ tel que $u + \delta u \in \mathcal{U}_{T,x_0}$ (qui est ouvert dans $\mathcal{B}$). 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

$$\begin{aligned} E_{T,x_0}(u+\delta u) - E_{T,x_0}(u) &= x_{u+\delta u}(T) - x_u(T) \\ &= \delta x(T) + \epsilon(T) = \delta x(T) + o(\delta u), \end{aligned}$$

c'est-à-dire que, si elle existe, la différentielle $E'_{T,x_0}(u)$ est bien définie par la formule (5.21). Pour conclure, il faut montrer que $E'_{T,x_0}(u)$ est effectivement une forme linéaire continue, c'est-à-dire que $\delta u \mapsto \delta x(T)$ définit une forme linéaire continue pour la topologie de $\mathcal{B}$ La formule de la résolvante du Lemme 5.1.11 donne

$$\langle E'_{T,x_0}(u), \delta u\rangle = \delta x(T) = R(T)\int_0^T R(s)^{-1}B_u(s)\delta u(s)\,ds,$$

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 $R(0) =$ Id. On a donc bien $|\langle E'_{T,x_0}(u), \delta u\rangle| \leq C\|\delta u\|_{L^\infty([0,T];\mathbb{R}^k)}$. □

Théorème 5.2.12 (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. Sans rentrer dans les détails, 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 [22]). □

Théorème 5.2.13 (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 5.2.14 On considère le cas particulier d'un **point d'équilibre** d'un système différentiel autonome, i.e., un couple $(x_0, u_0) \in \mathbb{R}^d \times \mathbb{R}^k$ 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. •

Exemple 5.2.1 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 $(x_0, u_0) = ((0,0)^*, 0)$. •

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 un terme de dérive $f \in L^1([0,T];\mathbb{R}^d)$. Le système de contrôle linéaire autonome s'écrit sous la forme

$$\begin{cases} \dot{x}_u(t) = Ax_u(t) + Bu(t) + f(t), \quad \forall t \in [0,T], \\ x_u(0) = x_0. \end{cases} \tag{6.1}$$

L'ensemble des contrôles admissibles est ici l'espace de Hilbert

$$V = L^2([0,T];\mathbb{R}^k).$$

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 (voir le Lemme 5.1.3).

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

$$J(u) = \frac{1}{2}\int_0^T Ru(t)\cdot u(t)\,dt + \frac{1}{2}\int_0^T Qe_{x_u}(t)\cdot e_{x_u}(t)\,dt + \frac{1}{2}De_{x_u}(T)\cdot e_{x_u}(T), \quad (6.2)$$

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 :

$$\text{trouver } \overline{u} \in V \text{ tel que } J(\overline{u}) = \inf_{u\in V} J(u). \quad (6.3)$$

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 $L^2([0,T];\mathbb{R}^k)$. On n'impose pas ici d'atteindre exactement la cible au temps final T (ni à aucun temps intermédiaire) mais on applique plutôt une méthode de pénalisation (voir la Sous-section 3.4.3). Une illustration générale du problème de contrôle optimal LQ est présentée à la Figure 6.1.

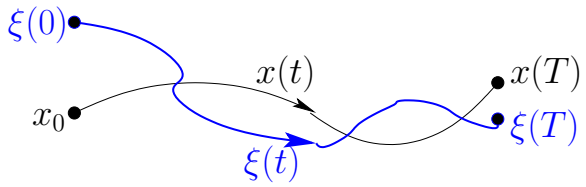

FIGURE 6.1 – Illustration du problème de contrôle optimal LQ : trajectoire cible et trajectoire optimale.

Remarque 6.1.1 Si on prend $Q = D = 0$ dans le critère (6.2), la solution (unique !) du problème (6.3) est alors triviale : $u \equiv 0$ sur $[0,T]$. •

Afin d'étudier les propriétés de la fonctionnelle J, il sera utile de décomposer

$$J(u) = J_R(u) + J_{QD}(u), \quad \forall u \in V,$$

avec

$$\begin{aligned} J_R(u) &= \frac{1}{2}\int_0^T Ru(t)\cdot u(t)\,dt, \\ J_{QD}(u) &= \frac{1}{2}\int_0^T Qe_{x_u}(t)\cdot e_{x_u}(t)\,dt + \frac{1}{2}De_{x_u}(T)\cdot e_{x_u}(T). \end{aligned} \quad (6.4)$$

Lemme 6.1.2 *La fonctionnelle J définie en* (6.2) *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, comme étudié à l'Exercice 2.3.5, la fonction $v \mapsto \frac{1}{2}Rv \cdot v$ est fortement convexe sur $\mathbb{R}^k$ avec $\alpha = \lambda_{\min}(R)$. L'intégration sur l'intervalle de temps $[0,T]$ ne change pas cette propriété puisque la matrice R est constante. Autrement dit, la fonctionnelle J_R est fortement convexe sur V de 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 $x_u(t) = e^{tA}x_0 + \int_0^t e^{(t-s)A}(Bu(s) + f(s))\,ds$, 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 semi-définies positives, l'application $y \mapsto \frac{1}{2}\int_0^T y(t)^* Q y(t)\,dt + \frac{1}{2}y(T)^* D y(T)$, définie de $C^0([0,T];\mathbb{R}^d)$ dans $\mathbb{R}$, est convexe (même raisonnement que ci-dessus avec l'utilisation de l'Exercice 2.3.5).

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* (6.3).

Démonstration. Il suffit de combiner le Théorème 2.3.9 (avec $K = V$) avec le Lemme 6.1.2. □

6.2 Différentielle du critère : état adjoint

6.2.1 Condition d'optimalité

L'objectif de cette sous-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 fonction objectif $J(u)$ est différentiable sur $V = L^2([0,T];\mathbb{R}^k)$ et, pour tout $u \in V$, sa dérivée directionnelle vaut, pour tout $\delta u \in L^2([0,T];\mathbb{R}^k)$,*

$$\langle J'(u), \delta u\rangle_V = \int_0^T Ru(t)\cdot\delta u(t)\,dt + \int_0^T Qe_{x_u}(t)\cdot\delta x(t)\,dt + De_{x_u}(T)\cdot\delta x(T) \quad (6.5)$$

avec δx la solution unique dans $AC([0,T];\mathbb{R}^d)$ de

$$\begin{cases} \dot{\delta x}(t) = A\delta x(t) + B\delta u(t) & \forall t \in [0,T], \\ \delta x(0) = 0. \end{cases} \quad (6.6)$$

Remarque 6.2.2 La formule de dérivée (6.5) est inutilisable en pratique ! Pour chaque δu il faut calculer δx pour obtenir une seule dérivée directionnelle. Autrement dit, on n'a pas une formule pour $J'(u)$ mais seulement pour une de ses composantes. •

Démonstration. On rappelle que $V = L^2([0,T];\mathbb{R}^k)$ est un espace de Hilbert pour le produit scalaire

$$\langle f, g\rangle_V = \int_0^T f(t)\cdot g(t)\,dt.$$

Comme $J = J_R + J_{QD}$ en vertu de (6.4), nous allons considérer séparément la différentiabilité des fonctionnelles J_R et J_{QD}.

La différentiabilité de J_R est immédiate puisque, en utilisant la symétrie de la matrice R, il vient, pour toute perturbation du contrôle $\delta u \in V$,

$$\begin{aligned} J_R(u+\delta u) &= \frac{1}{2}\int_0^T R(u(t)+\delta u(t))\cdot(u(t)+\delta u(t))\,dt \\ &= J_R(u) + \int_0^T Ru(t)\cdot\delta u(t)\,dt + J_R(\delta u) \\ &= J_R(u) + \langle Ru, \delta u\rangle_V + J_R(\delta u). \end{aligned}$$

Comme $\frac{J_R(\delta u)}{\|\delta u\|_V} \leq \frac{1}{2}\lambda_{\max}(R)\|\delta u\|_V$, où $\lambda_{\max}(R)$ désigne la plus grande valeur propre de R, on conclut que $J_R'(u) = Ru \in V$.

Pour différentier J_{QD}, on considère la trajectoire perturbée $x_{u+\delta u}$, associée au contrôle perturbé $u+\delta u$. Comme l'application $u \mapsto x_u$ est affine de V dans $AC([0,T];\mathbb{R}^d)$, en soustrayant l'équation pour x_u de celle pour $x_{u+\delta u}$, on a $x_{u+\delta u} = x_u + \delta x$ avec δx solution dans $AC([0,T];\mathbb{R}^d)$ de (6.6). La perturbation de la trajectoire δx est donc linéaire en δu et, par la formule explicite de la solution $\delta x(t) = \int_0^t e^{(t-s)A}B\delta u(s)ds$, on a

$$\|\delta x\|_{C^0([0,T];\mathbb{R}^d)} \leq C\|\delta u\|_V$$

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

$$\begin{aligned} J_{QD}(u+\delta u) = J_{QD}(u) &+ \int_0^T Qe_{x_u}(t)\cdot\delta x(t)\,dt + De_{x_u}(T)\cdot\delta x(T) \\ &+ \frac{1}{2}\int_0^T Q\delta x(t)\cdot\delta x(t)\,dt + \frac{1}{2}D\delta x(T)\cdot\delta x(T), \end{aligned}$$

ce qui montre que

$$\langle J_{QD}'(u), \delta u\rangle_V = \int_0^T Qe_{x_u}(t)\cdot\delta x(t)\,dt + De_{x_u}(T)\cdot\delta x(T), \tag{6.7}$$

d'où la formule (6.5). Remarquons qu'au membre de droite de (6.7), 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. □

Pour améliorer le Lemme 6.2.1 et obtenir une formule explicite de la dérivée $J'(u)$, on introduit la notion d'état adjoint qui joue le rôle d'un multiplicateur de Lagrange pour le problème d'optimisation (6.3). En optimisation sous contrainte de modèle, on a déjà rencontré cette notion d'état adjoint dans un cadre un peu plus simple, voir la Sous-section 3.6.2.

Définition 6.2.3 *L'**état adjoint** p associé au système LQ, c'est-à-dire au système de contrôle linéaire* (6.1) *et au critère* (6.2)*, est la solution unique dans $C^1([0,T];\mathbb{R}^d)$ de l'équation différentielle rétrograde en temps*

$$\begin{cases} \dot{p}(t) = -A^*p(t) - Qe_{x_u}(t), \quad \forall t \in [0,T], \\ p(T) = De_{x_u}(T). \end{cases} \tag{6.8}$$

où $x_u \in AC([0,T];\mathbb{R}^d)$ est la trajectoire associée au contrôle u, solution de (6.1)*.*

Rappelons que $e_{x_u} = x_u - \xi$ où $\xi(t)$ est la trajectoire cible. On vérifie aisément, puisque ξ et x_u sont continues en temps, que (6.8) admet bien une unique solution de classe C^1.

Remarque 6.2.4 L'état adjoint p ne vérifie pas une condition initiale mais une condition finale en T. Par ailleurs, dans la littérature, la convention est parfois de définir l'état adjoint comme un vecteur ligne $\widehat{p} = p^*$. 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 $\widehat{p}(T) = e_{x_u}(T)^*D$. •

Grâce à la notion d'état adjoint on obtient une formule simple pour la dérivée de la fonction objectif qui améliore sensiblement le Lemme 6.2.1 qui ne donnait qu'une composante ou dérivée directionnelle.

Proposition 6.2.5 *La fonctionnelle J est différentiable sur V et, pour tout $u \in V$, sa dérivée $J'(u) \in V$ est donnée par*

$$J'(u) = Ru + B^*p\,,$$

*où p est l'**état adjoint**, solution unique dans $C^1([0,T];\mathbb{R}^d)$ de* (6.8)*.*

Démonstration. Dans le Lemme 6.2.1 on a déjà montré que $J'_R(u) = Ru$. Comme $J = J_R + J_{QD}$, il reste à montrer que $J'_{QD}(u) = B^*p$ pour conclure. Autrement dit, dans la formule (6.7) il faut éliminer la perturbation de la trajectoire δx et la remplacer par la perturbation du contrôle δu. Pour cela on utilise l'état adjoint p solution de (6.8). On multiplie l'équation (6.6) (qui donne δx) par p et l'équation (6.8) (qui donne p) par δx et on additionne pour obtenir

$$\frac{d}{dt}(p \cdot \delta x) = -A^*p \cdot \delta x - Qe_{x_u} \cdot \delta x + p \cdot A\delta x + p \cdot B\delta u = -Qe_{x_u} \cdot \delta x + B^*p \cdot \delta u,$$

où les termes contenant A s'éliminent (l'adjoint est construit pour cela). On intègre alors en temps : comme $\delta x(0) = 0$, et prenant en compte la condition finale pour l'adjoint, on obtient

$$\int_0^T \frac{d}{dt}(p \cdot \delta x)\,dt = p(T) \cdot \delta x(T) = De_{x_u}(T) \cdot \delta x(T).$$

Par conséquent

$$De_{x_u}(T) \cdot \delta x(T) = \int_0^T B^*p(t) \cdot \delta u(t)\,dt - \int_0^T Qe_{x_u}(t) \cdot \delta x(t)\,dt$$

et la formule (6.7) pour $J'_{QD}(u)$ se simplifie en

$$\begin{aligned}\langle J'_{QD}(u), \delta u\rangle_V &= \int_0^T Qe_x(t)\cdot\delta x(t)\,dt + De_x(T)\cdot\delta x(T)\\ &= \int_0^T B^*p(t)\cdot\delta u(t)\,dt = \langle B^*p, \delta u\rangle_V.\end{aligned}$$

En conclusion, on a montré que $J'_{QD}(u) = B^*p$, ce qui conclut la preuve. □

Remarque 6.2.6 La Proposition 6.2.5 permet de mettre en oeuvre une méthode numérique de calcul du contrôle optimal. Il s'agit simplement d'appliquer un algorithme de gradient (sans contrainte) au problème de minimisation (fortement convexe) (6.3). L'algorithme (3.8) du gradient à pas fixe $\mu > 0$ s'écrit alors :

1. Initialisation : $u_0 \in L^2([0,T];\mathbb{R}^k)$, par exemple $u_0 = 0$.
2. Itérations $n \geq 0$:
   (a) calculer x_n en fonction de u_n,
   (b) calculer p_n en fonction de x_n et u_n,
   (c) mise à jour du contrôle $u_{n+1} = u_n - \mu(Ru_n + B^*p_n)$.

Dans ce qui précède, les fonctions $u_n(t), x_n(t), p_n(t)$ sont discrétisées en temps, c'est-à-dire qu'elles sont constantes sur des sous-intervalles $[t_{k-1}, t_k[$, $1 \leq k \leq k_T$, avec un (grand) entier $k_T \in \mathbb{N}$ et un (petit) pas de temps $\Delta t = T/k_T$ tel que $t_k = k\Delta t$. Les fonctions $x_n(t), p_n(t)$ sont calculées numériquement à l'aide d'un schéma de type différences finies, par exemple le schéma d'Euler implicite. Dans ce cas, si on note $x_n^k = x_n(t_k)$ et $u_n^k = u_n(t_k)$, on calcule :

• $x_n^0 = x_0$ et pour k croissant de 1 à k_T

$$\frac{x_n^k - x_n^{k-1}}{\Delta t} = Ax_n^k + Bu_n^k + f(t_k).$$

On stocke la solution $(x_n^k)_{0\leq k\leq k_T}$ puis on calcule $p_n^k = p_n(t_k)$:

• $p_n^{k_T} = D(x_n^{k_T} - \xi(T))$ et pour k décroissant de $k_T - 1$ à 0 (équation rétrograde)

$$\frac{p_n^{k+1} - p_n^k}{\Delta t} = -A^*p_n^k - Q(x_n^k - \xi(t_k)).$$

•

Théorème 6.2.7 (CNS d'optimalité) *Le contrôle $\overline{u} \in V$ est optimal pour le problème LQ* (6.3) ***si et seulement si*** *on a*

$$\overline{u}(t) = -R^{-1}B^*\overline{p}(t) \quad \forall t \in [0,T], \tag{6.9}$$

où l'état adjoint $\overline{p}(t)$ est la solution unique dans $C^1([0,T];\mathbb{R}^d)$ du système adjoint (6.8) *calculé pour la trajectoire $\overline{x} = x_{\overline{u}}$ associée au contrôle optimal $\overline{u}$. Autrement dit,*

$$\frac{d\overline{p}}{dt}(t) = -A^*\overline{p}(t) - Qe_{\overline{x}}(t), \quad \forall t \in [0,T], \qquad \overline{p}(T) = De_{\overline{x}}(T), \tag{6.10}$$

avec $e_{\overline{x}} = \overline{x} - \xi$ *et*

$$\frac{d\overline{x}}{dt}(t) = A\overline{x}(t) + B\overline{u}(t) + f(t), \quad \forall t \in [0,T], \qquad \overline{x}(0) = x_0. \tag{6.11}$$

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 (6.9) résultant de la convexité de la fonctionnelle J. □

Remarque 6.2.8 Le Théorème 6.2.7 fournit une méthode pour calculer le contrôle optimal $\overline{u}$ du problème LQ. En effet, une fois éliminé $\overline{u}$ grâce à la formule (6.9), on obtient 2 systèmes d'équations différentielles ordinaires linéaires couplées à résoudre pour obtenir $(\overline{x}, \overline{p})$ et donc, in fine, $\overline{u}$. La difficulté est qu'il s'agit d'un problème "aux deux bouts", c'est-à-dire muni d'une donnée initiale pour $\overline{x}$ et d'une donnée finale pour $\overline{p}$. En particulier, c'est un problème difficile à résoudre numériquement (il faut avoir recours à une méthode de tir, voir [33]). •

Remarque 6.2.9 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 $\overline{u} \in C^1([0,T]; \mathbb{R}^k)$. Comme $\overline{u}$ est continu, il n'y a pas ici de phénomène de commutation pour le contrôle optimal (cf. Exemple 5.1.3). •

Exercice 6.2.1 On sait déjà par le Corollaire 6.1.3 que le contrôle optimal $\overline{u}$ est unique, donc la trajectoire $\overline{x}$ et l'adjoint correspondant $\overline{p}$ aussi. Néanmoins il est instructif de montrer directement l'unicité de l'extrémale. Montrer qu'une extrémale $(\overline{x}, \overline{p}, \overline{u})$, définie par le Théorème 6.2.7, est unique.

Exemple 6.2.1 On considère un point matériel qui peut se déplacer sur une droite et dont on contrôle la vitesse (cf. l'Exemple 5.1.3). Le système de contrôle linéaire s'écrit, avec $d = k = 1$,

$$\dot{x}_u(t) = u(t), \quad \forall t \in [0,T], \qquad x_u(0) = x_0.$$

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 6.2.7, on déduit que le contrôle optimal est

$$\overline{u}(t) = -\overline{p}(t),$$

où l'état adjoint est solution de

$$\frac{d\overline{p}}{dt}(t) = -\overline{x}(t), \quad \forall t \in [0,T], \qquad \overline{p}(T) = 0.$$

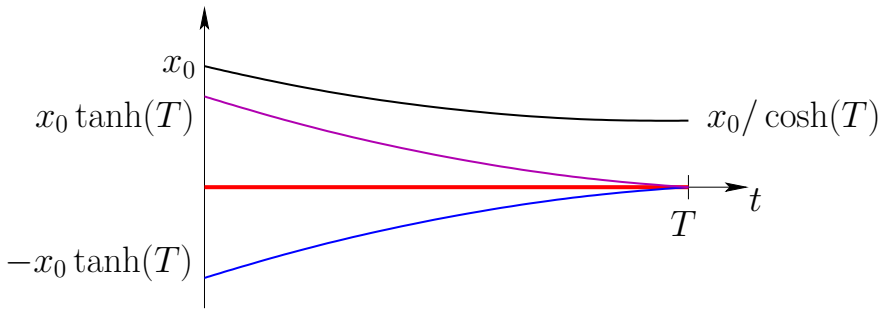

FIGURE 6.2 – Illustration de l'extrémale obtenue à l'Exemple 6.2.1 (mouvement d'un point matériel) : trajectoire $\overline{x}(t)$, état adjoint $\overline{p}(t)$, contrôle optimal $\overline{u}(t)$; la cible $\xi(t)$ est identiquement nulle.

On 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

$$\begin{aligned}\overline{x}(t) &= x_0\cosh(t) - \overline{p}(0)\sinh(t),\\ \overline{p}(t) &= -x_0\sinh(t) + \overline{p}(0)\cosh(t).\end{aligned}$$

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. •

Exercice 6.2.2 On considère à nouveau un point matériel en mouvement rectiligne, mais on contrôle l'accélération de ce point par un contrôle à valeurs dans $\mathbb{R}$:

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

Le critère à minimiser dans $V = L^2([0,T];\mathbb{R})$ est

$$J(u) = -x(T) + \frac{1}{2}\int_0^T u(t)^2\,dt,$$

c'est-à-dire que l'on veut maximiser la distance parcourue et minimiser la norme de l'accélération. Calculer l'adjoint et le contrôle optimal. En déduire que la distance optimale est $x(T) = T^3/6$.

Exercice 6.2.3 On considère le système de l'oscillateur harmonique pour un point matériel $x(t)$ à valeurs dans $\mathbb{R}$:

$$\ddot{x}(t) + \omega^2 x(t) = u(t),\ \forall t \in [0,T], \quad x(0) = x_0,\ \dot{x}(0) = v_0,$$

où le contrôle $u(t) \in \mathbb{R}$ est la force appliquée. Le critère à minimiser dans $V = L^2([0,T];\mathbb{R})$ est

$$J(u) = \frac{1}{2}x(T)^2 + \frac{1}{2}\dot{x}(T)^2 + \frac{R}{2}\int_0^T u(t)^2\,dt,$$

c'est-à-dire que l'on veut stabiliser vers zéro le point au temps final et minimiser la norme de l'accélération avec un poids $R > 0$. Calculer l'adjoint et le contrôle optimal. En déduire que la position finale optimale vérifie $x(T) = \mathcal{O}(T^{-1})$ et $\dot{x}(T) = \mathcal{O}(T^{-1})$ pour T grand.

6.2.2 Sur l'origine de l'état adjoint

De la manière dont est présentée la Proposition 6.2.5 l'état adjoint semble être une "astuce" de calcul qui permet de calculer, de manière un peu miraculeuse, la différentielle du critère $J(u)$. En particulier, la Définition 6.2.3 du système adjoint semble "tombée du ciel" et son interprétation n'est pas évidente. Le but de cette sous-section est d'expliquer comment on trouve la définition de l'état adjoint et de montrer en fait qu'il s'agit d'un multiplicateur de Lagrange. Cette interprétation est très proche du contenu de la Sous-section 3.6.2 qui était consacré à l'optimisation sous contrainte de modèle.

Le point de départ est de réécrire le problème d'optimisation (6.3) comme

$$\min_{\substack{u\in L^2([0,T];\mathbb{R}^k),\ x\in AC([0,T];\mathbb{R}^d)\\ C(u,x)=0}} \tilde{J}(u,x),$$

où u et x sont deux variables indépendantes, la nouvelle fonction objectif est

$$\tilde{J}(u,x) = \frac{1}{2}\int_0^T Ru(t)\cdot u(t)\,dt + \frac{1}{2}\int_0^T Qe_x(t)\cdot e_x(t)\,dt + \frac{1}{2}De_x(T)\cdot e_x(T), \quad (6.12)$$

avec $e_x = x - \xi$ et $C(u,x) = 0$ est la **contrainte** qui relie x à u, c'est-à-dire

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

Comme on l'a vu à de multiples reprises en optimisation, il est naturel d'introduire un Lagrangien qui, par construction (voir la Définition 2.5.9), est la somme de la fonction objectif et des contraintes multipliées par des multiplicateurs de Lagrange. La nouveauté ici est que la contrainte, notée $C(u,x)$, n'est pas à valeur vectorielle dans $\mathbb{R}^M$ mais est une fonction de t qui doit s'annuler sur $[0,T]$, en plus de la donnée initiale. Par conséquent, les résultats du cours ne s'appliquent pas en toute rigueur. On admettra néanmoins la généralisation qui suit et qui peut, bien sûr, se justifier avec un peu plus de travail. Puisque la contrainte est une fonction du temps, l'idée

est d'avoir un multiplicateur de Lagrange $p(t)$ qui est aussi une fonction du temps et de remplacer la sommation sur les différentes contraintes par une intégrale en temps. Autrement dit, pour 3 variables indépendantes $(u,x,p) \in L^2([0,T];\mathbb{R}^k) \times AC([0,T];\mathbb{R}^d) \times AC([0,T];\mathbb{R}^d)$, on définit le Lagrangien

$$\mathcal{L}(u,x,p) = \tilde{J}(u,x) - \int_0^T p \cdot (\dot{x} - Ax - Bu - f)\, dt - p(0) \cdot (x(0) - x_0)\,. \qquad (6.13)$$

On vérifie facilement que

$$\max_{p \in AC([0,T];\mathbb{R}^d)} \mathcal{L}(u,x,p) = \begin{cases} \tilde{J}(u,x_u) = J(u) & \text{si } x = x_u, \\ +\infty & \text{si } x \neq x_u, \end{cases}$$

ce qui montre que la définition (6.13) du Lagrangien est consistante, au sens où sa maximisation par rapport à p redonne la fonction objectif à minimiser sous contrainte (cf. le Lemme 2.5.10).

Nous allons maintenant calculer les trois dérivées partielles du Lagrangien (6.13), par rapport à (u,x,p), qui vont permettre de retrouver la définition de l'équation d'état (le système de contrôle), celle de l'équation adjointe et finalement la dérivée de la fonction objectif. Notons au passage que $AC([0,T];\mathbb{R}^d)$ n'est pas un espace de Hilbert, mais un espace de Banach ce qui nécessite de généraliser la Définition 2.4.1 de différentiation en suivant les explications de la Remarque 2.4.2. On note $\langle \cdot, \cdot \rangle$ le produit de dualité entre $AC([0,T];\mathbb{R}^d)$ et son dual.

Un calcul facile, car $p \mapsto \mathcal{L}(u,x,p)$ est affine, montre que, pour tout $\phi \in AC([0,T];\mathbb{R}^d)$, une première dérivée partielle du Lagrangien est donnée par

$$\langle \frac{\partial \mathcal{L}}{\partial p}(u,x,p), \phi \rangle = -\int_0^T \phi \cdot (\dot{x} - Ax - Bu - f)\, dt - \phi(0) \cdot (x(0) - x_0)\,. \qquad (6.14)$$

On en déduit que si (6.14) est nul pour toute fonction ϕ, alors $x = x_u$ est la solution unique du système de contrôle (6.1). Ce n'est pas une surprise car le Lagrangien (6.13) est précisément construit pour cela. Etudions maintenant une deuxième dérivée partielle du Lagrangien.

Lemme 6.2.10 *La fonction $x \mapsto \mathcal{L}(u,x,p)$ est différentiable sur $AC([0,T];\mathbb{R}^d)$ et, pour tout $\phi \in AC([0,T];\mathbb{R}^d)$, sa dérivée directionnelle vaut*

$$\langle \frac{\partial \mathcal{L}}{\partial x}(u,x,p), \phi \rangle = \int_0^T \phi \cdot (\dot{p} + A^* p + Qe_x)\, dt + \phi(T) \cdot (De_x(T) - p(T)). \qquad (6.15)$$

Par conséquent, si (6.15) *est nul pour toute fonction ϕ, alors p est la solution unique de l'équation adjointe* (6.8).

Remarque 6.2.11 Le Lemme 6.2.10 indique comment trouver le système adjoint de manière générale. On construit un Lagrangien $\mathcal{L}(u,x,p)$ du problème de contrôle puis on calcule sa dérivée par rapport à l'état x. Lorsque cette dérivée partielle s'annule, on trouve le système d'équations différentielles ordinaires vérifié par l'état adjoint p. •

Démonstration. On commence par intégrer par parties dans le membre de droite de (6.13)

$$\mathcal{L}(u,x,p) = \tilde{J}(u,x) + \int_0^T x\cdot(\dot{p}+A^*p)\,dt - p(T)\cdot x(T) + p(0)\cdot x_0 + \int_0^T p\cdot(Bu+f)\,dt,$$

de façon à éliminer la dérivée en temps $\dot{x}$. On dérive ensuite par rapport à x et, tenant compte de (6.12), pour tout $\phi \in AC([0,T];\mathbb{R}^d)$, on trouve

$$\langle \frac{\partial \mathcal{L}}{\partial x}(u,x,p), \phi\rangle = \int_0^T \phi\cdot(\dot{p}+A^*p+Qe_x)\,dt + (De_x(T)-p(T))\cdot\phi(T)$$

qui est précisément (6.15). Si on prend d'abord ϕ quelconque, mais s'annulant en T, on trouve que $\dot{p}+A^*p+Qe_x = 0$ p.p. dans $[0,T]$. Puis, prenant $\phi(T)$ quelconque, on obtient que $De_x(T)-p(T)=0$. Autrement dit, p est solution de l'équation adjointe (6.8). □

On peut pousser un peu plus loin l'analyse et calculer enfin la troisième dérivée partielle du Lagrangien.

Lemme 6.2.12 *La fonction $u \mapsto \mathcal{L}(u,x,p)$ est différentiable sur $L^2([0,T];\mathbb{R}^k)$ et sa différentielle vaut*

$$\frac{\partial \mathcal{L}}{\partial u}(u,x,p) = Ru + B^*p. \tag{6.16}$$

Autrement dit, la dérivée partielle du Lagrangien (6.13) *par rapport au contrôle u est précisément égale à la dérivée $J'(u)$ du critère de contrôle, si p est l'adjoint, solution de* (6.8).

Remarque 6.2.13 Cette façon de trouver l'adjoint et la dérivée du critère à partir du Lagrangien est un résultat profond (mais délicat) qui se généralise dans de nombreuses situations en théorie du contrôle optimal. On peut montrer que le résultat du Lemme 6.2.12 n'est pas un hasard (voir [2]). C'est en quelque sorte une généralisation du Corollaire 2.5.12 qui affirmait que les conditions d'optimalité d'un problème d'optimisation sous contrainte sont équivalentes à la stationnarité du Lagrangien. •

Démonstration. En tenant compte de la définition (6.12) de $\tilde{J}(u,x)$ dans la formule du Lagrangien, un simple calcul donne, pour tout $\delta u \in L^2([0,T];\mathbb{R}^k)$,

$$\langle \frac{\partial \mathcal{L}}{\partial u}(u,x,p), \delta u\rangle = \int_0^T Ru\cdot\delta u\,dt + \int_0^T p\cdot B\delta u\,dt,$$

d'où la formule (6.16) qui coincide avec celle pour $J'(u)$ obtenue à la Proposition 6.2.5. □

6.3 Principe du minimum : Hamiltonien

L'objectif de cette section est de reformuler le Théorème 6.2.7 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* (6.2) *est l'application $H : [0,T] \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$ telle que*

$$H(t,x,p,u) = p \cdot (Ax + Bu + f(t)) + \frac{1}{2} Ru \cdot u + \frac{1}{2} Q(x - \xi(t)) \cdot (x - \xi(t)).$$

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 des vecteurs colonnes) montre que

$$\begin{aligned}
\tfrac{\partial H}{\partial x}(t,x,p,u) &= A^* p + Q(x - \xi(t)),\\
\tfrac{\partial H}{\partial p}(t,x,p,u) &= Ax + Bu + f(t),\\
\tfrac{\partial H}{\partial u}(t,x,p,u) &= B^* p + Ru.
\end{aligned}$$

Autrement dit, la première dérivée partielle $\frac{\partial H}{\partial x}$ redonne, au signe près, le second membre de l'équation adjointe (6.8), tandis que la seconde dérivée partielle $\frac{\partial H}{\partial p}$ redonne le second membre du système de contrôle (6.1) et qu'enfin la troisième dérivée partielle $\frac{\partial H}{\partial u}$ redonne la formule pour $J'(u)$ de la Proposition 6.2.5 (on a vu des calculs similaires pour le Lagrangien dans la Sous-section 6.2.2 et la Remarque 7.2.7 discutera du lien entre Lagrangien et Hamiltonien). Si maintenant on considère l'extrémale $(\overline{x}, \overline{p}, \overline{u})$ obtenue au Théorème 6.2.7, on constate que

$$\frac{\partial H}{\partial u}(t, \overline{x}(t), \overline{p}(t), \overline{u}(t)) = 0. \tag{6.17}$$

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 (6.17) 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

$$\begin{aligned}
\frac{d\overline{x}}{dt}(t) &= \frac{\partial H}{\partial p}(t, \overline{x}(t), \overline{p}(t), \overline{u}(t)) = A\overline{x}(t) + B\overline{u}(t) + f(t), \quad \overline{x}(0) = x_0,\\
\frac{d\overline{p}}{dt}(t) &= -\frac{\partial H}{\partial x}(t, \overline{x}(t), \overline{p}(t), \overline{u}(t)) = -A^*\overline{p}(t) - Qe_{\overline{x}}(t), \qquad \overline{p}(T) = De_{\overline{x}}(T),
\end{aligned}$$

où $e_{\overline{x}}(t) = \overline{x}(t) - \xi(t)$.

Remarque 6.3.3 On aurait pu définir $\widehat{H} = -H$ et aboutir à un principe du maximum pour $\widehat{H}$ (ce qui est le cas parfois dans la littérature). •

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 on 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

$$\begin{aligned}\frac{d\mathcal{H}}{dt}(t) &= \frac{\partial H}{\partial x}\cdot\frac{d\overline{x}}{dt}(t) + \frac{\partial H}{\partial p}\cdot\frac{d\overline{p}}{dt}(t) + \frac{\partial H}{\partial u}\cdot\frac{d\overline{u}}{dt}(t)\\ &= -\frac{d\overline{p}}{dt}(t)\cdot\frac{d\overline{x}}{dt}(t) + \frac{d\overline{x}}{dt}(t)\cdot\frac{d\overline{p}}{dt}(t) + 0 = 0,\end{aligned}$$

ce qui conclut la preuve. □

Exemple 6.3.1 On reprend l'Exemple 6.2.1 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 $J(u) = \frac{1}{2}\int_0^T x_u(t)^2\,dt + \frac{1}{2}\int_0^T u(t)^2\,dt$. 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 d'introduire l'équation de Riccati dont la résolution permet 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***

$$\dot{P}(t) = -A^*P(t) - P(t)A + P(t)BR^{-1}B^*P(t) - Q, \ \forall t \in [0,T], \quad P(T) = D, \ (6.19)$$

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

$$J(\overline{u}) = \frac{1}{2}x_0^*P(0)x_0.$$

Remarque 6.4.2 La solution $P(t)$ de l'équation de Ricatti permet donc d'exprimer le contrôle optimal comme un feedback (ou rétroaction) de l'état $\overline{x}(t)$, sous la forme $\overline{u}(t) = K(t)\overline{x}(t)$ avec $K(t) = -R^{-1}B^*P(t)$. Dans un tel cas, on parle de contrôle en boucle fermée car le contrôle dépend de l'état à l'instant présent t et pas de l'historique sur l'intervalle $[0,t]$. En effet, la matrice $P(t)$ est calculée de manière rétrograde à partir du temps final T et ne dépend pas de ce qui s'est passé sur $[0,t]$. Au contraire, la formule précédente $\overline{u}(t) = -R^{-1}B^*\overline{p}(t)$ est appelée contrôle en boucle ouverte car il faut d'abord calculer l'état sur tout l'intervalle $[0,T]$, puis résoudre l'adjoint rétrograde de T à t. Autrement dit, en boucle ouverte le contrôle optimal est calculé à l'avance, indépendamment de l'état réel du système. Le contrôle en boucle fermée est une approche très importante en pratique car en cas de perturbations ou d'erreurs (de modèle, de données ou numériques dans les résolutions d'équations) il permet de corriger le contrôle en prenant en compte la valeur "mesurée" (et pas prédite) de l'état. C'est un gage de robustesse qui explique la préférence en ingénierie pour une boucle fermée, plutôt qu'ouverte, de contrôle. •

Démonstration. (1) Dépendance linéaire. Comme le problème LQ admet un unique contrôle optimal $\overline{u}$, après l'avoir remplacé par sa formule (6.9), 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 $\mathcal{X}(0) = \mathrm{Id}$.
(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 $\mathcal{X}(0) = \mathrm{Id}$. De plus, on a vu que

$$\frac{d}{dt}\big(\overline{p}(t)\cdot\overline{x}(t)\big) = -Q\overline{x}(t)\cdot\overline{x}(t) - \overline{p}(t)\cdot(BR^{-1}B^*\overline{p}(t)). \tag{6.20}$$

En intégrant de s à T, et comme $\overline{x}(s) = 0$, il vient

$$0 = D\overline{x}(T)\cdot\overline{x}(T) + \int_s^T \Big(Q\overline{x}(t)\cdot\overline{x}(t) + R^{-1}B^*\overline{p}(t)\cdot B^*\overline{p}(t)\Big)\,dt \geq 0.$$

Les matrices D, Q, R étant symétriques (semi-)définies positives et R définie positive, on en déduit que $B^*\overline{p}(t) = 0$ pour tout $t \in [s,T]$ et donc que $\overline{u}(t) = -R^{-1}B^*\overline{p}(t) = 0$ sur $[s,T]$. On en déduit que

$$\frac{d\overline{x}}{dt}(t) = A\overline{x}(t) \text{ sur } [s,T], \qquad \overline{x}(s) = 0,$$

d'où $\overline{x}(t) = 0$ sur $[s,T]$. Ainsi on obtient

$$\frac{d\overline{p}}{dt}(t) = -A^*\overline{p}(t) \text{ sur } [s,T], \qquad \overline{p}(T) = D\overline{x}(T) = 0,$$

ce qui entraine $\overline{p}(t) = 0$ sur $[s,T]$. Par conséquent, $(\overline{x},\overline{p})$ vérifie un système différentiel linéaire (sans terme source) 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

$$P(t) = \mathcal{P}(t)\mathcal{X}(t)^{-1}, \quad \forall t \in [0,T],$$

qui vérifie $\overline{p}(t) = P(t)\overline{x}(t)$ et, par construction, $P \in C^1([0,T];\mathbb{R}^{d\times d})$. 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 $t \mapsto \overline{p}(t) \cdot \overline{x}(t)$ est décroissante d'après (6.20), il vient

$$P(t)x \cdot x = \overline{p}(t) \cdot \overline{x}(t) \geq \overline{p}(T) \cdot \overline{x}(T) = D\overline{x}(T) \cdot \overline{x}(T) \geq 0,$$

ce qui montre que $P(t)$ est semi-définie positive. Enfin, si $P(t)x \cdot x = 0$ et que 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. D'après (6.20) il vient

$$\begin{aligned}
J(\overline{u}) &= \frac{1}{2}\int_0^T \left(Q\overline{x}(t) \cdot \overline{x}(t) + R\overline{u}(t) \cdot \overline{u}(t)\right) dt + \frac{1}{2}D\overline{x}(T) \cdot \overline{x}(T) \\
&= \frac{1}{2}\int_0^T \left(Q\overline{x}(t) \cdot \overline{x}(t) - B^*\overline{p}(t) \cdot \overline{u}(t)\right) dt + \frac{1}{2}\overline{p}(T) \cdot \overline{x}(T) \\
&= \frac{1}{2}\int_0^T -\frac{d}{dt}\big(\overline{p}(t) \cdot \overline{x}(t)\big)\, dt + \frac{1}{2}\overline{p}(T) \cdot \overline{x}(T) \\
&= \frac{1}{2}\overline{p}(0) \cdot \overline{x}(0) = \frac{1}{2}P(0)\overline{x}(0) \cdot \overline{x}(0) = \frac{1}{2}P(0)x_0 \cdot x_0,
\end{aligned}$$

ce qui conclut la preuve. □

Remarque 6.4.3 Il est possible de remplacer l'équation non-linéaire (quadratique) de Riccati (6.19) par une équaltion **linéaire**, mais de plus grande taille. En effet, considérons le système différentiel linéaire suivant (de taille $2d$) :

$$\frac{d}{dt}\begin{pmatrix} x(t) \\ p(t) \end{pmatrix} = \underbrace{\begin{pmatrix} A & -BR^{-1}B^* \\ -Q & -A^* \end{pmatrix}}_{=\mathcal{A}\in\mathbb{R}^{(2d)\times(2d)}} \begin{pmatrix} x(t) \\ p(t) \end{pmatrix}$$

On note $\mathcal{R}(t) = e^{(T-t)\mathcal{A}}$ la résolvante associée à ce système différentiel (de dimension $2d$) et telle que $\mathcal{R}(T) = \mathrm{Id}$. On pose

$$\mathcal{R}(t) = \begin{pmatrix} R_1(t) & R_2(t) \\ R_3(t) & R_4(t) \end{pmatrix} \in \mathbb{R}^{(2d)\times(2d)},$$

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 vérifie aussi $P(t) = \mathcal{P}_T(t)\mathcal{X}_T(t)^{-1}$ et peut donc également se calculer en résolvant le système linéaire vérifiée par $\mathcal{R}(t)$, qui est de taille $4d^2$, donc plus grande que la taille de l'équation non-linéaire de Riccati (6.19) qui est $\frac{d(d+1)}{2}$ (car P est symétrique). Notons qu'il faut en plus réaliser l'inversion de la matrice $\mathcal{X}_T(t)$. Cette approche est intéressante en pratique car elle évite de devoir résoudre un système différentiel non-linéaire. •

Exemple 6.4.1 On reprend l'Exemple 6.2.1 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 $J(u) = \frac{1}{2}\int_0^T x_u(t)^2\,dt + \frac{1}{2}\int_0^T u(t)^2\,dt$. 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 $J(\overline{u}) = \frac{1}{2}x_0^2P(0) = \frac{1}{2}x_0^2\tanh(T)$. •

Exercice 6.4.1 On considère l'équation suivante pour un point matériel $x(t)$ à valeurs dans $\mathbb{R}$:

$$\dot{x}(t) = \frac{1}{2\alpha}x(t) + u(t),\ \forall t \in [0,T], \quad x(0) = x_0,$$

avec un contrôle $u(t) \in \mathbb{R}$ et un paramètre $\alpha > 0$. Le critère à minimiser dans $V = L^2([0,T];\mathbb{R})$ est

$$J(u) = \frac{1}{2}x(T)^2 + \frac{\alpha}{2}\int_0^T u(t)^2\,dt,$$

c'est-à-dire que l'on veut stabiliser vers zéro le point au temps final et minimiser la norme du contrôle. Montrer que la solution de l'équation de Ricatti est constante et vérifier que la trajectoire optimale est exponentiellement décroissante.

Pour finir, étudions une application de l'équation de Riccati au problème de l'asservissement ou de la poursuite. On revient au cas général en ne supposant plus la dérive et la cible nulles, $f(t) \neq 0$ et $\xi(t) \neq 0$. Par contre, on suppose que la cible $\xi(t)$ est une trajectoire du système sans contrôle, c'est-à-dire une solution de

$$\begin{cases} \dot{\xi}(t) = A\xi(t) + f(t) & \forall t \in [0,T], \\ \xi(0) = \xi_0. \end{cases}$$

On note la différence $z(t) = x(t) - \xi(t)$ qui vérifie

$$\begin{cases} \dot{z}(t) = Az(t) + Bu(t) & \forall t \in [0,T], \\ z(0) = z_0 = x_0 - \xi_0. \end{cases} \tag{6.21}$$

On considère la fonction objectif

$$J(u) = \frac{1}{2}\int_0^T Ru(t) \cdot u(t)\, dt + \frac{1}{2}\int_0^T Qz_u(t) \cdot z_u(t)\, dt + \frac{1}{2}Dz_u(T) \cdot z_u(T),$$

qui consiste à vouloir minimiser conjointement l'erreur entre la trajectoire $x(t)$ et la cible $\xi(t)$ et le coût du contrôle. Remarquons que les données initiales ne sont pas les mêmes. On souhaite donc suivre (ou poursuivre) une trajectoire cible, d'où le nom de ce type de problème. Bien entendu, puisque la dérive et la cible sont nulles pour $z(t)$, on peut lui appliquer le Théorème 6.4.1 et trouver le contrôle optimal par rétroaction (ou feedback) sous la forme

$$\overline{u}(t) = K(t)\big(\overline{x}(t) - \xi(t)\big),$$

avec $K(t) = -R^{-1}B^*P(t)$. La solution $P(t)$ de l'équation de Riccati a été préalablement calculée. Mais comme elle est indépendante à la fois de la donnée initiale x_0 et de la cible $\xi(t)$, elle est pertinente même si la cible change ou si l'état est perturbé par des causes extérieures. C'est ce qu'on appelle la robustesse dans cet exemple. L'écart à la cible apparait explicitement dans la formule du contrôle optimal et uniquement à l'instant présent t.

Remarque 6.4.4 Lorsqu'on applique le contrôle optimal $u(t) = K(t)z(t)$, donné par l'équation de Ricatti, au système (6.21), on obtient l'équation $\dot{z}(t) = (A + BK(t))z(t)$ qui est une réminiscence de la méthode du placement de pôles. Cette dernière consiste à trouver une matrice K telle que toutes les valeurs propres de $(A + BK)$ aient une partie réelle strictement négative, ce qui garantit bien que la différence $z(t)$ entre la trajectoire et la cible décroit exponentiellement en temps (voir [33]). •

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 horizon temporel $T > 0$, on considère un état à valeurs dans $\mathbb{R}^d$ 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

$$\mathcal{U} = L^1([0,T]; U) \subset L^1([0,T]; \mathbb{R}^k).$$

L'objectif est de trouver un contrôle optimal $\overline{u} \in \mathcal{U}$ qui minimise le critère

$$J(u) = \int_0^T g(t, x_u(t), u(t))\, dt + h(x_u(T)), \tag{7.2}$$

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 :

$$\text{trouver } \overline{u} \in \mathcal{U} \text{ tel que } J(\overline{u}) = \inf_{u \in \mathcal{U}} J(u). \tag{7.3}$$

Nous allons formuler quelques hypothèses (raisonnables mais parfois améliorables) sur les différents ingrédients intervenant dans la formulation du problème de contrôle optimal (7.3), 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

(h1) $f \in C^0([0,T] \times \mathbb{R}^d \times U; \mathbb{R}^d)$ et f est de classe C^1 par rapport à x ;

(h2) $\exists C$, $|f(t,y,v)|_{\mathbb{R}^d} \leq C(1 + |y|_{\mathbb{R}^d} + |v|_{\mathbb{R}^k})$, $\forall t \in [0,T]$, $\forall y \in \mathbb{R}^d$, $\forall v \in U$;

(h3) 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, l'unicité et la non-explosion sur tout l'intervalle $[0,T]$ 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* (h1), (h2), (h3) *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 8.3.4). 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 (h3), on a bien $C_0 \in L^1([0,T]; \mathbb{R}_+)$. En outre, la fonction F est localement intégrable grâce à l'hypothèse (h2) 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 $x(t) = x_0 + \int_0^t f(s,x(s),u(s))\,ds$, on peut appliquer ce lemme avec $z(t) = |x(t)|_{\mathbb{R}^d}$ et $\psi(t) \equiv C$.

L'estimation (7.4) est satisfaite avec $\alpha = |x_0|_{\mathbb{R}^d} + C(T + \|u\|_{L^1([0,T];\mathbb{R}^k)})$ grâce à l'hypothèse (h2). 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*

$$\exists \alpha \geq 0, \quad \forall t \in [0,T], \quad z(t) \leq \alpha + \int_0^t \psi(s) z(s)\, ds. \tag{7.4}$$

Alors, on a $z(t) \leq \alpha e^{\int_0^t \psi(s)\, ds}$ pour tout $t \in [0,T]$.

Démonstration. Posons $\Psi(t) = \int_0^t \psi(s)\, ds$ et $v(t) = e^{-\Psi(t)} \int_0^t \psi(s) z(s)\, ds$. En utilisant (7.4), on constate que

$$\begin{aligned} \frac{dv}{dt}(t) &= -\psi(t) e^{-\Psi(t)} \int_0^t \psi(s) z(s)\, ds + e^{-\Psi(t)} \psi(t) z(t) \\ &= \psi(t) e^{-\Psi(t)} \left(z(t) - \int_0^t \psi(s) z(s)\, ds \right) \leq \alpha \psi(t) e^{-\Psi(t)}. \end{aligned}$$

Comme $v(0) = 0$ et $\Psi(0) = 0$, en intégrant cette majoration de 0 à t, il vient

$$e^{-\Psi(t)} \int_0^t \psi(s) z(s)\, ds = v(t) \leq \alpha \int_0^t \psi(s) e^{-\Psi(s)}\, ds = \alpha (1 - e^{-\Psi(t)}),$$

et en ré-arrangeant les termes, on obtient

$$\alpha + \int_0^t \psi(s) z(s)\, ds \leq \alpha e^{\Psi(t)}.$$

On conclut en utilisant à nouveau la borne (7.4) sur $z(t)$. □

Venons en maintenant aux hypothèses sur le critère. On suppose que

(h4) $g \in C^0([0,T] \times \mathbb{R}^d \times U; \mathbb{R})$ et g est de classe C^1 par rapport à x ; de plus, $h \in C^1(\mathbb{R}^d; \mathbb{R})$;

(h5) pour tout $R > 0$, $\exists C_R$, $|g(t,y,v)| \leq C_R(1 + |v|_{\mathbb{R}^k})$, $\forall t \in [0,T]$, $\forall y \in \overline{B}(0,R)$, $\forall v \in U$;

(h6) pour tout $R > 0$, $\exists C_R$, $|\frac{\partial g}{\partial x}(t,y,v)|_{\mathbb{R}^d} \leq C_R(1 + |v|_{\mathbb{R}^k})$, $\forall t \in [0,T]$, $\forall y \in \overline{B}(0,R)$, $\forall v \in U$;

(h7) 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 (h5), 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 (h7). Il est donc raisonnable de considérer le problème de minimisation (7.3). Les hypothèses (h4) et (h6) 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 (7.2). 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* (7.2) *est l'application $H : [0,T] \times \mathbb{R}^d \times \mathbb{R}^d \times U \to \mathbb{R}$ définie par*

$$H(t,x,p,u) = p \cdot f(t,x,u) + g(t,x,u). \tag{7.5}$$

*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**.*

Définition 7.2.2 *L'**état adjoint** $p \in AC([0,T];\mathbb{R}^d)$ associé au système de contrôle non-linéaire* (7.1) *et à la fonctionnelle J définie en* (7.2) *est la solution de*

$$\dot{p}(t) = -A(t)^* p(t) - b(t),\ \forall t \in [0,T], \quad p(T) = \frac{\partial h}{\partial x}(x(T)) \in \mathbb{R}^d, \tag{7.6}$$

où pour tout $t \in [0,T]$,

$$A(t) = \frac{\partial f}{\partial x}(t,x(t),u(t)) \in \mathbb{R}^{d\times d}, \qquad b(t) = \frac{\partial g}{\partial x}(t,x(t),u(t)) \in \mathbb{R}^d,$$

et $x = x_u \in AC([0,T];\mathbb{R}^d)$ est la trajectoire associée au contrôle u. (La matrice $A(t)$ est la Jacobienne de f, voir la Définition 5.2.9. 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 p est solution d'un système linéaire (à (x,u) fixés) instationnaire et rétrograde en temps. Ce système, ainsi que la condition finale sur $p(T)$, sont bien définis grâce aux hypothèses (h1) et (h4) de la section précédente. De plus, ce système admet une unique solution car la fonction b est bien intégrable en temps grâce à l'hypothèse (h6) et la fonction A est dans $L^1([0,T];\mathbb{R}^{d\times d})$ grâce à l'hypothèse (h3).

Théorème 7.2.4 (PMP) *On suppose vérifiées les hypothèses* (h1) *à* (h7) *de la Section 7.1. Si $\overline{u} \in \mathcal{U} = L^1([0,T];U)$ est un contrôle optimal, i.e., si $\overline{u}$ est une solution de* (7.3)*, alors en notant $\overline{x} = x_{\overline{u}} \in AC([0,T];\mathbb{R}^d)$ la trajectoire associée au contrôle $\overline{u}$ et $\overline{p} \in AC([0,T];\mathbb{R}^d)$ l'**état adjoint** correspondant, solution de* (7.6) *pour $(x,u) = (\overline{x},\overline{u})$, alors on a, p.p. $t \in [0,T]$,*

$$\overline{u}(t) \in \arg\min_{v\in U} H(t,\overline{x}(t),\overline{p}(t),v), \tag{7.7}$$

où le Hamiltonien $H : [0,T] \times \mathbb{R}^d \times \mathbb{R}^d \times U \to \mathbb{R}$ est défini en (7.5)*. Un triplet $(\overline{x},\overline{p},\overline{u})$ satisfaisant les conditions* (7.1)*,* (7.6)*,* (7.7) *est appelé une **extrémale**.*

Remarque 7.2.5 Dans le cas du système de contrôle non-linéaire (7.1) avec la fonctionnelle J définie en (7.2), 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 7.2.6 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 7.2.1 Appliquons le Théorème 7.2.4 au **système LQ** étudié au chapitre précédent. Pour simplifier, on omet le terme de dérive. On a

$$f(t,x,u) = Ax+Bu, \quad g(t,x,u) = \tfrac{1}{2}Ru\cdot u+\tfrac{1}{2}Qe_x(t)\cdot e_x(t), \quad h(x) = \tfrac{1}{2}De_x(T)\cdot e_x(T),$$

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

$$H(t,x,p,u) = p\cdot(Ax+Bu) + \tfrac{1}{2}Ru\cdot u + \tfrac{1}{2}Qe_x(t)\cdot e_x(t).$$

On a donc (noter l'unicité du minimiseur)

$$\overline{u}(t) = \arg\min_{v\in\mathbb{R}^k}\left(\overline{p}\cdot Bv + \frac{1}{2}Rv\cdot v\right),$$

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 (7.6) 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 6.2.7). •

Exemple 7.2.2 Examinons maintenant le cas du véhicule de Dubins, introduit à l'Exemple 1.3.3. La dynamique est donnée par le système différentiel pour $x = (X,Y,\theta)$

$$\begin{cases} \dot{X}(t) = v\cos(\theta(t)), \\ \dot{Y}(t) = v\sin(\theta(t)), \\ \dot{\theta}(t) = u(t), \end{cases}$$

avec une donnée initiale $(X(0),Y(0),\theta(0))$ et une vitesse $v > 0$ constante. Le contrôle u appartient à $\mathcal{U} = L^2([0,T];U)$ avec $U = [-1,1]$. Le critère à minimiser est seulement lié à une cible au temps final

$$J(u) = h(x_u(T)),$$

où h est une fonction régulière de $\mathbb{R}^3$ dans $\mathbb{R}$ et x_u est la trajectoire associée à u. Pour calculer l'état adjoint p, défini par (7.6), on remarque que $b(t) = 0$ et $A(t)$ est donnée par

$$A(t) = \begin{pmatrix} 0 & 0 & -v\sin(\theta(t)) \\ 0 & 0 & v\cos(\theta(t)) \\ 0 & 0 & 0 \end{pmatrix}.$$

Par conséquent, si on note $p = (p_X, p_Y, p_\theta)^*$, on en déduit que les deux premières composantes sont constantes, $p_X(t) = p_X(T) = \frac{\partial h}{\partial X}(x(T))$, $p_Y(t) = p_Y(T) = \frac{\partial h}{\partial Y}(x(T))$, et que la troisième vérifie une équation différentielle très simple

$$\dot{p}_\theta(t) = v\sin(\theta(t))p_X(T) - v\cos(\theta(t))p_Y(T) \quad \text{avec } p_\theta(T) = \frac{\partial h}{\partial \theta}(x(T)).$$

Le Hamiltonien est $H = p_X v\cos(\theta) + p_Y v\sin(\theta) + p_\theta u$, dont la minimisation en u conduit à

$$\overline{u}(t) = -1 \text{ si } p_\theta(t) > 0, \quad \overline{u}(t) = 1 \text{ si } p_\theta(t) < 0, \quad \overline{u}(t) \text{ indéterminé si } p_\theta(t) = 0.$$

Les deux premiers cas correspondent à un virage à gauche ou un virage à droite et la trajectoire est un arc de cercle. Le troisième cas correspond à un point de commutation (changement d'une valeur extrême à une autre pour le contrôle) si la fonction p_θ ne s'annule qu'en des temps discrets. Par contre, si p_θ s'annule sur un sous-intervalle de temps I, alors on a aussi $\dot{p}_\theta(t) = 0$ sur I mais, au vu de l'équation pour l'adjoint, cela n'est possible que pour une valeur constante de $\theta(t)$, déterminée par $p_X(T), p_Y(T)$. Cette situation correspond alors à une trajectoire rectiligne dans la direction de cette valeur de l'angle. Par conséquent, les trajectoires optimales du véhicule de Dubbins sont constituées d'arcs de cercle et de segments de droite dans une direction constante. •

Contre-exemple 7.2.1 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

$$J(u) = \int_0^1 x_u(t)^2\,dt + \int_0^1 (u(t)^2 - 1)^2\,dt, \quad U = [-1,1].$$

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 $\|x_n\|_{L^\infty(0,1)} \leq \frac{1}{2n}$ (cf. la Figure 7.1). On en déduit que $J(u_n) \leq \frac{1}{4n^2}$. 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. •

FIGURE 7.1 – Illustration du Contre-exemple 7.2.1 : contrôle issu d'une suite minimisante et trajectoire associée.

Contre-exemple 7.2.2 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

$$J(u) = \int_0^1 (x_u(t)^2 - 1)^2 \, dt, \quad U = [-1,1].$$

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 $\inf_{u\in\mathcal{U}} J(u) = \int_0^1 (t^2 - 1)^2 \, dt = \frac{8}{15}$. 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. •

Il existe des cas particuliers où la condition d'optimalité du PMP est aussi suffisante. Nous donnons un résultat positif dans ce sens dans le cas d'un système différentiel linéaire et d'une fonction objectif convexe.

Proposition 7.2.6 (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 contentons 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

$$\langle J'(\overline{u}), v - \overline{u}\rangle_{L^2([0,T];\mathbb{R}^k)} \geq 0, \quad \forall v \in K.$$

Grâce à l'introduction de l'état adjoint $\overline{p}$ solution de (7.6), ceci se récrit

$$\int_0^T \left(\overline{p}(t) \cdot B(v(t) - \overline{u}(t)) + \frac{\partial g}{\partial u}(t,\overline{x}(t),\overline{u}(t)) \cdot (v(t) - \overline{u}(t)) \right) dt \geq 0, \quad \forall v \in K.$$

Cette inégalité, toujours grâce au Théorème 2.5.1, équivaut au fait que $\overline{u}$ soit minimiseur sur K de la fonctionnelle

$$\tilde{J}(u) = \int_0^T \left(\overline{p}(t) \cdot Bu(t) + g(t, \overline{x}(t), u(t)) \right) dt.$$

En posant $\Phi(t, v) := \overline{p}(t) \cdot Bv + g(t, \overline{x}(t), v)$, nous avons donc établi que

$$\int_0^T \Phi(t, \overline{u}(t))\, dt \leq \int_0^T \Phi(t, u(t))\, dt, \qquad \forall u \in K.$$

Supposons par l'absurde que $\Phi(t, \overline{u}(t)) > \min_{v \in U} \Phi(t, v)$ 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

$$\int_0^T \min_{v \in U} \Phi(t, v)\, dt < \int_0^T \Phi(t, \overline{u}(t))\, dt.$$

En posant $\hat{u}(t) = \arg\min_{v \in U} \Phi(t, v)$ sur [0,T], on peut montrer (voir le Théorème 8.2.8 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

$$\int_0^T \Phi(t, \hat{u}(t))\, dt = \int_0^T \min_{v \in U} \Phi(t, v)\, dt < \int_0^T \Phi(t, \overline{u}(t))\, dt \leq \int_0^T \Phi(t, \hat{u}(t))\, dt.$$

Ainsi $\overline{u}(t)$ est minimiseur instantané de $v \mapsto \overline{p}(t) \cdot Bv + g(t, \overline{x}(t), v)$, ce qui n'est rien d'autre que minimiser le Hamiltonien par rapport à v. □

Remarque 7.2.7 Comme pour le cas du système LQ (voir la Sous-section 6.2.2) on peut se demander quelle est l'origine de la Définition 7.2.2 de l'état adjoint. De la même manière, pour deviner la système adjoint il faut d'abord introduire un Lagrangien $\mathcal{L}(u, x, p)$ qui est la somme de la fonction objectif et de la contrainte multipliée par un multiplicateur de Lagrange $p(t)$, qui deviendra l'état adjoint à l'optimalité. Plus précisément, si on note

$$\tilde{J}(u, x) = \int_0^T g(t, x(t), u(t))\, dt + h(x(T)),$$

alors le Lagrangien est

$$\mathcal{L}(u, x, p) = \tilde{J}(u, x) - \int_0^T p \cdot (\dot{x} - f(t, x(t), u(t)))\, dt - p(0) \cdot (x(0) - x_0)\,.$$

Un calcul facile permet de vérifier que la condition

$$\langle \frac{\partial \mathcal{L}}{\partial x}(u, x, p), \phi \rangle = 0 \quad \forall \phi \in AC([0,T]; \mathbb{R}^d)$$

redonne le système adjoint (7.6) dont p est solution.

Le lien entre Lagrangien et Hamiltonien est donné par l'égalité suivante

$$\mathcal{L}(u,x,p) = \int_0^T \Big(g(t,x(t),u(t)) + p\cdot f(t,x(t),u(t))\Big)dt + \mathbf{r},$$

où $\mathbf{r} = -\int_0^T p\cdot\dot{x}\,dt + h(x(T)) - p(0)\cdot(x(0)-x_0)$ est un reste. Si on néglige ce reste, ce qui correspond à ignorer les conditions initiales et finales et à ignorer les variations en temps, alors on trouve que le Lagrangien n'est que l'intégrale sur [0,T] de l'Hamiltonien. Autrement dit, si on "gèle le temps", alors les deux notions sont identiques. (Attention, le Lagrangien dont on parle ici est différent de celui introduit en mécanique Lagrangienne et Hamiltonienne et qui est la transformée de Legendre de l'Hamiltonien.) •

Remarque 7.2.8 Terminons cette section en indiquant qu'il existe de nombreuses généralisations du principe du minimum de Pontryaguine (PMP), tel qu'énoncé dans le Théorème 7.2.4. En particulier, nous n'avons rien dit ici de deux cas importants en pratique pour lesquels nous renvoyons à [8], [33], [32], [34]. En premier lieu, l'horizon temporel, ou temps final T, est toujours fixé et jamais variable d'optimisation dans ce chapitre. Bien évidemment, les problèmes de temps minimal sont très importants en pratique et le PMP s'étend à ce cas. En second lieu, il n'y a pas de contrainte de cible finale dans ce chapitre alors que c'est aussi une problématique très naturelle. Encore une fois, le PMP se généralise à cette situation. Dans ces deux cas, la définition de l'Hamiltonien change un peu mais surtout la condition finale de l'adjoint est différente. Nous n'en disons rien pour ne pas aller trop loin... •

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

$$\dot{x}_u(t) = Ax_u(t) + Bu(t), \quad \forall t\in[0,T], \qquad x_u(0) = x_0. \tag{7.8}$$

Soit U un sous-ensemble **convexe, fermé, non-vide** de $\mathbb{R}^k$. L'ensemble des contrôles admissibles est ici le sous-ensemble

$$K = L^2([0,T];U). \tag{7.9}$$

On s'intéresse au problème de minimisation sous contraintes :

$$\text{trouver } \overline{u} \in K \text{ tel que } J(\overline{u}) = \inf_{u \in K} J(u), \tag{7.10}$$

avec le critère quadratique

$$J(u) = \frac{1}{2}\int_0^T Ru(t)\cdot u(t)\,dt + \frac{1}{2}\int_0^T Qe_{x_u}(t)\cdot e_{x_u}(t)\,dt + \frac{1}{2}De_{x_u}(T)\cdot e_{x_u}(T), \tag{7.11}$$

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* (7.10)*, i.e., la fonctionnelle* J *définie par* (7.11) *admet un unique minimiseur sur le sous-ensemble* K *défini par* (7.9)*.*

Démonstration. Il suffit d'appliquer le Théorème 2.3.9. 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 sur V (cf. le Lemme 6.1.2). □

Dans la suite de cette section, on notera $\overline{u} \in K = L^2([0,T];U)$ l'unique contrôle optimal solution de (7.10) 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 7.2.1 (qui traitait le cas sans contraintes), on introduit l'état adjoint $\overline{p} \in C^1([0,T];\mathbb{R}^d)$ tel que

$$\frac{d\overline{p}}{dt}(t) = -A^*\overline{p}(t) - Qe_{\overline{x}}(t), \quad \forall t \in [0,T], \qquad \overline{p}(T) = De_{\overline{x}}(T),$$

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

$$H(t,x,p,u) = p\cdot(Ax+Bu) + \frac{1}{2}Ru\cdot u + \frac{1}{2}Q(x-\xi(t))\cdot(x-\xi(t)).$$

En appliquant le PMP (Théorème 7.2.4), 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.,

$$\overline{u}(t) \in \arg\min_{v\in U} H(t,\overline{x}(t),\overline{p}(t),v).$$

En inspectant l'expression de H, on voit que de manière équivalente, on a

$$\overline{u}(t) \in \arg\min_{v\in U}\left(v\cdot B^*\overline{p}(t)+\frac{1}{2}Rv\cdot v\right).$$

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

$$\overline{u}(t) = \arg\min_{v\in U}\left(v\cdot B^*\overline{p}(t)+\frac{1}{2}Rv\cdot v\right). \tag{7.12}$$

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* (7.12) *est une* ***condition nécessaire et suffisante*** *d'optimalité pour le problème* (7.10)*. 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 (cf. Lemme 6.2.1), une condition nécessaire et suffisante d'optimalité pour le problème (7.10) est l'inéquation d'Euler (Théorème 2.5.1)

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

En utilisant l'expression de la différentielle de J obtenue au Lemme 6.2.1, on en déduit que

$$\langle R\overline{u}+B^*\overline{p}, v-\overline{u}\rangle_V \geq 0, \quad \forall v\in K,$$

ou encore, en explicitant le produit scalaire dans $V=L^2([0,T];\mathbb{R}^k)$,

$$\int_0^T (v(t)-\overline{u}(t))\cdot(R\overline{u}(t)+B^*\overline{p}(t))\,dt \geq 0, \quad \forall v\in K=L^2([0,T];U).$$

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

$$\mathcal{J}_{\overline{p}}: V\to\mathbb{R}, \qquad \mathcal{J}_{\overline{p}}(v)=\int_0^T\left(v(t)\cdot B^*\overline{p}(t)+\frac{1}{2}Rv(t)\cdot v(t)\right)dt$$

est quadratique, différentiable et fortement convexe sur V. On pose pour tout $t \in [0,T]$,

$$u_\sharp(t) = \arg\min_{v\in U}\left(v \cdot B^*\overline{p}(t) + \frac{1}{2}Rv \cdot v\right).$$

De l'inéquation d'Euler dans $U \subset \mathbb{R}^k$, on déduit que pour tout $t \in [0,T]$,

$$(v - u_\sharp(t)) \cdot (Ru_\sharp(t) + B^*\overline{p}(t)) \geq 0, \quad \forall v \in U.$$

(2) Montrons que la fonction $u_\sharp(t)$ ainsi définie est lipschitzienne en t sur $[0,T]$. Soit $t_1, t_2 \in [0,T]$. En prenant $v = u_\sharp(t_2)$ en t_1 et $v = u_\sharp(t_1)$ en t_2 on a

$$\begin{aligned}(u_\sharp(t_2) - u_\sharp(t_1)) \cdot (Ru_\sharp(t_1) + B^*\overline{p}(t_1)) \geq 0,\\ (u_\sharp(t_1) - u_\sharp(t_2)) \cdot (Ru_\sharp(t_2) + B^*\overline{p}(t_2)) \geq 0.\end{aligned}$$

En posant $\delta u_\sharp = u_\sharp(t_2) - u_\sharp(t_1)$, il vient

$$R\delta u_\sharp \cdot \delta u_\sharp \leq \delta u_\sharp \cdot B^*(\overline{p}(t_1) - \overline{p}(t_2)).$$

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. $t \in [0,T]$:

$$\overline{u}(t) \cdot B^*\overline{p}(t) + \frac{1}{2}R\overline{u}(t) \cdot \overline{u}(t) \geq u_\sharp(t) \cdot B^*\overline{p}(t) + \frac{1}{2}Ru_\sharp(t) \cdot u_\sharp(t).$$

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

$$\dot{x}(t) = \begin{pmatrix}\dot{a}(t)\\ \dot{r}(t)\end{pmatrix} = \begin{pmatrix}\varphi(u(t))a(t)\\ \gamma u(t)a(t)\end{pmatrix}, \quad \forall t \in [0,T], \tag{7.13}$$

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

$$\varphi : [0,1] \to \mathbb{R}, \quad \varphi(v) = \alpha(1-v) - \beta. \tag{7.14}$$

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

$$J(u) = -r(T), \tag{7.15}$$

le problème de contrôle optimal est donc le suivant :

$$\text{Chercher } \overline{u} \in \mathcal{U} \text{ tel que } J(\overline{u}) = \inf_{u \in \mathcal{U}} J(u). \tag{7.16}$$

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 (7.16) rentre dans le cadre d'application du PMP en posant

$$f(x,u) = \begin{pmatrix} \varphi(u)a \\ \gamma u a \end{pmatrix}, \quad g(x,u) = 0, \quad h(x) = -r. \tag{7.17}$$

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

$$\begin{cases} \dfrac{d\overline{p}_a}{dt}(t) = -\varphi(\overline{u}(t))\overline{p}_a(t) - \gamma \overline{u}(t)\overline{p}_r(t), \\ \dfrac{d\overline{p}_r}{dt}(t) = 0, \end{cases} \qquad \forall t \in [0,T], \tag{7.18}$$

et la condition finale sur l'état adjoint est

$$\overline{p}(T) = (\overline{p}_a(T), \overline{p}_r(T))^* = (0,-1)^*. \tag{7.19}$$

On a donc

$$\frac{d\overline{p}_a}{dt}(t) = -\varphi(\overline{u}(t))\overline{p}_a(t) + \gamma \overline{u}(t), \quad \overline{p}_r(t) \equiv -1, \quad \forall t \in [0,T]. \tag{7.20}$$

Par ailleurs, le Hamiltonien est autonome (cf. la Définition 7.2.1) et s'écrit sous la forme

$$H(x,p,u) = p_a \varphi(u) a + \gamma p_r u a. \tag{7.21}$$

La condition de minimisation (7.7) s'écrit, en utilisant le fait que $\overline{a}(t) \neq 0$ pour tout $t \in [0,T]$,

$$\overline{u}(t) \in \arg\min_{v\in[0,1]} \psi(t)v, \tag{7.22}$$

où la **fonction de commutation** est donnée par

$$\psi(t) = -\overline{p}_a(t)\alpha - \gamma. \tag{7.23}$$

La solution du problème de minimisation (7.22) 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 $\varphi(u(t))\frac{\gamma}{\alpha} + \gamma u(t) = \frac{\gamma(\alpha-\beta)}{\alpha} \neq 0$, 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

$$\overline{p}_a(t) = -\frac{\gamma}{\beta}\left(1 - e^{\beta(t-T)}\right), \quad \forall t \in [t_*, T]. \tag{7.24}$$

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

$$t_* = \frac{1}{\beta}\ln(1 - \frac{\beta}{\alpha}) + T. \tag{7.25}$$

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 $t = T$;

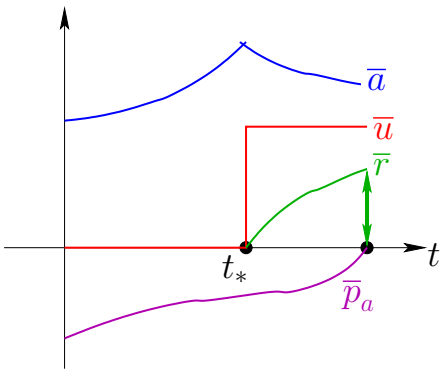

FIGURE 7.2 – Trajectoire, état adjoint et contrôle optimal pour le modèle de ruche.

— **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 7.2.4 qui établissait le principe du minimum de Pontryaguine (PMP). La preuve est esquissée au sens où certains points techniques de justification de manipulation de fonctions seulement mesurables (et pas plus régulières) ne sont que rapidement évoqués. En particulier, on utilise la notion de point de Lebesgue d'une fonction mesurable que l'on ne détaille pas plus ici. Par contre, toutes les idées principales et tous les arguments clés sont bien présents dans ce qui suit.

Démonstration du Théorème 7.2.4. Nous allons nous contenter de donner une esquisse de la preuve, en insistant sur les idées principales mais 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 (en taille !) 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

$$u_\delta(t) = \begin{cases} \overline{u}(t), & \forall t \in [0,T] \setminus I_\delta, \\ v, & \forall t \in I_\delta. \end{cases}$$

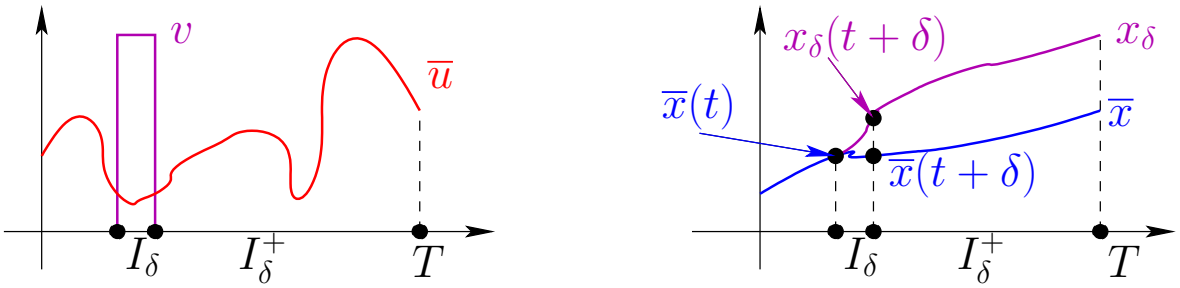

FIGURE 7.3 – Principe de la variation aiguille pour le contrôle optimal $\overline{u}$ (à gauche), trajectoire optimale et trajectoire perturbée (à droite).

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 $\psi = g$,

$$\lim_{\delta \to 0^+} \frac{1}{\delta} \int_{I_\delta} \psi(s, \overline{x}(s), \overline{u}(s))\, ds = \psi(t, \overline{x}(t), \overline{u}(t)).$$

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) On compare les trajectoires $\overline{x}$ et x_δ (voir la Figure 7.3). L'intervalle $[0,T]$ est découpée en trois parties successives : $[0,t]$, $I_\delta = [t,t+\delta]$ et $I_\delta^+ = [t+\delta,T]$. Tout d'abord, les trajectoires sont identiques sur $[0,t]$. Puis, comme $x_\delta(t) = \overline{x}(t)$, on a $x_\delta(s) = \overline{x}(s) + \mathcal{O}(\delta)$ pour tout $s \in I_\delta$. On peut alors invoquer la continuité de f en (t,x) et la propriété des points de Lebesgue afin d'obtenir les estimations suivantes plus précises :

$$x_\delta(t+\delta) = \overline{x}(t) + \int_{I_\delta} f(s, x_\delta(s), v)\, ds = \overline{x}(t) + \delta f(t, \overline{x}(t), v) + o(\delta),$$
$$\overline{x}(t+\delta) = \overline{x}(t) + \int_{I_\delta} f(s, \overline{x}(s), \overline{u}(s))\, ds = \overline{x}(t) + \delta f(t, \overline{x}(t), \overline{u}(t)) + o(\delta),$$

si bien que

$$x_\delta(t+\delta) - \overline{x}(t+\delta) = \delta\Big(f(t, \overline{x}(t), v) - f(t, \overline{x}(t), \overline{u}(t))\Big) + o(\delta).$$

Enfin, on compare $x_\delta(s)$ et $\overline{x}(s)$ pour $s \in I_\delta^+ = [t+\delta,T]$. Il est clair que $x_\delta(s) - \overline{x}(s) = \mathcal{O}(\delta)$ 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 linéarisée

$$\dot{y}_\delta(s) = \overline{A}(s) y_\delta(s), \quad \forall s \in I_\delta^+, \qquad y_\delta(t+\delta) = f(t, \overline{x}(t), v) - f(t, \overline{x}(t), \overline{u}(t)),$$

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 $\Psi_\delta(s) = o(\delta)$ uniformément sur I_δ^+, car

$$\Psi_\delta(s) = f(s, x_\delta(s), \overline{u}(s)) - f(s, \overline{x}(s), \overline{u}(s)) - \overline{A}(s)(x_\delta(s) - \overline{x}(s)).$$

En conclusion de cette première étape de la preuve, on a donc

$$x_\delta(s) - \overline{x}(s) = \delta y_\delta(s) + o(\delta) \text{ uniformément sur } I_\delta^+.$$

(3) On compare maintenant les 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

$$\begin{aligned}
J(u_\delta) - J(\overline{u}) &= \int_t^T g(s, x_\delta(s), u_\delta(s)) - g(s, \overline{x}(s), \overline{u}(s))\, ds + h(x_\delta(T)) - h(\overline{x}(T)) \\
&= \int_{I_\delta} g(s, x_\delta(s), v) - g(s, \overline{x}(s), \overline{u}(s))\, ds + \int_{I_\delta^+} g(s, x_\delta(s), \overline{u}(s)) - g(s, \overline{x}(s), \overline{u}(s))\, ds \\
&\quad + \delta \frac{\partial h}{\partial x}(\overline{x}(T)) \cdot y_\delta(T) + o(\delta) \\
&= \delta(g(t, \overline{x}(t), v) - g(t, \overline{x}(t), \overline{u}(t))) + \delta \int_{t+\delta}^T \overline{b}(s) \cdot y_\delta(s)\, ds \\
&\quad + \delta \frac{\partial h}{\partial x}(\overline{x}(T)) \cdot y_\delta(T) + o(\delta),
\end{aligned}$$

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, après division par δ, que

$$0 \le g(t, \overline{x}(t), v) - g(t, \overline{x}(t), \overline{u}(t)) + \int_{t+\delta}^T \overline{b}(s) \cdot y_\delta(s)\, ds + \frac{\partial h}{\partial x}(\overline{x}(T)) \cdot y_\delta(T) + o(1).$$

(4) Pour conclure on introduit l'état adjoint $\overline{p}$ qui va permettre d'éliminer la fonction y_δ. En effet, l'adjoint vérifie

$$\frac{d\overline{p}}{dt}(s) = -\overline{A}(s)^* \overline{p}(s) - \overline{b}(s) \text{ sur } [0,T], \quad \overline{p}(T) = \frac{\partial h}{\partial x}(\overline{x}(T)).$$

Si on multiplie par y_δ et qu'on utilise le système vérifié par y_δ, il vient

$$\begin{aligned}
\int_{t+\delta}^T \overline{b}(s) \cdot y_\delta(s)\, ds + \frac{\partial h}{\partial x}(\overline{x}(T)) \cdot y_\delta(T) &= \int_{t+\delta}^T \left(-\frac{d\overline{p}}{dt}(s) - \overline{A}(s)^* \overline{p}(s) \right) \cdot y_\delta(s)\, ds + \overline{p}(T) \cdot y_\delta(T) \\
&= \int_{t+\delta}^T -\frac{d}{dt}(\overline{p}(s) \cdot y_\delta(s))\, ds + \overline{p}(T) \cdot y_\delta(T) \\
&= \overline{p}(t+\delta) \cdot y_\delta(t+\delta) \\
&= \overline{p}(t+\delta) \cdot (f(t, \overline{x}(t), v) - f(t, \overline{x}(t), \overline{u}(t)))\,.
\end{aligned}$$

En faisant tendre δ vers 0, il vient par continuité de $\overline{p}$,

$$0 \le g(t, \overline{x}(t), v) - g(t, \overline{x}(t), \overline{u}(t)) + \overline{p}(t) \cdot (f(t, \overline{x}(t), v) - f(t, \overline{x}(t), \overline{u}(t)))\,,$$

et en utilisant la définition du Hamiltonien, on obtient

$$0 \le H(t, \overline{x}(t), \overline{p}(t), v) - H(t, \overline{x}(t), \overline{p}(t), \overline{u}(t)),$$

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 [17]). 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 [17]).

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 $\theta > 0$

$$0 \leq -2\langle x - x_K, y - x_K\rangle + \theta\|y - x_K\|^2.$$

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 [17]) 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 2.5.20 (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. □

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 [30]. 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 $t \in I$,

$$\mathcal{E}_\Phi(t) = \{(u, \alpha) \in \mathbb{R}^k \times \mathbb{R} \mid \Phi(t, u) \leq \alpha\}, \tag{8.8a}$$

$$\mathcal{D}_\Phi(t) = \{u \in \mathbb{R}^k \mid \Phi(t, u) < +\infty\}. \tag{8.8b}$$

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* $N_\alpha(t) = \{u \in \mathbb{R}^k \mid \Phi(t,u) \leq \alpha\}$ *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 $\{(u,\alpha) \in \mathbb{R}^k \times \mathbb{R} \mid f(u) \leq \alpha\}$ 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 $f(v) \geq f(u) - \epsilon$.

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 8.2.1 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 8.2.8 (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 8.2.9 (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. Il existe de très nombreux ouvrages sur ce sujet et nous renvoyons simplement aux cours [29] (pour les aspects théoriques) et [24] (pour les aspects numériques et applicatifs). On fixe un temps final $T > 0$ 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 donner un résultat classique dont la preuve est relativement simple.

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., il existe $C_0 > 0$ tel que*

$$\forall t \in [0,T],\ \forall x_1, x_2 \in \mathbb{R}^d, \quad |F(t,x_1) - F(t,x_2)|_{\mathbb{R}^d} \leq C_0 |x_1 - x_2|_{\mathbb{R}^d}.$$

Alors, il existe une unique solution $x \in C^1([0,T]; \mathbb{R}^d)$ au problème de Cauchy (8.10).

Démonstration. Le principe de la preuve consiste à observer que x est solution du problème de Cauchy (8.10) si et seulement si

$$x(t) = x_0 + \int_0^t F(s, x(s))\, ds, \quad \forall t \in [0,T].$$

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

$$\Phi(y)(t) = x_0 + \int_0^t F(s, y(s))\, ds, \quad \forall t \in [0,T].$$

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

$$\begin{aligned}
\|\Phi(y_1) - \Phi(y_2)\|_{Y*} &= \sup_{t\in[0,T]} \left(e^{-C_0 t} |\Phi(y_1)(t) - \Phi(y_2)(t)|_{\mathbb{R}^d} \right) \\
&\leq \sup_{t\in[0,T]} \left(e^{-C_0 t} \int_0^t |F(s, y_1(s)) - F(s, y_2(s))|_{\mathbb{R}^d} \, ds \right) \\
&\leq \sup_{t\in[0,T]} \left(e^{-C_0 t} C_0 \int_0^t |y_1(s) - y_2(s)|_{\mathbb{R}^d} \, ds \right) \\
&= \sup_{t\in[0,T]} \left(e^{-C_0 t} C_0 \int_0^t e^{C_0 s} e^{-C_0 s} |y_1(s) - y_2(s)|_{\mathbb{R}^d} \, ds \right) \\
&\leq \left(\sup_{t\in[0,T]} e^{-C_0 t} C_0 \int_0^t e^{C_0 s} \, ds \right) \|y_1 - y_2\|_{Y*} \\
&= \left(\sup_{t\in[0,T]} 1 - e^{-C_0 t} \right) \|y_1 - y_2\|_{Y*} = \left(1 - e^{-C_0 T}\right) \|y_1 - y_2\|_{Y*},
\end{aligned}$$

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. □

Les hypothèses du Théorème 8.3.1 sont trop fortes en pratique et on va donc donner deux généralisations de ce résultat, plus utiles mais aux preuves plus techniques qui ne seront pas détaillées ici. Tout d'abord, 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

$$\dot{x}(t) = f(t, x(t), u(t)), \quad \forall t \in [0,T], \qquad x(0) = x_0, \tag{8.11}$$

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 (8.11) se ramène à celle du problème de Cauchy (8.10) en posant

$$F(t, x) = f(t, x, u(t)), \quad \forall (t, x) \in [0,T] \times \mathbb{R}^d.$$

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).

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.,*

$$\forall x \in \mathbb{R}^d, \quad \exists \beta \in L^1([0,T]; \mathbb{R}_+), \quad \forall t \in [0,T], \quad |F(t, x)|_{\mathbb{R}^d} \leq \beta(t),$$

(iii) *l'application F est* ***globalement lipschitzienne*** *en x, i.e., il existe une fonction $C_0 \in L^1([0,T];\mathbb{R}_+)$ telle que*

$$\text{p.p. } t \in [0,T],\ \forall x_1, x_2 \in \mathbb{R}^d, \quad |F(t,x_1) - F(t,x_2)|_{\mathbb{R}^d} \leq C_0(t)|x_1 - x_2|_{\mathbb{R}^d}.$$

Alors, il existe une ***unique solution*** $x \in AC([0,T];\mathbb{R}^d)$ *au problème de Cauchy* (8.10) *Cette solution, qui est dérivable p.p. sur $[0,T]$, satisfait le système différentiel pour presque tout $t \in [0,T]$; elle vérifie également*

$$x(t) = x_0 + \int_0^t F(s,x(s))\,ds, \quad \forall t \in [0,T]. \tag{8.12}$$

On rappelle que $AC([0,T];\mathbb{R}^d)$ est l'espace des fonctions absolument continues, introduit à la Définition 5.1.2. Une différence notable entre les solutions données par le Théorème 8.3.1 et celles du Théorème 8.3.2 est que les premières vérifient l'équation différentielle ordinaire (8.10) en tout temps $t \in [0,T]$ alors que les secondes la vérifie presque partout dans $[0,T]$ (puisque la dérivée $\dot{x}(t)$ est une fonction simplement mesurable).

Remarque 8.3.3 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). Cependant, lorsque l'application F est non-linéaire en x, la propriété d'être globalement lipschitzienne est en général fausse (considérer, par exemple, $F(t,x) = x^2$ pour $x \in \mathbb{R}$). Il faut donc changer cette hypothèse. Mais il y a plus grave : il faut changer le résultat car, dans ce cas, il est bien connu que la solution x du problème de Cauchy (8.10) peut exploser en temps fini, c'est-à-dire avant le temps final T.

Contre-exemple 8.3.1 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 $|x_0| \leq 1$, on trouve $x(t) = \tanh(t + t_0)$ 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 $|x_0| > 1$, on trouve $x(t) = \coth(t+t_0)$ 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. •

Contre-exemple 8.3.2 Lorsque l'application F est seulement 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 donne donc une extension 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 [32]).

Théorème 8.3.4 (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.,*

$$\forall x \in \mathbb{R}^d, \quad \exists \beta \in L^1([0,T];\mathbb{R}_+), \quad \forall t \in [0,T], \quad |F(t,x)|_{\mathbb{R}^d} \leq \beta(t),$$

(iii) *l'application F est* ***localement lipschitzienne*** *en x, i.e., pour tout $x \in \mathbb{R}^d$, il existe $r > 0$ et $C_0 \in L^1([0,T];\mathbb{R}_+)$ tels que*

$$\text{p.p. } t \in [0,T],\ \forall x_1, x_2 \in B(x,r), \quad |F(t,x_1) - F(t,x_2)|_{\mathbb{R}^d} \leq C_0(t)|x_1 - x_2|_{\mathbb{R}^d},$$

où $B(x,r)$ désigne la boule ouverte de centre x et de rayon r.
Alors, il existe une ***unique solution maximale*** *$x \in AC(J;\mathbb{R}^d)$ au problème de Cauchy (8.10), où $J \subseteq [0,T]$ est l'intervalle défini, soit par $J = [0,T]$, soit par $J = [0,T_*[$ avec $T_* < T$ et $\lim_{t \uparrow T_*} |x(t)|_{\mathbb{R}^d} = +\infty$.*

A nouveau, les solutions données par le Théorème 8.3.4 vérifient l'équation différentielle ordinaire (8.10) seulement presque partout dans $[0,T]$.

Contre-exemple 8.3.3 Même en présence d'un contrôle, la solution peut exploser en temps fini. Prenons l'exemple du système de contrôle (8.11) en dimensions $d = 1$ et $k = 1$ avec l'application $f(t,x,u) = x^2 + u$ (qui ne dépend pas de t explicitement). Le problème de Cauchy est donc $\dot{x}(t) = x(t)^2 + u(t)$. 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] ALLAIRE G., *Conception optimale de structures,* Collection : Mathématiques et Applications, Vol. 58, Springer (2007).

[3] 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.

[4] 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.

[5] BONNANS J., *Optimisation continue,* Mathématiques appliquées pour le Master / SMAI, Dunod, Paris (2006).

[6] BONNANS J., GAUBERT S., *Recherche opérationnelle. Aspects mathématiques et applications,* Éditions de l'École Polytechnique, Palaiseau (2015).

[7] BONNANS J., GILBERT J.-C., LEMARECHAL C., SAGASTIZABAL C., *Optimisation numérique,* Mathématiques et Applications 27, Springer, Paris (1997).

[8] BONNANS J., ROUCHON P., *Commande et optimisation de systèmes dynamiques,* Éditions de l'École Polytechnique, Palaiseau (2005).

[9] BREZIS H., *Functional analysis, Sobolev spaces and partial differential equations.* Universitext. Springer, New York, 2011.

[10] CARPENTIER P., COHEN G., *Décomposition-coordination en optimisation déterministe et stochastique,* Mathématiques et Applications 81, Springer, Paris (2017).

[11] CHAMBOLLE A., POCK T., *An introduction to continuous optimization for imaging,* Acta Numerica, pp. 161-319 (2016).

[12] CHVÁTAL V., *Linear programming*, Freeman and Co., New York (1983).

[13] CULIOLI J.-C., *Introduction à l'optimisation,* Éditions Ellipses, Paris (1994).

[14] ERN A., *Contrôle de modèles dynamiques,* cours de 2ème année à l'Ecole Polytechnique (2019).

[15] EKELAND I., TEMAM R., *Analyse convexe et problèmes variationnels,* Dunod, Paris (1974).

[16] FLETCHER R., *Practical methods of optimization. Vol. 1.* John Wiley & Sons, Ltd., Chichester, 1980. Unconstrained optimization, A Wiley-Interscience Publication.

[17] GOLSE F., LASZLO Y., PACARD F., VITERBO C., *Analyse réelle,* cours de 1ère année à l'Ecole Polytechnique (2019).

[18] GRIVA I., NASH S., SOFER A., *Linear and Nonlinear Optimization,* Second Edition, SIAM, Philadelphia (2009).

[19] HERMES H., LASALLE J. P., *Functional analysis and time optimal control.* Academic Press, New York-London, 1969. Mathematics in Science and Engineering, Vol. 56.

[20] ISIDORI A., *Nonlinear control systems.* Communications and Control Engineering Series. Springer-Verlag, Berlin, third edition, 1995.

[21] KINGMA D., BA J., *Adam : A Method for Stochastic Optimization.* 3rd International Conference for Learning Representations, San Diego (2015). ArXiv preprint arXiv :1412.6980.

[22] LEE E. B., MARKUS L., *Foundations of optimal control theory.* Robert E. Krieger Publishing Co., Inc., Melbourne, FL, second edition, 1986.

[23] LIONS P.-L., *Contrôle de modèles dynamiques.* Cours polycopié. École Polytechnique, 2016.

[24] MASSOT M., SERIES L., BREDEN M., PICHARD T., *Introduction à l'analyse numérique : des fondements mathématiques à l'expérimentation avec Jupyter,* cours de 2ème année à l'Ecole Polytechnique (2020).

[25] NESTEROV Y., *Lectures on convex optimization,* Springer Optimization and Its Applications, 137, Springer, Cham, 2018.

[26] NOCEDAL J., WRIGHT S., *Numerical optimization,* Springer Series in Operations Research and Financial Engineering, New York (2006).

[27] PADBERG M., *Linear optimization and extensions*, Springer, Berlin (1999).

[28] POLYAK B., *Introduction to optimization*, Optimization software, New York (1987).

[29] RENARD D., *Calcul différentiel et analyse complexe,* cours de 2ème année à l'Ecole Polytechnique (2023).

[30] 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.

[31] SCHRIJVER A., *Theory of linear and integer programming*, Wiley, New York (1986).

[32] SONTAG E. D., *Mathematical control theory*, volume 6 of *Texts in Applied Mathematics.* Springer-Verlag, New York, second edition, 1998. Deterministic finite-dimensional systems.

[33] TRELAT E., *Contrôle optimal.* Mathématiques Concrètes. Vuibert, Paris, 2005. Théorie & applications.

[34] VINTER R., *Optimal control*. Systems & Control : Foundations & Applications. Birkhäuser Boston, Inc., Boston, MA, 2000.

Index