COMMANDE ET OPTIMISATION DE SYSTÈMES DYNAMIQUES

Frédéric Bonnans Pierre Rouchon

9 mars 2005

Avant propos

Les mécanismes de régulation et d'adaptation sont largement répandus dans la nature. Chez les organismes vivants, ils assurent le maintien de certaines variables essentielles comme le taux de sucre, la température, ... En ingénierie également les mécanismes d'asservissement et de recalage ont une longue histoire. Au temps des romains les niveaux d'eau dans les aqueducs étaient pilotés par un système complexe de vannes.

Les développements modernes ont débuté au 17ème siècle avec les travaux du savant hollandais Huyghens sur les horloges à pendules. Il était alors très important pour la marine de Louis XIV d'embarquer sur les bateaux des horloges les plus précises possible. La mesure du temps intervenait de façon cruciale dans les calculs de position. Huyghens s'est ainsi intéressé à la régulation en vitesse des horloges. Les idées élaborées par Huyghens et bien d'autres comme le savant anglais Robert Hooke furent utilisées dans la régulation en vitesse des moulins à vent. Une idée centrale fut alors d'utiliser un système mécanique à boules tournant autour d'un axe et dont la rotation était directement proportionnelle à celle du moulin. Plus les boules tournent vite et plus elles s'éloignent de l'axe. Elles actionnent ainsi par un système de renvois ingénieux les ailes du moulin de façon à réduire le couple dû au vent. En langage moderne, il s'agit d'un régulateur proportionnel.

La révolution industrielle vit l'adaptation par James Watt du régulateur à boules pour les machines à vapeur. Plus les boules tournent vite, plus elles ouvrent une soupape qui laisse s'échapper la vapeur. La pression de la chaudière baissant, la vitesse diminue. Le problème était alors de maintenir la vitesse de la machine constante malgré les variations de charge. Le mathématicien et astronome anglais Georges Airy fut le premier à tenter une analyse du régulateur à boules de Watt. Ce n'est qu'en 1868 que le physicien écossais James Clerk Maxwell publia une première analyse mathématique convaincante et expliqua ainsi certains comportement erratiques observés parmi les 75 000 régulateurs en service à cet époque. Ses travaux furent le point de départ de nombreux autres sur la stabilité, sa caractérisation ayant été obtenue indépendamment par les mathématiciens A. Hurwitz et E.J. Routh.

Durant les années 1930, les recherches aux "Bell Telephone Laboratories" sur les amplificateurs sont à l'origine d'idées encore enseignées aujourd'hui. Citons par exemple les travaux de Nyquist et Bode caractérisant à partir de la réponse fréquentielle en boucle ouverte celle de la boucle fermée. Pendant la seconde guerre mondiale, ces techniques furent utilisées et très activement développées en particulier lors de la mise au point de batteries anti-aériennes. Le mathématicien Nobert Wiener a donné le nom de "cybernétique" à toutes ces techniques.

Tous ces développements se faisaient dans le cadre des systèmes linéaires avec une seule commande et une seule sortie : on disposait d'une mesure sous la forme d'un signal électrique. Amplifié, filtré et convenablement traité, ce signal devenait

alors un signal de contrôle. Ce n'est qu'après les années 50 que les développements théoriques et technologiques (calculateurs numériques) permirent le traitement des systèmes multi-variables linéaires et non linéaires avec plusieurs entrées et plusieurs sorties. Citons les contributions de Rudolf Kalman avec la théorie de l'automatique en variable d'état et du filtrage, de Richard Bellmann avec la programmation dynamique, et l'école de L. Pontryagin avec le principe du même nom pour la commande optimale.

Ces contributions continuent encore aujourd'hui à alimenter les recherches en théorie des systèmes. L'objectif de ce cours est de souligner le lien entre les apports théoriques et les applications. Partant de ces dernières, nous développons au fil des chapitres les concepts fondamentaux de l'automatique et de la commande optimale, sans négliger les aspects concrets et l'illustration sur des exemples.

Frédéric Bonnans et Pierre Rouchon
Février 2005

Table des matières

I Stabilité, Commandabilité et Observabilité **11**

1 Introduction **13**
1.1 Un exemple emprunté à la robotique 13
1.2 Le plan 18
1.3 Problème 19

2 Étude de cas **21**
2.1 Le bio-réacteur 21
2.1.1 Étude à $D > 0$ fixé 22
2.1.2 Stabilisation (globale) par feedback (borné) 27
2.2 L'avion à décollage vertical 29
2.2.1 Modèle de simulation 30
2.2.2 Modèle de commande 31
2.2.3 Commande linéaire 32
2.2.4 Commande non-linéaire 35
2.3 Pendule inversé sur un rail 36
2.4 Moteur électrique à courant continu 37
2.4.1 Stabilité en boucle ouverte 38
2.4.2 Estimation de la vitesse et de la charge 39
2.4.3 Le contrôleur 40
2.4.4 L'observateur-contrôleur 41
2.4.5 Robustesse par rapport à la dynamique rapide du courant . . . 41
2.4.6 Boucle rapide et contrainte de courant 43

3 Systèmes dynamiques explicites **45**
3.1 Espace d'état, champ de vecteurs et flot 45
3.1.1 Un modèle élémentaire de population 45
3.1.2 Existence, unicité, flot 47
3.1.3 Remarque sur l'espace d'état 55
3.1.4 Résolution numérique 56
3.1.5 Comportements asymptotiques 58
3.1.6 L'étude qualitative ou le contenu des modèles 62

3.2 Points d'équilibre 62
3.2.1 Stabilité et fonction de Lyapounov 63
3.2.2 Les systèmes linéaires 68
3.2.3 Lien avec le linéaire tangent 71
3.3 Systèmes dynamiques discrets 74
3.3.1 Point fixe et stabilité 74
3.3.2 Les systèmes linéaires discrets 75
3.4 Stabilité structurelle et robustesse 76
3.5 Théorie des perturbations 79
3.5.1 Les perturbations singulières 80
3.5.2 Moyennisation 85
3.6 Problèmes 87

4 Commandabilité et observabilité 91
4.1 Commandabilité non linéaire 92
4.1.1 Définition 92
4.1.2 Intégrale première 93
4.2 Commandabilité linéaire 95
4.2.1 Matrice de commandabilité 95
4.2.2 Invariance 97
4.2.3 Un exemple 99
4.2.4 Critère de Kalman et forme de Brunovsky 100
4.2.5 Planification et suivi de trajectoires 103
4.2.6 Linéarisation par bouclage 106
4.3 Observabilité non linéaire 110
4.3.1 Définition 110
4.3.2 Critère 111
4.3.3 Observateur, estimation, moindre carré 113
4.4 Observabilité linéaire 114
4.4.1 Le critère de Kalman 114
4.4.2 Observateurs asymptotiques 116
4.4.3 Observateur réduit de Luenberger 117
4.5 Observateur-contrôleur linéaire 118
4.6 Problèmes 119

5 Annexe : Systèmes semi-implicites et inversion 127
5.1 Systèmes semi-implicites 129
5.1.1 Un exemple 129
5.1.2 Le cas général 132
5.1.3 Linéaire tangent 135
5.1.4 Résolution numérique 136
5.2 Inversion et découplage 137
5.2.1 Un exemple 138
5.2.2 Le cas général 140

II Analyse Fréquentielle 145

6 Représentation fréquentielle 149
6.1 Système dynamique linéaire . . . 149
6.2 Transformations de Laplace et de Fourier . . . 150
6.3 Calcul symbolique . . . 152
6.4 Réalisation en variables d'état d'un système . . . 154
6.5 Systèmes mono-entrée, mono-sortie . . . 157
6.6 Systèmes à déphasage minimal . . . 158
6.7 Diagramme de Bode . . . 160
6.8 Systèmes du second ordre . . . 162

7 Stabilisation des systèmes bouclés 165
7.1 Bouclage sur la sortie. Système équivalent . . . 165
7.2 Stabilité du système bouclé . . . 166
7.3 Critère de Nyquist . . . 167
7.4 Aspects pratiques du calcul . . . 169
7.5 Analyse basée sur l'abaque de Black . . . 171
7.6 Synthèse P.I.D., avance et retard de phase . . . 174
7.7 Loop shaping . . . 177

8 Systèmes positifs réels 179
8.1 Positivité de l'opérateur d'entrée-sortie . . . 179
8.2 Caractérisations de la positivité . . . 179
8.3 Coercivité de la fonction potentiel . . . 182
8.4 Critère de Popov et critère du cercle . . . 183

I Méthodes Numériques en Commande Optimale 7

1 Temps minimal : systèmes linéaires 9
1.1 Introduction . . . 9
1.2 Un problème d'alunissage . . . 9
1.3 Existence de solutions . . . 10
1.3.1 Position du problème . . . 10
1.3.2 Résultats d'existence . . . 12
1.4 Conditions d'optimalité . . . 14
1.4.1 Séparation de l'ensemble accessible de la cible . . . 14
1.4.2 Critère linéaire sur l'état final . . . 16
1.4.3 Etat adjoint et principe du minimum . . . 19
1.5 Exemples et classes particulières . . . 20
1.5.1 Contraintes de bornes sur la commande . . . 20
1.5.2 Cas de l'oscillateur harmonique . . . 23
1.5.3 Stabilisation d'un pendule inversé . . . 24

1.5.4 Cibles épaisses 25

2 Temps minimal : systèmes non linéaires 27
2.1 Présentation du problème 27
2.1.1 Un exemple 27
2.1.2 Spécification du problème 28
2.1.3 Existence de solutions 29
2.2 Conditions d'optimalité 30
2.2.1 Un résultat général 30
2.2.2 Arc singulier 31
2.3 Applications 34
2.3.1 Pendule 34
2.3.2 Avion à trajectoire horizontale 35
2.4 Démonstration du résultat principal 37
2.5 Notes 43

3 Commande optimale : l'approche HJB 45
3.1 Cadre 45
3.2 Valeur fonction de l'état 46
3.2.1 Principe de programmation dynamique 46
3.2.2 Equation de Hamilton-Jacobi-Bellman 48
3.2.3 Continuité uniforme de la valeur 50
3.3 Commande optimale 52
3.4 Solution de viscosité 53
3.4.1 Notion de solutions de viscosité 54
3.4.2 Théorème de comparaison 56
3.5 Temps d'arrêt et commande impulsionnelle 59
3.5.1 Problèmes avec temps d'arrêt 60
3.5.2 Commande impulsionnelle 61
3.6 Notes 64

4 Résolution numérique de l'équation HJB 65
4.1 Motivation : problème continu 65
4.2 Schémas décentrés et extensions 67
4.2.1 Dimension d'espace $n = 1$ 67
4.2.2 Forme de point fixe contractant 68
4.2.3 Dimension d'espace quelconque 69
4.2.4 Discrétisation par triangulation 71
4.3 Convergence des schémas et essais numériques 72
4.3.1 Un argument élémentaire de convergence 72
4.3.2 Estimation d'erreur 73
4.3.3 Equation eikonale 76
4.3.4 Problème d'alunissage 78
4.4 Notes 78

5 Commande optimale stochastique **79**
5.1 Chaînes de Markov commandées 79
5.1.1 Quelques exemples 79
5.1.2 Chaînes de Markov et valeurs associées 79
5.1.3 Quelques lemmes 81
5.1.4 Principe de Programmation dynamique 82
5.1.5 Problèmes à horizon infini 83
5.1.6 Algorithmes numériques 85
5.1.7 Problèmes de temps de sortie 87
5.1.8 Problèmes avec décision d'arrêt 88
5.1.9 Un algorithme implémentable 90
5.2 Problèmes en temps et espace continus 92
5.2.1 Position du problème 92
5.2.2 Problème discrétisé en temps 93
5.2.3 Schémas monotones : dimension 1 95
5.2.4 Différences finies classiques 97
5.2.5 Différences finies généralisées 100
5.2.6 Analyse de la condition de consistance forte 102
5.3 Notes 104

Troisième partie

Méthodes Numériques en Commande Optimale

Chapitre 9

Temps minimal : systèmes linéaires

9.1 Introduction

Lors de la conception du transfert d'un système dynamique commandé vers un point de l'espace d'état, il est nécessaire de prendre en compte plusieurs critères, en général en conflit les uns avec les autres, dont les principaux sont :
- Le temps de transfert,
- L'énergie dépensée,
- L'écart par rapport à une trajectoire de référence,
- La robustesse par rapport à des perturbations,
- La complexité du problème de calcul de la commande,
- La simplicité de mise en œuvre en temps réel.

Les poids respectifs de ces critères dépendent de chaque application. Dans les chapitres suivants, nous allons nous concentrer sur le problème de transfert en temps minimal.

Le plan du chapitre est le suivant. Nous discutons l'exemple du problème d'alunissage en section 1.2. L'existence de solutions est analysée en section 1.3, et les conditions d'optimalité en section 1.4. Enfin la théorie est appliquée à plusieurs exemples en section 1.5.

9.2 Un problème d'alunissage

Dans sa phase finale, et en négligeant la gravité, une manœuvre d'alunissage peut se modéliser par l'équation

$$\ddot{h}(t) = m^{-1}u(t), \quad t \geq 0, \tag{9.1}$$

où h est l'altitude, $m > 0$ la masse de l'engin, et u la poussée nette (après déduction de la pesanteur). On notera $v := \dot{h}$, et on impose la contrainte $u(t) \in [-1, 1]$ à tout instant. Le problème est d'amener l'engin à vitesse et altitude nulle en un temps minimal.

La situation physique est celle où l'altitude initiale est positive. La solution intuitive est de fixer d'abord $u = -1$ pour se rapprocher le plus vite possible de la cible, jusqu'à atteindre un point où on commute à $u = 1$ (freinage maximum).

Nous allons résoudre graphiquement ce problème de transfert par une commande ne prenant que les valeurs ± 1, et changeant de signe au plus une fois. La théorie développée ultérieurement permettra de montrer que que pour ce problème, ces commandes réalisent le transfert en temps minimal (voir la remarque 1.31).

Soit (h_0, v_0) la condition initiale. Calculons d'abord les commandes permettant d'atteindre la cible avec une commande constante égale à ± 1. Si $u(t)$ vaut 1 pour tout $t \geq 0$, alors

$$h(t) = h_0 + tv_0 + \tfrac{1}{2}t^2, \quad v(t) = v_0 + t, \quad t \geq 0. \tag{9.2}$$

La trajectoire atteint la cible au temps $T > 0$ ssi $v_0 = -T$ et $h_0 = \frac{1}{2}T^2$. Si $u(t)$ vaut -1 pour tout $t \geq 0$, alors

$$h(t) = h_0 + tv_0 - \tfrac{1}{2}t^2, \quad v(t) = v_0 - t, \quad t \geq 0. \tag{9.3}$$

La trajectoire atteint la cible au temps $T > 0$ ssi $v_0 = T$ et $h_0 = -\frac{1}{2}T^2$. Les deux demi paraboles sont tracées en trait plein sur la figure 1.1.

Si la condition initiale se trouve sous la courbe en traits pleins, la trajectoire obtenue avec $u = 1$ permet d'atteindre le lieu des point pouvant être transférés à 0 par une commande égale à -1 ; si la condition initiale se trouve au dessus, la trajectoire obtenue avec $u = -1$ permet d'atteindre le lieu des point pouvant être transférés à 0 par une commande égale à 1. Il est facile de vérifier que toutes les commandes égales à ± 1 et changeant de signe au plus une fois sont de ce type.

La courbe en traits pleins est le *lieu de changement de signe* ; elle partage l'espace d'état en deux zones où la commande est constante. Nous avons réalisé (comme cela sera justifié ultérieurement) la *synthèse*, c'est à dire le calcul de la commande optimale en tout point de l'espace d'état : la commande s'exprime comme fonction de retour d'état, ou *feedback*

$$u(h, v) = \begin{cases} 1 & \text{si} \quad v \leq 0 \quad \text{et} \quad h \leq \frac{1}{2}v^2, \\ 1 & \text{si} \quad v > 0 \quad \text{et} \quad h < -\frac{1}{2}v^2, \\ -1 & \text{sinon.} \end{cases} \tag{9.4}$$

9.3 Existence de solutions

9.3.1 Position du problème

Considérons le système dynamique linéaire

$$\dot{x}(t) = Ax(t) + Bu(t); \quad t \geq 0, \quad x(0) = x^0, \tag{9.5}$$

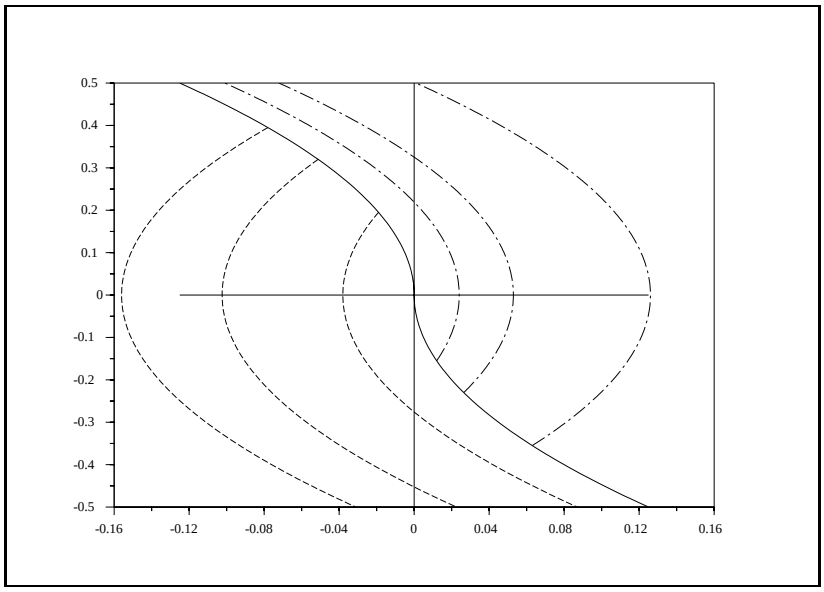

FIG. 9.1 – Trajectoires en temps minimal

avec $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, et A et B de taille respectivement $n \times n$ et $n \times m$. La commande, fonction mesurable $\mathbb{R}_+ \to \mathbb{R}^m$, doit respecter une contrainte du type

$$u(t) \in U, \quad \text{p.p.} \quad t \geq 0, \tag{9.6}$$

où U est un ensemble *convexe, compact* et tel que[1] $0 \in \operatorname{int} U$. On appellera solution de (1.5) toute fonction $x(t)$ continue en temps, à valeurs dans $\mathbb{R}^n$, telle que

$$x(t) = x^0 + \int_0^t [Ax(s) + Bu(s)]\mathrm{d}s, \quad \text{pour tout } \ t > 0. \tag{9.7}$$

On peut vérifier (par le même argument de point fixe contractant que dans le cas de seconds membres continus) que (1.5) possède une solution unique.

Soit C une partie convexe fermée et non vide de l'espace d'état $\mathbb{R}^n$, appelée la *cible*. On considère le *problème de transfert en temps minimal* d'un état initial $x^0 \notin C$ donné à la cible :

$$\inf_{(x,u,T)} T; \quad x(T) \in C; \quad x(0) = x^0; \quad (x,u) \quad \text{satisfont (1.5)-(1.6).} \tag{9.8}$$

[1] On notera $\operatorname{int} U$ l'intérieur de U, défini comme l'ensemble des $u \in U$ tels que, pour $\rho > 0$ assez petit, la boule $B(u,\rho)$ de centre u et rayon ρ est contenue dans U.

Remarque 9.1 La présence de la contrainte sur la commande est essentielle. En effet, si le système est commandable, le transfert de x^0 à un point quelconque de la cible est possible en un temps arbitrairement petit en l'absence de cette contrainte.

On dira que le problème en temps minimal (1.8) est *réalisable* s'il existe une commande transférant l'état initial à la cible. Le *temps minimal* noté $T(x^0)$ est la valeur de l'infimum dans (1.8), et vaut par définition $+\infty$ si le problème n'est pas réalisable.

On dit que la commande $\bar{u}$, fonction mesurable de $[0, T(x^0)]$ à valeurs dans U p.p., est une *commande en temps minimal* si elle réalise le transfert de l'état initial à la cible.

Rappelons la formule

$$x(t) = e^{tA}x^0 + \int_0^t e^{(t-s)A}Bu(s)ds, \quad t \geq 0, \tag{9.9}$$

où $e^A := \sum_{i=0}^{\infty} A^i/i!$. Etant donnés $t \geq 0$ et $x^0 \in I\!R^n$, on désigne par $\mathcal{R}(t, x^0)$ l'*ensemble des états accessibles* au temps t en partant de x^0 au temps $t = 0$. Autrement dit,

$$\mathcal{R}(t, x^0) = \left\{ e^{tA}x^0 + \int_0^t e^{(t-s)A}Bu(s)ds;\ u(s) \in U,\ \text{p.p. } s \in [0, t] \right\}. \tag{9.10}$$

Soit $T > 0$. On vérifie facilement que $\cup_{0 \leq t \leq T} \mathcal{R}(t, x^0)$ est borné. Il est clair que $\mathcal{R}(T, x^0)$ est convexe ; les propriétés de fermeture sont étudiées dans la section suivante à l'occasion de l'analyse de l'existence de solutions pour le problème (1.8).

9.3.2 Résultats d'existence

Théorème 9.2 *Si le problème en temps minimal* (1.8) *est réalisable, alors il existe une commande optimale.*

La démonstration du théorème nécessite un résultat d'analyse fonctionnelle que nous admettrons (voir Brézis [19]) :

Lemme 9.3 *Soit E une partie convexe fermée d'un espace de Hilbert F. De toute suite bornée $\{e_i\}$ dans E, on peut extraire une sous suite $\{e_j\}_{j \in J}$ qui converge faiblement vers un certain $\bar{e} \in E$, au sens où, pour toute forme linéaire continue L sur F, on a $\lim_{j \in J} L(e_j) = L(\bar{e})$.*

Lemme 9.4 *Soient $\bar{\tau} > 0$, $\tau_k \to \bar{\tau}$, et $x_k \in \mathcal{R}(\tau_k, x^0)$. Alors tout point d'adhérence x^d de $\{x_k\}$ appartient à $\mathcal{R}(\tau, x^0)$.*

Démonstration. On peut supposer que $x_k \to x^d$. Notons u_k une commande à valeurs p.p. dans U telle que l'état associé noté x_{u_k} vérifie $x_{u_k}(\tau_k) = x_k$. Comme U est compact, l'ensemble $\cup_{0 \leq t \leq \bar{\tau}+1} \mathcal{R}(t, x^0)$ est borné. L'équation d'état implique

donc que $\|\dot{x}_{u_k}\|_{L^\infty(0,\tau_k,\mathbb{R}^n)}$ est uniformément bornée par $L > 0$. On en déduit que ces fonctions sont lipschitziennes de constante L, et donc $x_{u_k}(\bar{\tau}) \to x^d$.

Par ailleurs la restriction de u_k à $[0, \bar{\tau}]$ est bornée dans l'ensemble convexe fermé $L^2(0, \bar{\tau}, U)$. Extrayant si nécessaire une sous suite on déduit du lemme 1.3 la convergence faible de cette restriction vers un certain $\bar{u} \in L^2(0, \bar{\tau}, U)$. En particulier

$$x_k(\bar{\tau}) = \int_0^{\bar{\tau}} e^{(t-s)A} B u_k(s) ds \to \int_0^{\bar{\tau}} e^{(t-s)A} B \bar{u}(s) ds. \tag{9.11}$$

Comme $x^k(\bar{\tau}) \to x^d$, ceci implique que x^d est la valeur de l'état associé à $\bar{u}$ à l'instant $\bar{\tau}$ d'où la conclusion. ■

Démonstration du théorème 1.2. Posons $\bar{T} := T(x^0)$. Par définition du temps minimal, il existe une suite décroissante $\{T_k\} \to \bar{T}$ telle que $\mathcal{R}(T_k, x^0) \cap C \neq \emptyset$, et donc il existe des commandes u^k, fonctions mesurables de $[0, T_k]$ à valeurs dans U p.p., telles que les états associés x^k vérifient $x^k(T_k) \in C$. Extrayant une sous-suite, on peut supposer que la suite bornée $\{x^k(T_k)\}$ converge vers un point x^d ; on conclut avec le lemme 1.4. ■

Notons l'ensemble des instants pour lesquels on peut atteindre la cible par

$$\mathcal{T}(x^0) := \left\{ t \geq 0; \quad \mathcal{R}(t, x^0) \cap C \neq \emptyset \right\}. \tag{9.12}$$

Cet ensemble, fermé d'après le lemme 1.4, a une structure simple dans deux cas particuliers.

Définition 9.5 On dira que la cible C est *viable* si, pour tout $x^d \in C$, il existe une commande à valeur p.p. dans U telle que le système (1.5) avec état initial $x(0) = x^d$ vérifie $x(t) \in C$ pour tout $t \geq 0$.

La cible est viable si elle est réduite à 0, et plus généralement si, pour tout $x^d \in C$, il existe $u \in U$ tel que $Ax^d + Bu = 0$. On peut donner des caractérisations de la viabilité basées sur la notion d'espace tangent à C, voir par exemple H. Frankowska [30, Section 1.3.5].

Proposition 9.6 *Si l'état initial est nul, ou si C est viable, alors $\mathcal{T}(x^0)$ est de la forme $[T(x^0), \infty[$.*

Démonstration. Notons d'abord que $T(x^0) \in \mathcal{T}(x^0)$ d'après le théorème 1.2. Si $x^0 = 0$, tout état accessible en temps t par une commande admissible $u = u(s)$ peut aussi être atteint en un temps $t' > t$ avec une commande u' nulle sur $[0, t' - t[$ et égale à $u'(s) = u(s - (t' - t))$ sur $[t' - t, t']$.

Si u transfère x^0 à $x^d \in C$ en un temps t, la viabilité de C implique l'existence d'une commande transférant x^0 à un point de C en tout temps $t' > t$, d'où la conclusion. ■

Remarque 9.7 L'oscillateur harmonique, présenté dans la section 1.5.2, est un exemple de système pour lequel, en général, si C n'est pas réduit à $\{0\}$, alors $\mathcal{T}(x^0)$ n'est pas de la forme $[T(x^0), \infty[$.

9.4 Conditions d'optimalité

Cette section établit des conditions nécessaires d'optimalité pour un problème de transfert en temps minimal du type (1.8). Ces conditions, suffisantes dans certains cas, permettront de résoudre complètement un certain nombre d'exemples.

9.4.1 Séparation de l'ensemble accessible de la cible

Dans cette section, on notera $\bar{T} := T(x^0)$ le temps minimal de transfert de x^0 à C. On suppose que $x^0 \notin C$, et donc $\bar{T} > 0$. Les conditions d'optimalité sont fondées sur la notion de *séparation d'ensembles convexes*.

Définition 9.8 On dit qu'une forme linéaire q sur $\mathbb{R}^n$ sépare deux parties C_1 et C_2 de $\mathbb{R}^n$ si $q \neq 0$ et

$$q \cdot x_1 \leq q \cdot x_2, \quad \text{pour tout} \quad x_1 \in C_1,\ x_2 \in C_2. \tag{9.13}$$

Théorème 9.9 *Il existe une forme linéaire séparant C de $\mathcal{R}(\bar{T}, x^0)$. Autrement dit, il existe $q \in \mathbb{R}^n$ non nulle telle que*

$$q \cdot y \leq q \cdot x, \quad \textit{pour tout} \quad y \in C \quad \textit{et} \quad x \in \mathcal{R}(\bar{T}, x^0). \tag{9.14}$$

Démonstration. Soit $\{T_k\}$ une suite strictement croissante de limite $\bar{T}$, telle que $T_0 > 0$. Par définition du temps minimal, $\mathcal{R}(T_k, x^0) \cap C = \emptyset$. Nous allons séparer C de $\mathcal{R}(T_k, x^0)$, puis passer à la limite. Notons $\mathrm{dist}(\cdot, C)$ la distance (euclidienne) à l'ensemble C :

$$\mathrm{dist}(x, C) := \inf\{\|x - y\|; \quad y \in C\}. \tag{9.15}$$

Cette fonction continue atteint son minimum sur le compact $\mathcal{R}(T_k, x^0)$ en un point x^k. Puisque C est fermé, il existe $y^k \in C$ tel que $\mathrm{dist}(x^k, C) = \|y^k - x^k\|$. Posons

$$q^k := (x^k - y^k)/\|x^k - y^k\|. \tag{9.16}$$

Montrons que

$$q^k \cdot y \leq q^k \cdot y^k \leq q^k \cdot x^k \leq q^k \cdot x, \quad \text{pour tout} \quad y \in C,\ \ x \in \mathcal{R}(T_k, x^0). \tag{9.17}$$

La seconde inégalité est conséquence directe de (1.16). La première traduit le fait que y^k est la projection de x^k sur C. Enfin il est facile de vérifier que x^k est la projection de y^k sur $\mathcal{R}(T_k, x^0)$, ce que traduit la troisième inégalité.

Or $\{x^k\}$ est bornée, et $\{y^k\}$ l'est donc aussi. Extrayant une sous suite si nécessaire, on peut supposer que x^k converge vers x^d, avec $x^d \in \mathcal{R}(\bar{T}, x^0)$ d'après le lemme 1.4, que y^k converge vers $\bar{y}$, avec $\bar{y} \in C$ puisque C est fermé, et enfin que q^k converge vers $\bar{q}$, forme linéaire de norme 1.

De plus, tout $x \in \mathcal{R}(\bar{T}, x^0)$ est limite d'une suite de points de $\mathcal{R}(T_k, x^0)$: il suffit de prolonger la commande transférant à x en un temps $\bar{T}$ sur $[T_k, \bar{T}]$.

Passant à la limite dans (1.17), nous obtenons donc

$$\bar{q} \cdot y \leq \bar{q} \cdot \bar{y} \leq \bar{q} \cdot x^d \leq \bar{q} \cdot x, \quad \text{pour tout} \quad y \in C \quad \text{et} \quad x \in \mathcal{R}(\bar{T}, x^0), \tag{9.18}$$

d'où le résultat. ■

Remarque 9.10 On peut vérifier que les points x^d et $\bar{y}$ construits dans la démonstration précédente coïncident.

La *frontière* d'une partie K de $\mathbb{R}^n$ est notée $\partial K := K \setminus \operatorname{int} K$.

Remarque 9.11 La démonstration n'utilise pas le fait que $\bar{T}$ est le temps minimal de transfert, mais seulement l'existence d'une suite $\{T_k\}$ qui converge vers $\bar{T}$, et telle que $\mathcal{R}(T_k, x^0) \cap C = \emptyset$. La propriété de séparation est donc satisfaite par tout élément de la frontière $\partial\mathcal{T}(x^0)$ de $\mathcal{T}(x^0)$. Ce n'est donc pas une condition suffisante d'optimalité si $\partial\mathcal{T}(x^0) \neq \{T(x^0)\}$, autrement dit si $\mathcal{T}(x^0) \neq [T(x^0), \infty[$. On verra dans le théorème 1.23 que sous certaines hypothèses supplémentaires ces conditions sont suffisantes.

Lemme 9.12 *Tout état final $x(\bar{T})$ associé à une commande en temps minimal appartient aux frontières des ensembles C et $\mathcal{R}(\bar{T}, x^0)$.*

Démonstration. Il suffit de combiner le théorème 1.9 et le lemme qui suit[2]. ■

Lemme 9.13 *Soit une partie $\mathcal{C}$ convexe de $\mathbb{R}^n$ contenant y. Alors $y \in \operatorname{int}\mathcal{C}$ ssi il n'existe pas de forme linéaire séparant y de $\mathcal{C}$.*

Démonstration. Montrons d'abord que, si $y \in \operatorname{int}\mathcal{C}$, il n'existe pas de forme linéaire séparant y de $\mathcal{C}$. Soit $\rho > 0$ tel que $B(y, \rho) \subset \mathcal{C}$. S'il existe une forme linéaire q séparant y de $\mathcal{C}$, posons $\varepsilon := \rho/\|q\|$. Alors $y - \varepsilon q \in \mathcal{C}$, et donc avec (1.13), $0 \geq \varepsilon\|q\|^2$ ce qui donne la contradiction recherchée.
b) Soit maintenant $y \in \partial\mathcal{C}$; il faut construire une forme linéaire séparant y de $\mathcal{C}$.

Notons $\bar{\mathcal{C}}$ la fermeture de $\mathcal{C}$. Montrons que $y \in \partial\bar{\mathcal{C}}$. Dans le cas contraire, puisque $y \in \bar{\mathcal{C}}$, on aurait $y \in \operatorname{int}\bar{\mathcal{C}}$, ce qui, grâce à la convexité de $\mathcal{C}$, impliquerait $y \in \operatorname{int}\mathcal{C}$, contraire à l'hypothèse.

Il existe donc une suite $y^k \to y$, avec $y^k \notin \bar{\mathcal{C}}$ pour tout k. Notons z^k la projection (orthogonale) de y^k sur $\bar{\mathcal{C}}$, et $q^k := z^k - y^k$. Puisque $y^k \notin \bar{\mathcal{C}}$, on a $q^k \neq 0$. Si $x \in \bar{\mathcal{C}}$ et $\alpha \in]0, 1]$, on a $z^k + \alpha(x - z^k) \in \bar{\mathcal{C}}$, et donc

$$0 \leq \lim_{\alpha \downarrow 0} \frac{\|z^k + \alpha(x - z^k) - y^k\|^2 - \|z^k - y^k\|^2}{2\alpha} = q^k \cdot (x - z^k). \tag{9.19}$$

[2]On peut admettre en première lecture ce lemme classique d'analyse convexe.

Or $q^k \cdot (z^k - y^k) = \|q^k\|^2 > 0$, donc $q^k \cdot (x - y^k) \geq 0$ pour tout $x \in \bar{\mathcal{C}}$. Ceci prouve que q^k sépare y^k de $\mathcal{C}$. Extrayant une sous suite si nécessaire, on peut supposer que $q^k/\|q^k\|$ converge vers $q \in I\!R^n$, de norme 1. Passant (à x fixé) à la limite dans l'inégalité

$$\frac{q^k}{\|q^k\|} \cdot y^k \leq \frac{q^k}{\|q^k\|} \cdot x, \quad \text{pour tout} \quad x \in \mathcal{C}, \tag{9.20}$$

on obtient la relation désirée. ■

A vrai dire, l'appartenance à la frontière de l'ensemble accessible n'est une information utile que si le système est commandable. Dans le cas contraire, le lemme ci-dessous nous indique en effet que tout point accessible en temps T est un point frontière de $\mathcal{R}(T, x^0)$.

Lemme 9.14 *Pour tout $T > 0$, l'ensemble $\mathcal{R}(T, x^0)$ est d'intérieur non vide ssi le système est commandable.*

Démonstration. Si le système n'est pas commandable, soit $w \in I\!R^n$ un élément non nul du noyau à gauche de la matrice de commandabilité. Nous savons que la forme linéaire $x \to w \cdot e^{-tA}x$ est une intégrale première ; donc $\mathcal{R}(T)$ est d'intérieur vide.

Si le système est commandable, soit $\rho > 0$ tel que $B(0, \rho) \subset U$, et soit e_j un vecteur de base de $I\!R^n$. Il existe une commande continue u^j amenant l'état 0 à l'état e_j en un temps T. Posons $M := \max_j |u^j|_{L^\infty(0,T)}$. Alors $\pm\rho M^{-1} u_j$ est admissible pour tout j, et amène x^0 à $e^{TA}x_0 \pm \rho M^{-1} e_j$ en un temps T ; donc $\mathcal{R}(T, x^0) \supset e^{TA}x_0 + \rho M^{-1} E$, où E désigne l'enveloppe convexe de $\{\pm e_1, \ldots, \pm e_n\}$. Puisque E est d'intérieur non vide, il en est de même pour $\mathcal{R}(T, x^0)$. ■

9.4.2 Critère linéaire sur l'état final

Dans cette section nous allons oublier (provisoirement) les problèmes de transfert en temps minimal, pour nous consacrer à l'étude du problème suivant :

$$\text{Inf } q \cdot x(T); \quad (x, u) \quad \text{satisfont (1.5)-(1.6)}, \tag{9.21}$$

où $q \in I\!R^n$ et l'horizon T sont donnés. La propriété de séparation (1.14) implique en effet qu'une commande en temps minimal est solution d'un tel problème, lorsque q est la forme linéaire séparante et $T = T(x^0)$.

Ce problème est convexe : il a un critère linéaire et des contraintes ponctuelles sur la commande. On peut caractériser ses solutions par un système d'optimalité faisant intervenir le *pseudo-hamiltonien* $H : I\!R^n \times I\!R^m \times I\!R^n \to I\!R$ défini par

$$H(x, u, p) := p \cdot (Ax + Bu), \tag{9.22}$$

et l'*état adjoint* $p \in C([0, T], I\!R^n)$, solution de

$$\begin{cases} -\dot{p}(t) &= H_x(x(t), u(t), p(t)) = A^\top p(t), \quad t \in [0, T], \\ p(T) &= q. \end{cases} \tag{9.23}$$

On dira que la commande u, fonction mesurable de $[0,T]$ vers U, vérifie le *Principe du minimum* pour le problème (1.21) si elle satisfait la relation

$$H(x(t),u(t),p(t)) = \inf_{v\in U} H(x(t),v,p(t)), \quad \text{p.p.} \quad t\in[0,T]. \tag{9.24}$$

Noter que (1.24) équivaut à $p(t)\cdot B(v-u(t))\geq 0$, pour tout $v\in U$, p.p. $t\in[0,T]$.

Théorème 9.15 *Une commande u, fonction mesurable de $[0,T]$ vers U, est solution de* (1.21) *ssi elle vérifie le principe du minimum.*

Démonstration. Soit u' une autre commande à valeur p.p. dans U. Posons

$$u''(t) := u(t) \text{ si } H(x(t),u(t),p(t)) \leq H(x(t),u'(t),p(t)), \quad u'(t) \quad \text{sinon.} \tag{9.25}$$

Alors u'' est mesurable, à valeur dans U p.p. ; notons x'' l'état associé. Puisque $x(0) = x''(0) = x^0$, on a avec (1.5) et (1.23), après simplification,

$$\begin{aligned} 0 &\leq q\cdot(x''(T)-x(T)) \\ &= \int_0^T \frac{\mathrm{d}}{\mathrm{d}t}[p(t)\cdot(x''(t)-x(t))]\,\mathrm{d}t = \int_0^T p(t)\cdot B(u''(t)-u(t))\,\mathrm{d}t. \end{aligned} \tag{9.26}$$

Or $p(t)\cdot B(u''(t)-u(t))\leq 0$ p.p., donc $p(t)\cdot Bu'(t)\geq p(t)\cdot Bu''(t)\geq p(t)\cdot Bu(t)$ p.p. comme on voulait le montrer. ■

On utilisera le lemme suivant dont la démonstration est immédiate.

Lemme 9.16 *Soient a et b deux fonction réelles d'une variable u. Alors*

$$\left|\inf_{u\in\mathcal{U}} a(u) - \inf_{u\in\mathcal{U}} b(u)\right| \quad \leq \quad \sup_{u\in\mathcal{U}} |a(u)-b(u)|, \tag{9.27}$$

et

$$\inf_{u\in\mathcal{U}} a(u) - \inf_{u\in\mathcal{U}} b(u) \quad \leq \quad \sup_{u\in\mathcal{U}} (a(u)-b(u)). \tag{9.28}$$

Proposition 9.17 *Si une commande u satisfait le principe du minimum sur $[0,T]$, alors l'application $t\to H(x(t),u(t),p(t))$ est essentiellement constante*[3]*.*

Démonstration. Posons

$$h(t) := \inf_{v\in U} H(x(t),v,p(t)). \tag{9.29}$$

Le lemme 1.16 implique

$$|h(t')-h(t)| \leq |p(t')\cdot Ax(t') - p(t)\cdot Ax(t)| + \sup_{v\in U} |(p(t')-p(t))\cdot Bv|\,. \tag{9.30}$$

[3]Autrement dit, constante à un ensemble de mesure nulle près.

On en déduit facilement l'existence de $M > 0$ tel que

$$|h(t') - h(t)| \leq M \left(\|x(t') - x(t)\| + \|p(t') - p(t)\| \right). \tag{9.31}$$

Or x et p sont lipschitziens, donc h l'est aussi, et a en conséquence une dérivée dans $L^\infty(0,T)$. De plus $h(t) = h(0) + \int_0^T \dot{h}(t)\mathrm{d}t$, pour tout $t \in [0,T]$. Montrons que $\dot{h}(t) = 0$ presque partout. Soit t_0 un point où h est dérivable. Le principe du minimum implique

$$\begin{aligned} \dot{h}(t_0) &\leq \lim_{t>t_0} \frac{H(x(t), u(t_0), p(t)) - H(x(t_0), u(t_0), p(t_0))}{t - t_0} \\ &= D_x H(x(t_0), u(t_0), p(t_0))\dot{x}(t_0) + D_p H(x(t_0), u(t_0), p(t_0))\dot{p}(t_0) = 0. \end{aligned} \tag{9.32}$$

De même, avec $t < t_0$ on montre que $\dot{h}(t_0) \geq 0$ et donc $\dot{h}$ est nulle p.p., de sorte que h est constante. Or $h(t) = H(x(t), u(t), p(t))$ p.p. d'après le principe du minimum, d'où la conclusion. ■

Soit $p(t)$ solution de (1.23). Alors $B^\top p(t)$ est une fonction analytique de t, donc soit est identiquement nulle, soit a un nombre fini de zéros sur $[0,T]$. Dans ce dernier cas on déduit du principe du minimum nombre de renseignements sur la solution de (1.21).

Définition 9.18 On dit que U est *strictement convexe* si, étant donné deux points *distincts* u_1 et u_2 de U, le segment[4] $]u_1, u_2[$ appartient à l'intérieur de U.

Exemple 9.19 Dans $\mathbb{R}^n$, la boule unité fermée pour la norme ℓ^p est strictement convexe si $1 < p < \infty$, mais pas si $p = 1$ ou $p = \infty$.

Théorème 9.20 *Soit p solution de* (1.23)*, avec $q \neq 0$. Alors*
(i) *Si le système est commandable, l'application $t \to B^\top p(t)$ n'est pas identiquement nulle.*
(ii) *Si $B^\top p(t)$ n'est pas identiquement nulle, toute solution u du problème à coût linéaire* (1.21) *est telle que $u(t) \in \partial U$ p.p. $t \in [0,T]$.*
(iii) *Si de plus l'ensemble U est strictement convexe, alors* (1.21) *a une solution unique, continue en tout instant t, sauf peut-être ceux (en nombre fini) où $B^\top p(t)$ est nul.*

Démonstration. (i) Supposons au contraire $B^\top p(t)$ identiquement nulle. Alors $0 = B^\top p(\bar{T}) = B^\top \dot{p}(\bar{T}) = \cdots$, d'où $q \cdot BA^i = 0$, pour $i = 1, \ldots, n-1$; autrement dit, q appartient au noyau à gauche de la matrice de commandabilité. Si le système est commandable, ceci implique $q = 0$, ce qui est impossible.
(ii) D'après le théorème 1.15, $u(t)$ doit minimiser la forme linéaire $v \to p(t) \cdot Bv$ sur U à tout instant. Sauf en un nombre fini de points, cette forme linéaire est non nulle, ce qui implique que $u(t)$ est point frontière de U.

[4] Ce segment est par définition $\{\alpha u_1 + (1-\alpha)u_2;\ \alpha \in]0,1[\}$.

(iii) Le minimum d'une forme linéaire sur un ensemble strictement convexe compact existe et est unique. Il est facile de vérifier qu'il dépend continûment de la forme linéaire si cette dernière n'est pas nulle, ce qui assure le point (iii). ■

9.4.3 Etat adjoint et principe du minimum

Revenons maintenant au problème de temps minimal (1.8). On dira que la commande u, fonction mesurable de $[0,T]$ vers U, vérifie le *Principe du minimum pour le problème* (1.8) si elle satisfait les relations suivantes :

$$\begin{cases} \dot{x}(t) &= Ax(t)+Bu(t), \quad t \geq 0, \\ x(0) &= x_0, \end{cases} \tag{9.33}$$

$$\begin{cases} -\dot{p}(t) &= A^\top p(t), \quad t \in [0,T], \\ p(T) &= q. \end{cases} \tag{9.34}$$

$$H(x(t),u(t),p(t)) = \inf_{v\in U} H(x(t),v,p(t)), \quad \text{p.p.} \quad t\in[0,T], \tag{9.35}$$

$$q\cdot y \leq q\cdot x(T), \quad \text{pour tout } \ y\in C; \quad x(T)\in C; \quad q\neq 0. \tag{9.36}$$

Le pseudo-hamiltonien dans (1.35) est toujours défini par (1.22). On reconnaît l'équation d'état et d'état adjoint, ainsi que la propriété de minimisation du pseudo-hamiltonien. Enfin (1.36) est conséquence de la propriété de séparation de la section 1.4.1. Définissons une *normale extérieure* à C en $z\in C$ comme un élément $q\in I\!R^n$ tel que

$$q\cdot y \leq q\cdot z, \quad \text{pour tout } \ y\in C. \tag{9.37}$$

Alors (1.36) dit que q est une normale extérieure *non nulle* à C en $x(T)$.

Des théorèmes 1.9 et 1.20, on déduit immédiatement le *résultat principal de ce chapitre*, qui exprime des conditions nécessaires d'optimalité :

Théorème 9.21
(i) *Toute solution u du problème de temps minimal* (1.8) *satisfait le principe du minimum* (1.33)-(1.36)*, avec $T = T(x^0)$, et $t \to H(x(t),u(t),p(t))$ a une valeur constante p.p. le long de la trajectoire optimale.*
(ii) *Si le système est commandable, toute solution u de* (1.8) *satisfait p.p. $u(t)\in\partial U$.*
(iii) *Si le système est commandable, et U est strictement convexe, alors* (1.8) *a une solution unique, continue en tout instant t, sauf peut-être ceux (en nombre fini) où $B^\top p(t)$ est nul.*

Exemple 9.22 Etudions le cas où U est la boule unité euclidienne fermée, qui est strictement convexe. Le minimum de $v\to r\cdot v$ sur U, pour $r\neq 0$, est atteint en $-r/\|r\|$. Donc si $B^\top p(t)$ n'est pas identiquement nulle, la commande en temps minimal vaut p.p. $u(t) = -B^\top p(t)/\|B^\top p(t)\|$.

Discutons maintenant la suffisance du principe du minimum.

Théorème 9.23 *On suppose U strictement convexe, le système commandable, et la cible viable. Alors une commande transférant le système de x^0 à C en en temps T réalise le transfert en temps minimal si et seulement elle satisfait le principe du minimum* (1.33)-(1.36).

Démonstration. D'après le théorème 1.20, ces conditions sont nécessaires. Réciproquement, supposons que la commande u satisfait (1.33)-(1.36). Le théorème 1.15 affirme que (1.33)-(1.35) caractérise les solutions du problème convexe (1.21). Soit u^* une autre commande transférant x^0 à la cible en un temps $T^* < T$. Prolongeant u^* sur $[T^*, T]$ grâce à la viabilité de C, par une commande encore notée u^*. On obtient le transfert de x_0 en un point $x^* \in C$ avec la commande u^*. Alors (1.36) implique que u^* est aussi solution de (1.21). Comme ce dernier a, en raison du théorème 1.21(iii), une solution unique, $u = u^*$ comme il fallait le montrer. ■

Remarque 9.24 La démonstration n'exclut pas l'inégalité $T(x^0) < T$. Si une commande satisfait le principe du minimum, le temps de transfert est donc le premier instant où l'état associé appartient à la cible.

Remarque 9.25 La remarque 1.11 montre que, si la cible n'est pas viable, le principe du minimum n'est pas une condition suffisante d'optimalité.

9.5 Exemples et classes particulières

Nous allons voir que les résultats précédents permettent de donner une solution explicite au problème de commande en temps optimal dans quelques cas particuliers importants.

9.5.1 Contraintes de bornes sur la commande

Nous reprenons dans cette section le problème de temps minimal, dans le cas où l'ensemble U est la boule unité de $\mathbb{R}^m$ muni de la norme infinie :

$$U = \{u \in \mathbb{R}^m;\ \ |u_i| \leq 1,\ i = 1, \ldots, m\}. \tag{9.38}$$

Cet ensemble est convexe et compact, d'intérieur contenant 0. Il n'est en revanche pas strictement convexe si $m > 1$. Le principe du minimum implique

$$u_i(t) = \begin{cases} -1 & \text{si } (B^\top p(t))_i > 0, \\ 1 & \text{si } (B^\top p(t))_i < 0. \end{cases} \tag{9.39}$$

Si $(B^\top p(t))_i = 0$, le principe du minimum ne donne pas d'informations sur $u_i(t)$.

Puisque p est solution de l'équation linéaire homogène (sans second membre) (1.23) de dimension n, il est de la forme

$$\pi_1(t)e^{\alpha_1 t} + \cdots + \pi_r(t)e^{\alpha_r t}, \tag{9.40}$$

où $\alpha_1, \dots, \alpha_r$ sont les opposées des valeurs propres *distinctes* de A (donc $r \leq n$) de multiplicité μ_i, et $\pi_i(t)$ est un polynôme de degré d_i, avec $d_i \leq \mu_i - 1$. Les fonctions $(B^\top p(t))_i$ sont également de la forme (1.40). Elles sont donc, sur $[0, \bar{T}]$, soit identiquement nulles, soit nulles en un nombre fini de points, et dans ce dernier cas le principe du minimum détermine u_i (sauf en ces points).

Lemme 9.26 *Soit u une commande amenant x^0 à x^d en un temps minimal $\bar{T}$, et p un état adjoint associé. Soit $i \in \{1, \dots, m\}$. Alors, soit $(B^\top p(t))_i$ est identiquement nul, soit u_i change de signe un nombre fini de fois. Dans ce dernier cas, toutes les commandes transférant l'état de x^0 à x^d en temps minimal ont même composante i, sauf peut-être aux instants de changement de signe.*

Si les valeurs propres de A sont réelles, on peut donner une estimation du nombre des points de changement de signe :

Lemme 9.27 *Toute fonction $\psi(t)$ non nulle, de la forme* (1.40)*, avec $\alpha_1, \dots, \alpha_r$ réels distincts et $\pi_i(t)$ polynôme réels de degré d_i, a au plus $d_1 + \cdots + d_r + r - 1$ zéros.*

Démonstration. Procédons par récurrence sur r. Si $r = 1$, $\psi(t) = \pi_1(t)e^{\alpha_1 t}$ a les mêmes zéros que π_1 ; ce dernier étant un polynôme de degré d_1, au au plus $d_1 = d_1 + r - 1$ racines sur $[0, T]$. Supposons maintenant le résultat vrai pour $r - 1$. Alors $\psi(t)$ a les même zéros que la fonction

$$e^{-\alpha_1 t}\psi(t) = \pi_1(t) + \pi_2(t)e^{(\alpha_2 - \alpha_1)t} + \cdots + \pi_r(t)e^{(\alpha_r - \alpha_r)t}. \tag{9.41}$$

La dérivée d'ordre $d_1 + 1$ de cette fonction est de la forme

$$\frac{\mathrm{d}^{(\mathrm{d}_1+1)}\psi(\mathrm{t})}{\mathrm{d}t^{(d_1+1)}} = \bar{\pi}_2(t)e^{(\alpha_2 - \alpha_1)t} + \cdots + \bar{\pi}_r(t)e^{(\alpha_r - \alpha_1)t}, \tag{9.42}$$

avec $\bar{\pi}_2(t), \dots, \bar{\pi}_n(t)$ polynômes de degré d_i. D'après notre construction par récurrence, elle a au plus $d_2 + \cdots + d_r + r - 2$ zéros. Or, entre deux zéros d'une fonction se trouve au moins un zéro de sa dérivée. Si la fonction ψ avait plus de $d_1 + \cdots + d_r + r - 1$ zéros, sa dérivée d'ordre $d_1 + 1$ aurait donc plus de $d_2 + \cdots + d_r + r - 2$ zéros, d'où une contradiction. ■

Proposition 9.28 *Supposons les valeurs propres de A réelles. Soit u une commande amenant x^0 à x^d en un temps minimal T, et p un état adjoint associé. Soit $i \in \{1, \dots, m\}$. Alors, soit $(B^\top p(t))_i$ est identiquement nul, soit u_i change de signe au plus $n - 1$ fois.*

Démonstration. Soient $\alpha_1, \dots, \alpha_r$ les opposées des valeurs propres *distinctes* de A de multiplicité μ_i. Alors $(B^\top p(t))_i$ est de la forme (1.40), avec $d_i \leq \mu_i - 1$, et a donc au plus $d_1 + \cdots + d_r + r - 1$ zéros. Mais

$$d_1 + \cdots + d_r + r - 1 \leq \mu_1 + \cdots + \mu_r - 1 = n - 1. \tag{9.43}$$

■

Discutons quelques exemples qui éclairciront les résultats ci-dessus.

Exemple 9.29 Considérons le problème de transfert en temps minimal de $x^0 = (1,1)^\top$ à $x^d = 0$, avec la dynamique

$$\dot{x}_1 = u_1, \quad \dot{x}_2 = 2u_2, \tag{9.44}$$

et les contraintes $|u_i(t)| \leq 1$, $t \in [0,T]$, $i = 1,2$. Il est clair que le temps minimal de transfert est $T = 1$; toute commande optimale u est telle que $u_1(t) = -1$ sur $[0,T]$; par contre on n'a pas d'unicité de $u_2(t)$. Comment cela se traduit-il sur le système d'optimalité ?

L'ensemble accessible au temps $T = 1$ est $\mathcal{R}(T,x^0) = [0,2] \times [-1,3]$. Les formes linéaires séparant 0 de $\mathcal{R}(T,x^0)$ sont de la forme $q = (q_1,0)$ avec $q_1 > 0$. Les états adjoints associés sont $p(t) = q = (q_1,0)$. Le principe du minimum impose donc $u_1(t) = -1$ sur $[0,1]$, mais n'impose rien sur u_2, sinon d'être à valeurs dans $[-1,1]$, et tel que $x_2(T) = 0$.

Exemple 9.30 Soit, pour $n \geq 1$, le système dynamique

$$\frac{\mathrm{d}^n}{\mathrm{d}\,t^n} z(t) = u(t), \quad t \in [0,T]. \tag{9.45}$$

Considérons le problème de transfert en temps minimal vers la position de repos ($z(t)$ nulle ainsi que ses dérivées jusqu'à l'ordre $n-1$) sous la contrainte $|u(t)| \leq 1$. Traduisons (1.45) en

$$\frac{\mathrm{d}}{\mathrm{d}t} x_i(t) = x_{i+1}(t), \quad i = 1,\dots,n-1, \quad \frac{\mathrm{d}}{\mathrm{d}t} x_n(t) = u(t), \quad t \in [0,T]. \tag{9.46}$$

La dynamique de l'état adjoint est

$$-\frac{\mathrm{d}}{\mathrm{d}t} p_i(t) = p_{i-1}(t), \quad i = 2,\dots,n, \quad -\frac{\mathrm{d}}{\mathrm{d}t} p_1(t) = 0, \quad t \in [0,T]. \tag{9.47}$$

En particulier, $\mathrm{d}^n p_n(t)/\mathrm{d}\,t^n = 0$, donc $p_n(t)$ est un polynôme de degré au plus $n-1$.

Le système est commandable, et U est strictement convexe, donc (théorème 1.20) $B^\top p(t) = p_n(t)$ n'est pas identiquement nulle et la commande optimale est unique. La dynamique a pour seule valeur propre 0. La proposition 1.28 implique que cette commande optimale change de signe au plus $n-1$ fois.

On peut vérifier que la réciproque est vraie : toute commande amenant x^0 à 0 (en un temps T a priori quelconque) et changeant de signe au plus $n-1$ fois est optimale. En effet, soient $t_1,\dots,t_r$ les instants de changement de signe, avec $r \leq n-1$. Posons $p(t) = \pm(t-t_1) \times \cdots \times (t-t_r)$. Alors p est un polynôme de degré $r \leq n-1$, donc satisfait l'équation de l'état adjoint, avec la condition finale $q = p(T) \neq 0$, et (suivant le signe choisi dans $\pm$) la commande satisfait le principe du minimum. L'optimalité de la commande est alors conséquence du théorème 1.23.

Remarque 9.31 La discussion précédente montre que les commandes construites dans l'étude du problème d'alunissage (section 1.2) sont optimales. Le cas $n = 3$, nettement plus complexe, est traité dans Lee et Markus [42, Chapitre 2].

9.5.2 Cas de l'oscillateur harmonique

Considérons maintenant le problème de transfert en temps minimal de x^0 à $x^d = 0$ de l'oscillateur harmonique

$$\ddot{z}(t) + \omega^2 z(t) = u(t), \quad t \in [0, T], \tag{9.48}$$

où $\omega > 0$, sous la contrainte $|u(t)| \leq 1$. La dynamique avec une commande $u(t) = u_0$ constante est périodique, de la forme

$$z(t) = \omega^{-2} u_0 + r \cos(\omega t + \varphi), \quad t \in [0, T]. \tag{9.49}$$

La trajectoire décrit, dans l'espace d'état $(z, v = \dot{z})$ un cercle de centre $(\omega^{-2} u_0, 0)$ et rayon r. Celui-ci, ainsi que la phase φ, sont déterminées par les conditions initiales. Le cercle est parcouru dans le sens des aiguilles d'une montre.

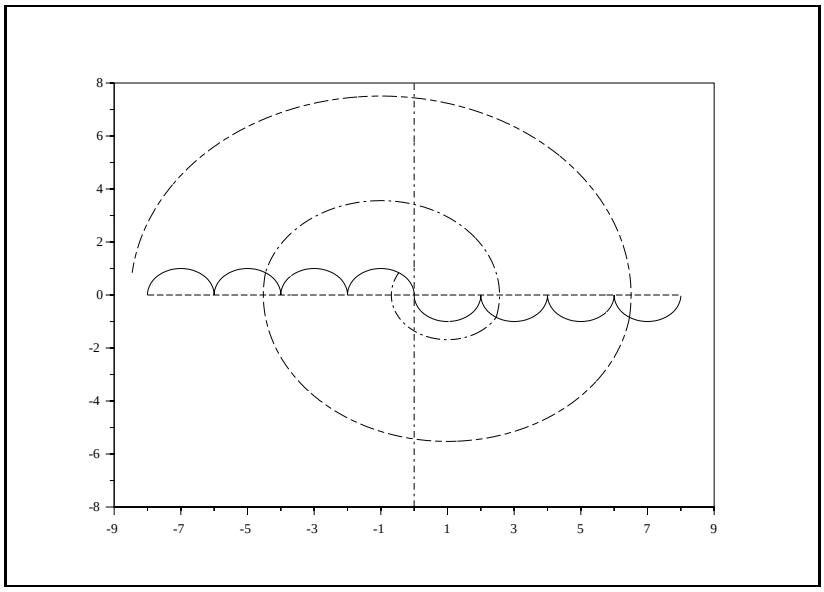

FIG. 9.2 – Oscillateur harmonique : trajectoires en temps minimal

Explicitons la dynamique quand la commande satisfait le principe du minimum. L'équation de l'état adjoint $p = (p_z, p_v)$ est

$$-\dot{p}_z(t) = -\omega^2 p_v(t); \quad -\dot{p}_v(t) = p_z(t); \quad t \in [0, T], \tag{9.50}$$

et donc p_v est de la forme

$$p_v(t) = r' \cos(\omega t + \varphi'). \tag{9.51}$$

Les instants de changement de signe de la commande sont espacés de π/ω, et la trajectoire de transfert en temps minimal est une sucession de demi-tours (sauf le dernier qui s'arrête quand la cible est atteinte) autour des points $(1, 0)$ et $(-1, 0)$, successivement. Le lieu de changement de signe est marqué en traits pleins sur la figure 1.2 ; il est formé d'une union de demi-cercles de rayon ω^{-2}. On a représenté en pointillé une trajectoire en temps minimal dans le cas $\omega = 1$. On vérifie aisément que, partant d'une condition initiale (z_0, v_0) quelconque, il existe une seule commande satisfaisant le principe du minimum, qui permet d'atteindre la cible. Il s'agit donc de la commande optimale. Celle-ci vaut $u = 1$ en dessous du lieu de changement de signe, et $u = -1$ au dessus.

9.5.3 Stabilisation d'un pendule inversé

La linéarisation de l'équation d'un problème de stabilisation du pendule inversé conduit à l'équation

$$\ddot{h}(t) = h(t) - u(t), \quad t \in [0, T]. \tag{9.52}$$

On considère le problème d'atteinte du point de vitesse et position nulles en un temps minimal. On notera $v = \dot{h}$ la vitesse. Le système non commandé a pour valeurs propres ± 1 et n'est donc pas stable. Il faut déterminer à partir de quels points on peut atteindre la cible. Pour cela on peut s'appuyer sur les portraits de phase quand u est constant. Celui-ci est la translation de celui obtenu quand $u = 0$ (voir la figure 1.3).

D'après la proposition 1.28 une trajectoire optimale atteint la cible avec $u(t) = \pm 1$ et au plus un changement de signe.

Points pouvant atteindre la cible avec $u = \pm 1$ constant. Quand u est constant, $h(t)$ est de la forme

$$h(t) = \alpha e^t + \beta e^{-t} + u. \tag{9.53}$$

Atteindre la cible au temps T signifie que

$$\alpha e^T + \beta e^{-T} = -u; \quad \alpha e^T - \beta e^{-T} = 0. \tag{9.54}$$

De là $\alpha = -\frac{1}{2} u e^{-T}$ et $\beta = -\frac{1}{2} u e^T$. On en déduit l'expression du point initial :

$$\begin{array}{rcl} h(0) = \alpha + \beta + u & = & u - u \cosh T; \\ v(0) = \alpha - \beta & = & u \sinh T. \end{array} \tag{9.55}$$

Pour $u = \pm 1$ on obtient le lieu tracé en traits pleins sur la figure 1.4.

La courbe est tangente en 0 à l'axe vertical à la cible, et a pour asymptotes les droite $h + v = \pm 1$. En effet, on a $\cosh^2 T - \sinh^2 T = 1$, et donc $\cosh T - \sinh T = (\cosh T + \sinh T)^{-1} = o(1)$ pour T grand.

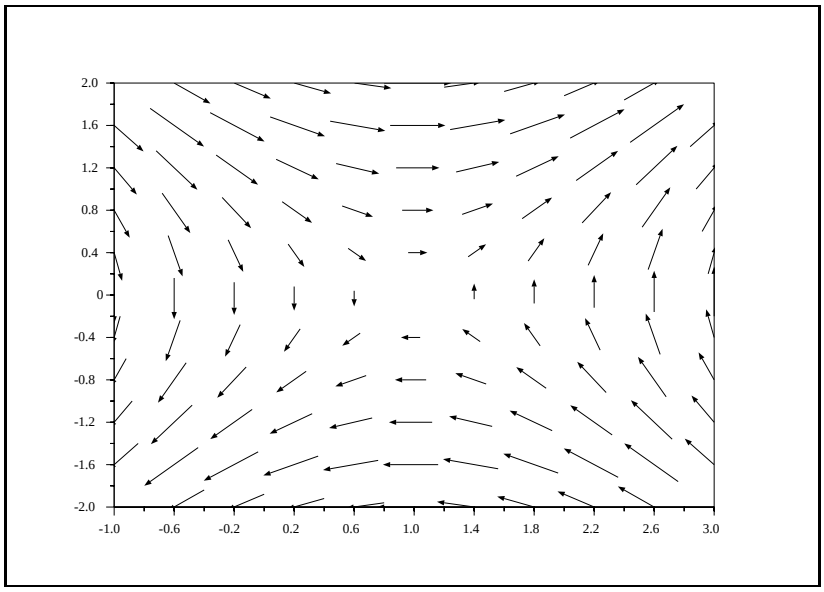

FIG. 9.3 – Pendule inversé : portrait de phase, $u = 1$

Points ne pouvant atteindre la cible Notons $v = \dot{h}$ et $\xi := h+v$. Alors $\dot{\xi} = \xi - u$. Donc si $|\xi| \geq 1$, $|\xi|$ ne peut diminuer au cours du temps. Ceci interdit d'atteindre la cible.

Points atteignant la cible avec un changement de signe de la commande On a déjà construit les trajectoires optimales sans changement de signe. Il suffit d'examiner quand les trajectoires obtenues avec $u = \pm 1$ rencontrent celles-ci. Or ces courbes sont obtenues par translation de $(u, 0)$ de celles pour $u = 0$ (cf les portraits de phase, questions 3). La figure 1.4 donne la représentation des trajectoires optimales.

9.5.4 Cibles épaisses

Soit x^d l'état final d'une trajectoire en temps minimal $\bar{T}$. Dans les exemples précédents, la cible était réduite à un point et la condition de séparation (1.36) se réduisait donc à la séparation de x^d et $\mathcal{R}(\bar{T}, x^0)$. Dans le cas dit de la cible épaisse, il faut prendre en compte le fait que q est une normale extérieure à C en x^d.

Exemple 9.32 Soit C égal à la boule unité fermée associée à la norme euclidienne. On sait (lemme 1.12) que $x^d \in \partial C$, soit $\|x^d\| = 1$. Toute normale extérieure à C en x^d est de la forme αx^d, avec $\alpha \in I\!R_+$. Or $q \neq 0$, et seule la direction de q importe et non son module. On peut donc supposer que $q = x(T)$. Le principe du minimum

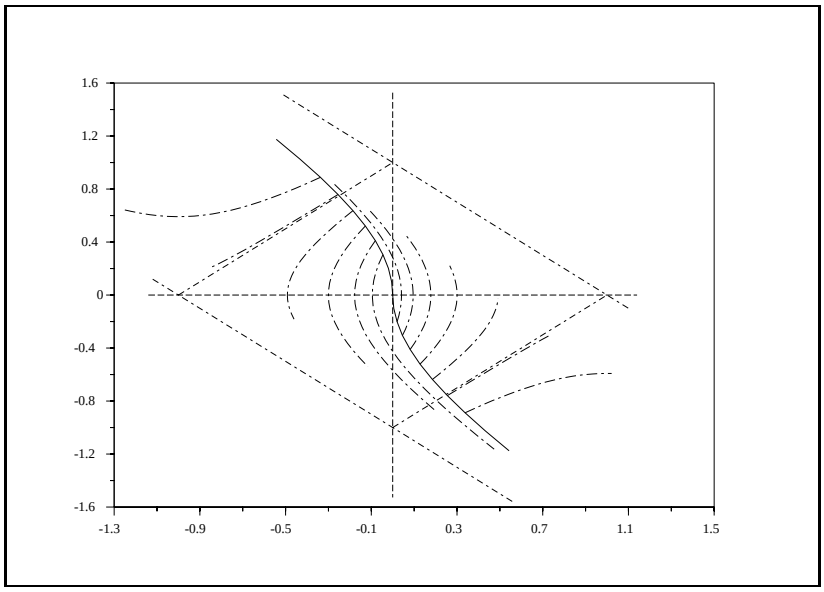

FIG. 9.4 – Synthèse des trajectoires en temps minimal

(1.33)-(1.36) équivaut alors à

$$\begin{cases} \dot{x}(t) &= Ax(t) + Bu(t), \quad t \geq 0, \\ x(0) &= x_0, \quad \|x(T)\| = 1, \end{cases} \tag{9.56}$$

$$\begin{cases} -\dot{p}(t) &= A^\top p(t), \quad t \in [0,T], \\ p(T) &= x(T). \end{cases} \tag{9.57}$$

$$H(x(t),u(t),p(t)) = \inf_{v\in U} H(x(t),v,p(t)), \qquad \text{p.p.} \quad t \in [0,T]. \tag{9.58}$$

Exemple 9.33 Supposons encore C égal à la boule unité fermée associée à la norme euclidienne, l'équation d'état étant $\ddot{z} = u$, avec $U = [-1,1]$. Notons $v = \dot{z}$, $x(T) = x^d = (z^d, v^d)$, et donc $q = (z^d, v^d)$. On a $-\dot{p}_z = 0$, $-\dot{p}_v = p_z$ et donc

$$p_z = z^d; \quad p_v = v^d + (\bar{T} - t)z^d. \tag{9.59}$$

La commande optimale vaut donc -1 et 1, respectivement, si (z^d, v^d) est dans le premier (resp. troisième) cadrant, et ne peut changer de signe que si v^d et z^d sont de signe différents. Intégrant en temps rétrograde, à partir du temps T avec $x(T)$ quelconque de norme 1, on obtient le lieu de changement de signe. McCausland [47, Section 6.6] donne une étude détaillée de ce problème.

Chapitre 10

Temps minimal : systèmes non linéaires

Ce chapitre aborde le problèmes de transfert en temps minimal en présence d'une dynamique non linéaire. L'ensemble accessible n'est plus convexe. Une linéarisation non standard de la dynamique, basée sur des perturbations en aiguilles, permettra cependant une extension du principe du minimum.

Par ailleurs, dans le cas d'une dynamique linéaire, la commande optimale est, si le système est commandable, p.p. sur la frontière des commandes admissibles. Il n'en est plus de même quand la dynamique est non linéaire, même si la commande entre linéairement dans l'équation d'état, comme le montre l'exemple de la section 2.1.1. Ceci nous amènera à introduire la théorie des arcs singuliers.

10.1 Présentation du problème

10.1.1 Un exemple

Nous allons discuter le problème du transfert en temps minimal vers une position donnée d'un avion dont la trajectoire est horizontale et rectiligne.

Les variables d'état sont la position y, la vitesse v, et la masse m de l'engin. Les forces en jeu sont liées à la gravité g, supposée constante, la traînée D, et la portance L (drag et lift). La portance doit équilibrer la gravité, soit $L = mg$; la traînée est liée à la portance via l'incidence, et cette relation a pour expression

$$D = Av^2 + B\frac{L^2}{g^2v^2}, \tag{10.1}$$

où A et B sont deux constantes positives. Eliminant la portance, il vient

$$D = D(v, m) = Av^2 + B\frac{m^2}{v^2}. \tag{10.2}$$

La commande u est le débit d'éjection des gaz, et la poussée est $c\,u$ avec $c > 0$ constant. L'équation d'état est donc

$$\dot{y}(t) = v(t); \quad \dot{v}(t) = \frac{cu - D(v,m)}{m(t)}; \quad \dot{m}(t) = -u. \tag{10.3}$$

Nous mènerons autant que possible les calculs avec une traînée $D = D(v,m)$ sans utiliser l'expression (2.2) qui varie d'un avion à l'autre. L'état initial est noté (y^0, v^0, m^0) et la cible est

$$C = \{(y,v,m); \quad y \geq y^d; \quad m \geq m^d\}. \tag{10.4}$$

On suppose que $y^d > y^0$, $m^0 > m^d$ et que $v^0 > 0$ (si $v^0 < 0$ le problème n'a pas de sens).

Cet exemple permet d'illustrer un phénomène typique des systèmes non linéaires. Il n'est pas nécessairement optimal de rechercher des vitesses élevées en raison du terme de traînée. Il peut donc y avoir une phase du vol où la commande en temps minimal se trouvera hors des bornes. Nous allons vérifier qu'il en est ainsi, et montrer comment calculer la trajectoire optimale, en section 2.3.2.

10.1.2 Spécification du problème

Nous considérons le système dynamique non linéaire

$$\dot{x}(t) = f(t, x(t), u(t)), \quad t \geq 0; \quad x(0) = x^0, \tag{10.5}$$

avec $x(t) \in I\!R^n$, $u(t) \in I\!R^m$, et $f : I\!R \times I\!R^n \times I\!R^m \to I\!R^n$. On supposera f lipschitzienne et dérivable, de dérivée lipschitzienne. Ici encore, on définit une solution de (2.5) comme une fonction $x(t)$ continue en temps, à valeurs dans $I\!R^n$, telle que

$$x(t) = x^0 + \int_0^t f(s, x(s), u(s))\mathrm{d}s, \quad \text{pour tout } \ t > 0, \tag{10.6}$$

et on peut vérifier l'existence et l'unicité de la solution. Comme dans le chapitre précédent, on prendra en compte une contrainte sur la commande du type

$$u(t) \in U, \quad t \geq 0, \tag{10.7}$$

où U est un ensemble *convexe, compact* et tel que $0 \in \operatorname{int} U$. Le problème de transfert en temps minimal de l'état initial x^0 à un point de la cible C, supposée convexe fermée et non vide, s'écrit

$$\inf_{(x,u,T)} T; \quad x(T) \in C; \quad (x,u) \quad \text{satisfont (2.6)-(2.7).} \tag{10.8}$$

10.1.3 Existence de solutions

Malheureusement, sous les hypothèses précédentes, il peut ne pas exister de solution au problème, comme le montre l'exemple suivant.

Exemple 10.1 Soit le système dynamique

$$\dot{x} = \sin 2\pi u, \quad \dot{y} = \cos 2\pi u, \quad \dot{z} = \pi^2(x^2 + y^2) - 1, \tag{10.9}$$

avec état initial $(0, 0, 1)$. On considère le problème du transfert en temps minimal à la cible $z = 0$, sous la contrainte $u \in [0, 1]$ (qui se ramène au cas $0 \in \operatorname{int} U$ par translation). L'expression de la dérivée de z implique que le temps minimal de transfert ne peut être inférieur à 1, et que le transfert en temps $T = 1$ est impossible.

Soit k un entier positif. A la commande $u(t) = kt$ (modulo 1) est associé l'état

$$x(t) = \frac{1 - \cos 2\pi kt}{2\pi k}; \quad y(t) = \frac{\sin 2\pi kt}{2\pi k}; \tag{10.10}$$

et

$$z(t) = 1 + \int_0^t \left[\frac{1 - \cos 2\pi kt}{2k^2} - 1\right] \mathrm{d}t = 1 - t + \frac{t}{2k^2} - \frac{\sin 2\pi kt}{4\pi k^3}. \tag{10.11}$$

Cette expression permet de vérifier que $z(t_k) = 0$ pour un temps t_k tendant vers 1 quand $k \to \infty$. L'infimum des temps de transfert est donc 1 et n'est jamais atteint.

Nous allons néanmoins donner un résultat d'existence pour la classe, importante dans les applications, des problèmes pour lesquels la *dynamique est affine par rapport à la commande*. Soient $g_1, \ldots, g_q$ des *champs de vecteurs*[1]. Supposant pour simplifier l'exposé que la dynamique est *autonome* (indépendante du temps), on se place donc dans le cas où f est de la forme

$$f(x, u) = g_0(x) + \sum_{i=1}^{q} u_i g_i(x). \tag{10.12}$$

Théorème 10.2 *On suppose la dynamique affine en la commande, les champs de vecteurs étant lipschitziens et bornés. Si le problème* (2.8) *est réalisable, il a au moins une solution.*

Démonstration. Soient u_k une suite minimisante et x_k les états associés. Le lemme 1.3 permet (extrayant une sous suite si nécessaire) d'affirmer que u_k a une limite faible $\bar{u}$, telle que $\bar{u}(t) \in U$ p.p. D'après les hypothèses sur g, la suite x_k est uniformément lipschitzienne, au sens où il existe $L > 0$ telle que

$$\|x_k(t') - x_k(t)\| \leq L|t' - t|, \quad \text{pour tout } t,\, t' \in [0, T]. \tag{10.13}$$

[1] Un champ de vecteurs est une application de $\mathbb{R}^n$ dans lui même.

Extrayant une sous-suite si nécessaire, on en déduit[2] que x_k converge uniformément sur $[0,T]$ vers une fonction $\bar{x}$, lipschitzienne de constante L. De plus, $f(x_k(t),u_k(t)) - f(\bar{x}(t),u_k(t))$ converge uniformément vers 0. En conséquence, pour tout $t \in [0,T(x^0)]$,

$$\begin{aligned}\bar{x}(t) - x_0 &= \lim_k x_k(t) - x_0 = \lim_k \int_0^t f(x_k(s),u_k(s))\mathrm{d}s = \lim_k \int_0^t f(\bar{x}(s),u_k(s))\mathrm{d}s \\ &= \int_0^t g_0(\bar{x}(s))\mathrm{d}s + \sum_{i=1}^n \int_0^t (u_k)_i(s) g_i(\bar{x}(s))\mathrm{d}s = \int_0^t f(\bar{x}(s),\bar{u}(s))\mathrm{d}s\end{aligned}$$

où la dernière égalité est conséquence de la convergence faible. De plus $\bar{x}(T(x^0)) = \lim_k x_k(T_k)$ appartient à C puisque C est fermé. Donc $\bar{u}$ réalise le transfert en temps minimal. ■

10.2 Conditions d'optimalité

10.2.1 Un résultat général

Introduisons le *pseudo-hamiltonien* $H : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ défini par

$$H(t,x,u,p) := p \cdot f(t,x,u). \tag{10.14}$$

Dans le cas autonome on notera $f(x,u)$ la dynamique et $H(x,u,p)$ le pseudo-hamiltonien. On dit que la commande $u \in L^\infty(0,T,U)$ satisfait le *Principe du minimum pour le problème* (2.8) si elle satisfait les relations suivantes :

$$\begin{cases} \dot{x}(t) &= f(t,x(t),u(t)), \quad t \geq 0, \\ x(0) &= x_0, \end{cases} \tag{10.15}$$

$$\begin{cases} -\dot{p}(t) &= H_x(t,x(t),u(t),p(t)), \quad t \in [0,T], \\ p(T) &= q, \end{cases} \tag{10.16}$$

$$H(t,x(t),u(t),p(t)) = \inf_{v \in U} H(t,x(t),v,p(t)),\ u(t) \in U, \quad \text{p.p. } t \in [0,T], \tag{10.17}$$

$$q \cdot y \leq q \cdot x(T), \quad \text{pour tout } \ y \in C; \quad x(T) \in C; \quad q \neq 0. \tag{10.18}$$

Théorème 10.3 *Toute solution du problème* (2.8) *satisfait le principe du minimum.*

Démonstration. La démonstration étant technique, nous la reportons en section 2.4 pour discuter sans attendre les conséquences de ce résultat. ■

Remarque 10.4 Si la commande u satisfait le principe du minimum pour le problème (2.8), l'application $t \to H(t,x(t),u(t),p(t))$ est essentiellement constante. On le vérifie facilement en étendant la démonstration de la proposition 1.17.

[2]Par exemple en appliquant le théorème d'Ascoli-Arzela, concernant les familles équicontinues de fonctions.

Exemple 10.5 Dans le cas de systèmes dynamiques autonomes affines en la commande, donc de dynamique donnée par (2.12), une commande u satisfaisant le principe du minimum vérifie presque partout en $t \in [0,T]$

$$\sum_{i=1}^{q} u_i(t)p(t)\cdot g_i(x(t)) \leq \sum_{i=1}^{q} v_i p(t)\cdot g_i(x(t)), \quad \text{pour tout } \ v \in U. \tag{10.19}$$

En particulier, on a p.p. $u(t) \in \partial U$ quand $p(t)\cdot g_i(x(t)) \neq 0$ pour au moins un i.

Remarque 10.6 Si la dynamique est linéaire et autonome on retrouve les résultats du chapitre précédent. Il en résulte que les conditions du théorème 2.3 ne sont pas des conditions suffisantes d'optimalité (remarque 1.25).

10.2.2 Arc singulier

Nous étudions dans cette section des systèmes dynamiques autonomes affines en la commande, dans le cas d'une seule commande :

$$\dot{x} = g_0(x(t)) + u(t)g_1(x(t)), \tag{10.20}$$

les champs de vecteurs g_0 et g_1 étant de classe C^∞, et en supposant $U = [-1,1]$. Le hamiltonien est fonction affine de la commande, de pente $p(t)\cdot g_1(x(t))$. Une commande satisfaisant le principe du minimum vérifie donc

$$\bar{u}(t) = \begin{cases} -1 & \text{si} \quad p(t)\cdot g_1(x(t)) > 0, \\ 1 & \text{si} \quad p(t)\cdot g_1(x(t)) < 0. \end{cases} \tag{10.21}$$

Nous avons vu dans l'exemple 2.1.1 que la commande en temps minimal peut se trouver hors des bornes sur un intervalle de temps $]\tau_1,\tau_2[$. Dans ce cas, l'application $v \to H(x(t),v,p(t))$ est constante, et donc

$$p(t)\cdot g_1(x(t)) = 0, \tag{10.22}$$

de sorte que le principe du minimum ne semble donner aucune information sur la commande optimale. On appelle *arc singulier* la courbe $(x(t),u(t),p(t))$ sur $]\tau_1,\tau_2[$.

Dans la plupart des applications, on peut obtenir une expression de la commande en fonction de l'état et de l'état adjoint en dérivant autant de fois que nécessaire l'application $t \to p(t)\cdot g_1(x(t))$.

Le calcul sera grandement simplifié par l'utilisation des *crochets de Lie*. Rappelons que cette opération associe à une paire de champs de vecteurs (X,Y) différentiables un nouveau champ de vecteur

$$[X,Y] := X'Y - Y'X. \tag{10.23}$$

Autrement dit, $[X,Y]$ a pour composantes i, avec $1 \leq i \leq n$, la quantité

$$[X,Y]_i(x) = \sum_{j=1}^{n} X'_{ij}(x)Y_j(x) - Y'_{ij}(x)X_j(x). \tag{10.24}$$

On notera, pour $k \geq 1$,

$$adX.Y := ad^1X.Y := [X,Y]; \quad ad^{k+1}X.Y := [X, ad^kX.Y]. \tag{10.25}$$

On notera aussi $g_i(t) := g_i(x(t))$, $[g_0,g_1](t) := [g_0(x(t)), g_1(x(t))]$.

Théorème 10.7 *Soit une commande u vérifiant le principe du minimum. Alors, p.p. $t \in [0,T]$ on a*

$$\frac{\mathrm{d}}{\mathrm{d}t}H'_u(x(t),u(t),p(t)) = -p(t)\cdot[g_0,g_1](t), \tag{10.26}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}H'_u(x(t),u(t),p(t)) = p(t)\cdot(ad^2g_0.g_1(t) - u(t)\,ad^2g_1.g_0(t)). \tag{10.27}$$

De plus, soit un instant t faisant partie d'un arc singulier, tel que

$$p(t)\cdot ad^2g_1.g_0(t) \neq 0.$$

Alors la commande est donné en fonction de l'état et de l'état adjoint par la formule

$$u(t) = \frac{p(t)\cdot ad^2g_0.g_1(t)}{p(t)\cdot ad^2g_1.g_0(t)}. \tag{10.28}$$

Démonstration. L'équation de l'état adjoint s'écrit ici

$$-\dot p(t) = (g_0'(x(t)) + u(t)g_1'(x(t)))^\top p(t). \tag{10.29}$$

Notons $\Delta^1 := \dfrac{\mathrm{d}}{\mathrm{d}t}H'_u(y(t),u(t),p(t))$ et $\Delta^2 = \dfrac{\mathrm{d}}{\mathrm{d}t}\Delta^1$. Pour simplifier les calculs, on omettra l'argument $x(t)$ des champs de vecteurs, et le temps en argument. Il vient

$$\begin{aligned}
\Delta^1 &= \frac{\mathrm{d}}{\mathrm{d}t}\langle p, g_1\rangle = \langle \dot p, g_1\rangle + \langle p, g_1'\dot x\rangle \\
&= -\langle (g_0' + ug_1')^\top p, g_1\rangle + \langle p, g_1'(g_0+ug_1)\rangle \\
&= -\langle p, g_0'g_1 + ug_1'g_1\rangle + \langle p, g_1'g_0 + ug_1'g_1\rangle \\
&= -\langle p, g_0'g_1 - g_1'g_0\rangle = -\langle p, [g_0,g_1]\rangle,
\end{aligned}$$

d'où (2.26), et

$$\begin{aligned}
\Delta^2 &= -\frac{\mathrm{d}}{\mathrm{d}t}\langle p, [g_0,g_1]\rangle = -\langle \dot p, [g_0,g_1]\rangle - \langle p, \frac{\mathrm{d}}{\mathrm{d}t}[g_0,g_1]\rangle \\
&= \langle p, (g_0' + u\,g_1')[g_0,g_1]\rangle - \langle p, [g_0,g_1]'(g_0+ug_1)\rangle \\
&= \langle p, g_0'[g_0,g_1] - [g_0,g_1]'g_0 + u\,[g_1'[g_0,g_1] - [g_0,g_1]'g_1]\rangle \\
&= \langle p, [g_0,[g_0,g_1]] + u\,[g_1,[g_0,g_1]]\rangle.
\end{aligned}$$

Mais $[g_0, g_1] = -[g_1, g_0]$, donc $[g_1, [g_0, g_1]] = -[g_1, [g_1, g_0]]$ d'où (2.27). Enfin si t fait partie d'un arc singulier, les deux membres de (2.27) sont nuls, d'où (2.28). ■

Notons

$$\mathcal{T} := \{t \in [0, T]; \quad H'_u(x(t), u(t), p(t)) = p(t) \cdot g_1(x(t)) = 0\}. \tag{10.30}$$

Les formules (2.26)-(2.27) permettent, dans certains cas, d'assurer que $\mathcal{T}$ est de cardinal fini (ce qui exclut la présence d'arcs singuliers).

Lemme 10.8 *Soit $t_0 \in \mathcal{T}$. Si $p(t_0) \cdot [g_0, g_1](t_0) \neq 0$, alors t est un point isolé de $\mathcal{T}$. Si cette propriété est satisfaite pour tout $t \in \mathcal{T}$, alors $\mathcal{T}$ est de cardinal fini.*

Démonstration. Comme $H'_u(x(t), u(t), p(t))$ a en t_0 une dérivée non nulle, t_0 est un zéro isolé de $t \mapsto H'_u(x(t), u(t), p(t))$. Si cela est vrai pour tout $t \in \mathcal{T}$, puisque $\mathcal{T}$ est fermé, il ne peut avoir de points d'accumulation donc est de cardinal fini. ■

Proposition 10.9 *Supposons la dimension de l'espace d'état égale à 2 et, pour tout $x \in I\!R^2$, les champs g_1 et $[g_0, g_1]$ linéairement indépendants. Alors une trajectoire extrémale ne peut avoir d'arc singulier, et une commande en temps minimal change de signe un nombre fini de fois.*

Démonstration. Soit $t \in \mathcal{T}$, donc $p(t) \cdot g_1(t) = 0$. Nécessairement $p(t) \neq 0$, sinon (par intégration de l'équation de l'état adjoint) p serait nul sur $[0, T]$, or on sait que $0 \neq q = p(T)$. Comme $n = 2$, l'indépendance linéaire de g_1 et $[g_0, g_1]$ implique $p(t) \cdot [g_0, g_1](t) \neq 0$. On conclut avec le lemme 2.8. ■

Remarque 10.10 En d'autres termes, si $n = 2$, un arc singulier est contenu dans le *lieu singulier* de l'espace d'état défini par l'équation (on note '$\wedge$' le produit vectoriel)

$$G(x) := g_1(x) \wedge [g_0, g_1](x) = 0. \tag{10.31}$$

Notons que, sur un arc singulier, dérivant la relation précédente, il vient

$$G'(x(t))(g_0(x(t)) + u(t)g_1(x(t))) = 0. \tag{10.32}$$

Si $G'(x(t))g_1(x(t)) \neq 0$, on en tire une expression de la commande *en fonction de l'état*.

Proposition 10.11 *Supposons $n = 3$ et, notant encore $G(x) := g_1(x) \wedge [g_0, g_1](x)$, l'indépendance linéaire de $g_1(x)$ et $[g_0, g_1](x)$ pour tout x. Si t appartient à un arc singulier, et*

$$G(x(t)) \cdot ad^2 g_1.g_0(t) \neq 0, \tag{10.33}$$

on peut exprimer la commande à l'instant t en fonction de l'état :

$$u(t) = \frac{G(x(t)) \cdot ad^2 g_0.g_1(t)}{G(x(t)) \cdot ad^2 g_1.g_0(t)}. \tag{10.34}$$

Démonstration. L'appartenance de t à un arc singulier implique les relations (2.22) et(2.26), et donc $p(t)$ est colinéaire à $G(x(t))$. On conclut avec la relation (2.28) du théorème 2.7. ■

Remarque 10.12 On trouvera d'autres aspects de la théorie des arcs singuliers dans Bryson et Ho [20], en particulier des conditions d'optimalité d'ordre élevé et des conditions dites de jonction, qui concernent les bords de l'arc singulier.

10.3 Applications

10.3.1 Pendule

Considérons le problème de commande du pendule

$$\ddot{\theta} + g \sin\theta = u, \tag{10.35}$$

avec $g > 0$ pesanteur, $\theta \in I\!R$ angle du pendule, et la contrainte $u \in U = [-1, 1]$. Introduisant la vitesse angulaire ω, on obtient la forme suivante :

$$\dot{\theta} = \omega; \quad \dot{\omega} = u - g \sin\theta, \tag{10.36}$$

d'où l'expression du pseudo-hamiltonien

$$H(\theta, \omega, u, p_\theta, p_\omega) = p_\theta \omega + p_\omega (u - g \sin\theta), \tag{10.37}$$

et de la dynamique de l'état adjoint

$$-\dot{p}_\theta = -g p_\omega \cos\theta, \quad -\dot{p}_\omega = p_\theta. \tag{10.38}$$

Sur un arc singulier, $p_\omega = 0$, donc p_θ aussi, ce qui est impossible. Il n'existe donc pas d'arc singulier.

La commande optimale est

$$u(t) = \begin{cases} -1 & \text{si} \quad p_\omega(t) > 0, \\ 1 & \text{si} \quad p_\omega(t) < 0. \end{cases} \tag{10.39}$$

Le long d'une trajectoire en temps minimal, on a donc

$$\begin{cases} \ddot{\theta} + g \sin\theta &= -\dfrac{p_\omega}{|p_\omega|}, \\ \ddot{p}_\omega + g p_\omega \cos\theta &= 0. \end{cases} \tag{10.40}$$

Si le temps minimal est assez petit, on obtient des trajectoires en temps minimal en intégrant l'équation d'état en temps rétrograde à partir de la cible, avec par exemple $u = 1$ sur un intervalle $[0, \tau]$, puis avec $u = -1$. Le tracé correspondant se trouve en figure 2.1.

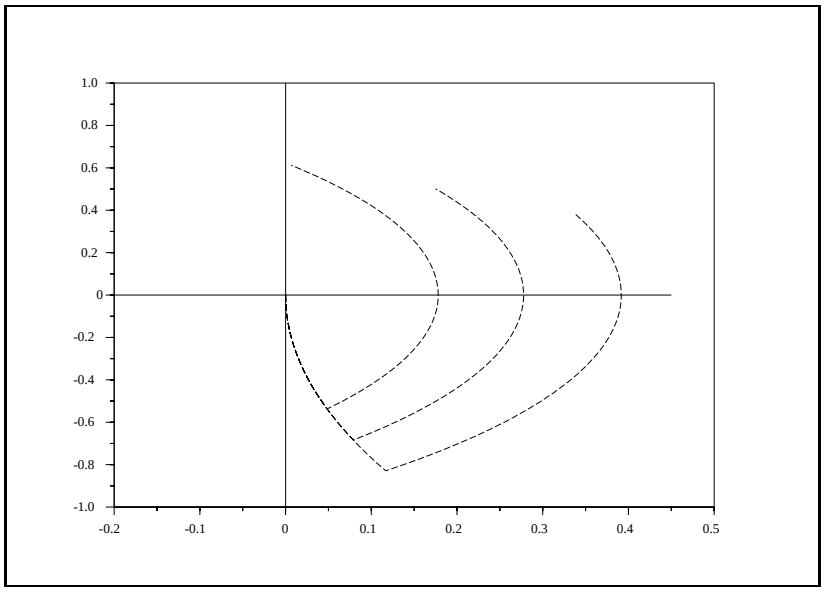

FIG. 10.1 – Commande du pendule : quelques trajectoires en temps minimal

Remarque 10.13 On trouvera dans Lee et Markus [42, Chapitre 7] une étude du problème (assez complexe !) de détermination de la commande optimale en des points éloignés de la cible.

D'un point de vue pratique, une heuristique consiste à prendre d'abord une commande réduisant le plus possible l'énergie mécanique, soit $u = -\omega/|\omega|$ puis, quand on est assez près de la cible, prendre la commande optimale (calculée ci-dessus). Enfin au voisinage immédiat de la cible on préferera un bouclage linéaire, pour éviter les oscillations rapides entre ± 1 qu'engendreraient inévitablement les bruits et vibrations diverses.

On voit sur cet exemple l'intérêt pratique de combiner différentes approches.

10.3.2 Avion à trajectoire horizontale

Nous reprenons le problème décrit dans la section 2.1 : transfert en temps minimal, vers une position donnée, d'un avion dont la trajectoire est horizontale et rectiligne. Nous allons calculer, sur un arc singulier, l'expression de la commande optimale, en fonction de l'état.

La dynamique est linéaire par rapport à la commande, et la théorie de l'arc singulier développée en section 2.2.2 s'applique donc. Cependant, plutôt que de calculer les crochets de Lie correspondants, *il est beaucoup plus simple d'effectuer des dérivations*

directes[3].

On note (p_y, p_v, p_m) les coordonnées de l'état adjoint. Il vient avec (2.3), omettant les arguments quand on le peut :

$$H(y, v, m, u, p) = vp_y + p_v \frac{cu - D(v, m)}{m} - up_m, \tag{10.41}$$

et donc l'équation de l'état adjoint est

$$-\dot{p}_y = 0, \quad -\dot{p}_v = p_y - p_v \frac{D'_v}{m}, \quad -\dot{p}_m = p_v \frac{D - cu - mD'_m}{m^2}. \tag{10.42}$$

Notons que p_y est constant, donc $p_y(t) = q_y$. On a aussi

$$H'_u = c\frac{p_v}{m} - p_m. \tag{10.43}$$

Posons

$$\Delta(v, m) := \frac{mD'_m(v, m) - D(v, m) - cD'_v(v, m)}{c}. \tag{10.44}$$

Lemme 10.14 *Sur un arc singulier, la commande est solution de*

$$(\Delta + c\Delta'_v - m\Delta'_m)\, u = \Delta\,(\Delta - D'_v) + D\Delta'_v. \tag{10.45}$$

Démonstration. Sur l'arc singulier, on a avec (2.43), en omettant le temps en argument :

$$H'_u = c\frac{p_v}{m} - p_m = 0. \tag{10.46}$$

Dérivant en temps cette relation, il vient

$$0 = cu\frac{p_v}{m^2} - c\frac{p_y}{m} + c\frac{p_v D'_v}{m^2} + p_v \frac{D - cu - mD'_m}{m^2} = c\frac{mp_y - p_v\Delta}{m^2}. \tag{10.47}$$

Cette relation, qui conformément à la théorie ne dépend pas de u, équivaut à

$$mp_y - p_v\Delta = 0. \tag{10.48}$$

Dérivant cette relation par rapport au temps, il vient

$$-up_y + \Delta\left(p_y - p_v\frac{D'_v}{m}\right) - p_v\left(\Delta'_v\frac{cu - D}{m} - u\Delta'_m\right) = 0, \tag{10.49}$$

soit

$$\left(p_y + p_v\left(\Delta'_v\frac{c}{m} - \Delta'_m\right)\right)u = \Delta\left(p_y - p_v\frac{D'_v}{m}\right) + p_v\Delta'_v\frac{D}{m}. \tag{10.50}$$

[3]L'important est de comprendre le principe des calculs qui suivent, plus que le détail qui est quelque peu pénible. Dans la pratique on réalise les calculs avec des outils de calcul formel.

Les relations (2.46) et (2.48) sont linéairement indépendantes (par rapport à p). Le long d'un arc singulier, p est donc proportionnel à la base du noyau des relations (2.46)-(2.48) d'expression

$$\mathcal{G}(x) = (\Delta, \quad m, \quad c)^\top . \tag{10.51}$$

Combinant avec (2.50), on obtient (2.45). ■

Remarque 10.15 Pour fournir une approximation numérique de la solution, on peut procéder comme suit. L'intuition physique suggère une première phase à débit maximum (si la vitesse initiale est faible) ou nul (si elle est élevée), suivie d'un arc singulier se terminant quand le réservoir est vide. Il suffit donc (dans chacun des deux cas) d'essayer différentes valeurs de l'instant d'entrée dans l'arc singulier. Dans les calculs on prendra garde aux bornes que doit respecter le débit de gaz dans l'arc singulier.

Remarque 10.16 On trouvera une analyse détaillée d'un problème similaire, mais un peu plus simple (on maximise la portée au lieu du temps de transfert) avec le tracé du lieu des trajectoires, dans Leitmann [43, Section 2.9].

10.4 Démonstration du résultat principal

Cette section est consacrée à la démonstration du théorème 2.3, dont la clé réside dans l'estimation de l'écart entre deux états associés à des commandes voisines, grâce à une linéarisation non standard de l'équation d'état. On introduit la *distance d'Ekeland* sur l'espace $L^\infty(0,T,U)$:

$$\delta(u,v) := \mathrm{mes}(\{v(t) \neq u(t)\}). \tag{10.52}$$

Soient u, u_1 et u_2 dans $L^\infty(0,T,U)$, x, x_1 et x_2 leurs état associé. Posons $w := x_2 - x_1$. On note z la solution de la *linéarisation non standard* de l'équation d'état :

$$\left\{ \begin{array}{lcl} \dot{z}(t) & = & f_x(t,x(t),u(t))z(t) + f(t,x(t),u_2(t)) - f(t,x(t),u_1(t)), \\ & & \text{p.p. } t \in [0,T], \\ z(0) & = & 0. \end{array} \right. \tag{10.53}$$

Lemme 10.17 *Soient $(u,u_1,u_2,x,x_1,x_2,w,z)$ comme ci-dessus. Si $\delta(u_i,u) \to 0$, $i = 1,2$, alors*

$$\text{(i)} \ \ \|w\|_{L^\infty(0,T,\mathbb{R}^n)} = O(\delta(u_2,u_1)), \quad \text{(ii)} \ \ \|w-z\|_{L^\infty(0,T,\mathbb{R}^n)} = o(\delta(u_2,u_1)). \tag{10.54}$$

Démonstration. L'application f est lipschitzienne, donc, p.p. $t \in [0,T]$:

$$\begin{array}{rcl} \|\dot{w}(t)\| & \leq & \|f(t,x_2(t),u_2(t)) - f(t,x_2(t),u_1(t))\| \\ & & + \|f(t,x_2(t),u_1(t)) - f(t,x_1(t),u_1(t))\| \\ & \leq & O(\|u_2(t)-u_1(t)\|) + O(\|w(t)\|). \end{array}$$

Comme U est compact, on a $\|u_2 - u_1\|_{L^1(0,T,U)} = O(\delta(u_2, u_1))$ et l'inégalité de Gronwall implique (2.54)(i). Par ailleurs, on peut écrire

$$\dot{w}(t) = f(t, x(t), u_2(t)) - f(t, x(t), u_1(t)) + A_2(t) - A_1(t), \tag{10.55}$$

où pour $i = 1, 2$, notant $\bar{x}_i := x_i - x$:

$$A_i(t) \quad = \quad f(t, x_i(t), u_i(t)) - f(t, x(t), u_i(t)) = \int_0^1 f_x(t, x(t) + \theta \bar{x}_i(t), u_i(t)) \bar{x}_i(t) \mathrm{d}\theta,$$

et donc

$$A_2(t) - A_1(t) = \int_0^1 f_x(t, x(t) + \theta \bar{x}_2(t), u_2(t)) w(t) \mathrm{d}\theta + \int_0^1 \left[f_x(t, x(t) + \theta \bar{x}_2(t), u_2(t)) - f_x(t, x(t) + \theta \bar{x}_1(t), u_1(t)) \right] \mathrm{d}\theta \bar{x}_1(t).$$

Soient $A_3(t)$ et $A_4(t)$ les membres de droite de chaque ligne. La convergence uniforme de x_1 et x_2 vers x et l'estimation $\|w\|_{L^\infty(0,T,\mathbb{R}^n)} = O(\delta(u_2, u_1))$ impliquent :

$$A_3(t) \quad = \quad f_x(t, x(t), u(t)) w(t) + o(\delta(u_2, u_1)) + o(\|u_2(t) - u_1(t)\|), \tag{10.56}$$

$$A_4(t) \quad = \quad o(\delta(u_2, u_1)) + o(\|u_2(t) - u_1(t)\|). \tag{10.57}$$

Au total, posant $y := z - w$, il vient

$$\dot{y}(t) = f_x(t, x(t), u(t)) y(t) + o(\delta(u_2, u_1)) + o(\|u_2(t) - u_1(t)\|), \tag{10.58}$$

d'où (2.54)(ii) avec l'inégalité de Gronwall. ■

Définition 10.18 Soient $u \in L^\infty(0, T, U)$ et x l'état associé. On dit que $y \in \mathbb{R}^n$ est une *variation finale* associée à u, s'il existe une suite de commandes $u_k \in L^\infty(0, T, U)$, et une suite numérique $\varepsilon_k \downarrow 0$ telles que, notant x_k l'état associé à u_k, on a $(x_k(T) - x(T))/\varepsilon_k \to y$. On note $\mathcal{C}_T(u)$ l'ensemble des variations finales.

Il est clair que $\mathcal{C}_T(u)$ est un cône fermé. Construisons un type particulier de variation admissible.

Définition 10.19 (i) La *perturbation en aiguille* associée à $t_0 \in]0, T[$ et $w \in U$, indicée par $\gamma > 0$, est la famille de commandes admissibles v_γ, d'état associé x_γ, définie par

$$v_\gamma(t) = w \ \text{ si } \ |t - t_0| \leq \gamma, \ \ u(t) \ \text{ sinon.} \tag{10.59}$$

(ii) Soit $z \in L^1(0, T, \mathbb{R}^n)$. On dit que $t_0 \in]0, T[$ est un *point de Lebesgue* de z si

$$z(t_0) = \lim_{\gamma \downarrow 0} \frac{1}{2\gamma} \int_{t_0 - \gamma}^{t_0 + \gamma} z(t) \mathrm{d}t. \tag{10.60}$$

On sait que (2.60) est satisfaite presque partout, voir par exemple Rudin [51, théorème 7.7]. En particulier, presque tout $t_0 \in]0,T[$ est un point de Lebesgue de $t \mapsto f(t,x(t),u(t))$.

Lemme 10.20 *Soient $u \in L^\infty(0,T,U)$, x l'état associé, et $t_0 \in]0,T[$ un point de Lebesgue de $f(t,x(t),u(t))$. Alors la perturbation en aiguille associée à $t_0 \in]0,T[$ et $w \in U$ est telle que la variation finale $y = \lim(x_\gamma(T) - x(T))/(2\gamma)$ existe. On l'appelle* variation en aiguille *associée à $t_0 \in]0,T[$ et $w \in U$. Si de plus p est solution de l'équation adjointe* (2.16) *(avec une condition terminale q quelconque), alors*

$$q \cdot y = H(t,x(t_0),w,p(t_0)) - H(t,x(t_0),u(t_0),p(t_0)). \tag{10.61}$$

Démonstration. On applique le lemme 2.17, avec $u_2 = v_\gamma$ et $u_1 = u$. Puisque $\delta(v_\gamma,u) \le 2\gamma$, on a $x_\gamma(T) - x(T) = z_\gamma(T) + o(\gamma)$, où z_γ est solution de

$$\left\{ \begin{array}{rcl} \dot z_\gamma(t) & = & f_x(t,x(t),u(t))z_\gamma(t) + f(t,x(t),v_\gamma(t)) - f(t,x(t),u(t)), \\ & & \text{p.p. } t \in [0,T], \\ z(0) & = & 0, \end{array} \right. \tag{10.62}$$

et donc pour $q \in I\!R^n$ quelconque et p solution de l'équation adjointe (2.16),

$$\begin{array}{rcl} q \cdot z_\gamma(T) & = & p(T) \cdot z_\gamma(T) = \displaystyle\int_0^T [\dot p(t) \cdot z_\gamma(t) + p(t) \cdot \dot z_\gamma(t)]\,\mathrm{d}t \\ & = & \displaystyle\int_0^T p(t)\,(f(t,x(t),v_\gamma(t)) - f(t,x(t),u(t)))\,\mathrm{d}t. \end{array} \tag{10.63}$$

Revenant à la définition de v_γ et utilisant le fait que $t_0 \in]0,T[$ est point de Lebesgue de $f(t,x(t),u(t))$, on obtient (2.61) par passage à la limite, d'où la conclusion. ■

On note $\hat{\mathcal{C}}_T(u)$ le cône convexe engendré par les combinaisons linéaires positives de variations finales en aiguille. Autrement dit,

$$\hat{\mathcal{C}}_T(u) := \left\{ \sum_{i\in I} a_i z_i;\ I \text{ fini}, a_i \ge 0,\ z_i \in C_T(u),\ i \in I \right\}. \tag{10.64}$$

Lemme 10.21 *Les conditions suivantes sont équivalentes :*
(i) *La commande u satisfait le principe du minimum, l'égalité dans* (2.17) *étant satisfaite pour tout point de Lebesgue de $t \mapsto f(t,x(t),u(t))$,*
(ii) *Il existe $q \neq 0$, normale extérieure à C en $x(T)$, telle que $q \cdot y \ge 0$, pour toute variation finale en aiguille y,*
(iii) $0 \notin \mathrm{int}\left(x(T) + \hat{\mathcal{C}}_T(u) - C\right)$.

Démonstration. Le principe du minimum fournit une normale extérieure $q \neq 0$ à C en $x(T)$; si y est une variation finale en aiguille, alors $q \cdot y \ge 0$ d'après le lemme 2.20 combiné à (2.17), donc (i) implique (ii). Si (ii) est satisfait, soit p solution de

(2.16) (cette équation différentielle linéaire rétrograde a une solution unique). Alors le lemme 2.20 implique le principe du minimum, l'égalité dans (2.17) étant satisfaite pour tout point de Lebesgue de $t \mapsto f(t, x(t), u(t))$; donc (ii) implique (i).

L'équivalence de (ii) et (iii) résulte du lemme 1.13, en notant que (ii) équivaut à la séparation de $\{0\}$ et de l'ensemble convexe $\Big(x(T) + \hat{\mathcal{C}}_T(u) - C\Big)$. ■

Dans la suite on va prouver la nécessité du principe du minimum en montrant que, si u réalise le transfert en temps minimal, la condition (iii) du lemme 2.21 est satisfaite. Les démonstrations par l'asurde s'appuieront sur la condition opposée

$$0 \in \text{int}\left(x(T) + \hat{\mathcal{C}}_T(u) - C\right), \tag{10.65}$$

et sur l'étude de l'ensemble $\hat{\mathcal{C}}_T(u)$.

Lemme 10.22 *Soit $u \in L^\infty(0, T, U)$. Alors $\hat{\mathcal{C}}_T(u) \subset \mathcal{C}_T(u)$.*

Démonstration. Soient $y = \sum_{i=1}^k a_i y_i$, avec $a_i > 0$ pour tout i, et y_i variation finale associée à la perturbation en aiguille associée à $t_i \in]0, T[$ et $w_i \in U$. Supposons d'abord les instants t_i distincts. On construit alors la perturbation de la commande de la manière suivante

$$v_\gamma(t) = w_i \ \text{ si } \ |t - t_i| \leq a_i\gamma, \ \ i = 1, \ldots, k; \ \ u(t) \ \text{ sinon.} \tag{10.66}$$

On conclut facilement avec le lemme 2.17, par des calculs similaires à ceux de la démonstration du lemme 2.20.

Donnons maintenant l'idée de la preuve du cas général en traitant le cas de deux points égaux $t_1 = t_2$, avec $k = 2$. On pose dans ce cas

$$v_\gamma(t) = \begin{cases} w_1 \ \text{ si } \ t \in [t_1 - 2a_1\gamma, t_1], \\ w_2 \ \text{ si } \ t \in]t_1, t_1 + 2a_2\gamma], \\ u(t) \ \text{ sinon.} \end{cases} \tag{10.67}$$

On conclut une fois de plus avec le lemme 2.17, par des calculs similaires à ceux de la démonstration du lemme 2.20. ■

On parlera encore de perturbation en aiguille associée à la variation en aiguille $y \in \hat{\mathcal{C}}_T(u)$. Ces variation finales en aiguille notées encore v_γ, dont la démonstration donne le principe de construction, sont telles que $\delta(v_\gamma, u) = O(\gamma)$, et leur état associé x_γ vérifie $x_\gamma(T) = x(T) + 2\gamma y + o(\gamma)$.

Démontrons d'abord le théorème 2.3 dans le cas où C est d'intérieur non vide.

Lemme 10.23 *Soit u solution du problème* (2.8)*. Si C est d'intérieur non vide, alors $0 \notin \text{int}\Big(x(T) + \hat{\mathcal{C}}_T(u) - C\Big)$, ce qui assure la conclusion du théorème 2.3 en raison du lemme 2.21.*

Démonstration. a) Posons $E := x(T)+\hat{\mathcal{C}}_T(u)-\operatorname{int} C$, et montrons que $0 \notin E$. Si $0 \in E$, il existe $y_0 \in \hat{\mathcal{C}}_T(u) \cap (\operatorname{int} C - x(T))$. La perturbation en aiguille v_γ correspondante est telle que son état associée x_γ vérifie $x_\gamma(T) = x(T)+2\gamma y_0+o(\gamma)$, donc $x_\gamma(T) \in \operatorname{int} C$ si $\gamma > 0$ est assez petit. Pour un tel $\gamma > 0$, quand $t < T$ est proche de T, on a encore $x_\gamma(t) \in \operatorname{int} C$, ce qui contredit l'optimalité de u.
b) Il est clair que E est ouvert, donc (puisqu'il est convexe) $E = \operatorname{int} \bar{E}$, et que $0 \in \bar{E} \backslash E$. Du lemme 1.13 on déduit l'existence d'une forme linéaire $q \neq 0$ séparant 0 de $\bar{E}$, donc de E. Autrement dit, $q \cdot (w - x(T)) \leq q \cdot y$, pour tout $w \in \operatorname{int} C$ et $y \in \hat{\mathcal{C}}_T(u)$. Cette inégalité reste vraie quand $w \in C$; donc q sépare 0 et $(x(T) + \hat{\mathcal{C}}_T(u) - C)$. Du lemme 1.13 découle $0 \notin \operatorname{int}(x(T) + \hat{\mathcal{C}}_T(u) - C)$, d'où la conclusion. ■

Etudions maintenant le cas où la cible est réduite à un point. Etant donné $E \subset \mathbb{R}^n$, on note $\operatorname{conv} E$ l'ensemble des combinaisons linéaires convexes (à coefficients positifs de somme un) d'éléments de E ; c'est le plus petit ensemble convexe contenant E.

Lemme 10.24 *Soit u solution du problème* (2.8)*. Si C est réduit à un point, alors $0 \notin \operatorname{int} \hat{\mathcal{C}}_T(u)$, ce qui assure la conclusion du théorème 2.3 en raison du lemme 2.21.*

Démonstration. Notons a_0, $a_1, \ldots,$ diverses constantes positives. On peut supposer que $C = \{0\}$. Si la conclusion n'est pas satisfaite, de (2.65) on déduit qu'il existe r variations finales en aiguille y^1 à y^r telles que

$$2\varepsilon B \subset \operatorname{conv}\left\{y^1, \ldots, y^r\right\}. \tag{10.68}$$

Pour $\gamma > 0$, et $h \in \mathbb{R}^r_+$, on note $u_{\gamma,h}$ la perturbation en aiguille associée à la variation finale $y_h := \sum_{i=1}^r h_i y_i$, bien définie pour $\gamma < a_0$ et $\|h\| \leq a_1$, et $x_{\gamma,h}$ l'état associé. Notons aussi $S_1 := \{h \in \mathbb{R}^n_+;\ \|h\| \leq a_1\}$. On va montrer que, pour γ assez petit, si $\tau < T$ est proche de T, alors

$$0 \in \{x_{\gamma,h}(\tau); h \in S_1\}. \tag{10.69}$$

Bien entendu (2.69) contredit l'optimalité de u d'où la conclusion.

Montrons donc que (2.69) est satisfait. Pour $i = 1, \ldots, r$, notons $y_i(\tau)$ la variation au temps τ associée à la perturbation en aiguille associée à y_i (donc $y_i = y_i(T)$). Comme $y_i(\tau)$ est fonction continue de τ, (2.68) implique que pour τ proche de T on a

$$\varepsilon B \subset \operatorname{conv}\left\{y^1(\tau), \ldots, y^r(\tau)\right\}. \tag{10.70}$$

Etant donné $\gamma > 0$, essayons de résoudre en $h \in S_1$ l'équation $x_{\gamma,h}(\tau) = 0$ par l'"algorithme" de linéarisation suivant :

$$h_0 = 0; \quad \sum_{i=1}^r (h_i^{k+1} - h_i^k) y^i(\tau) = -x_{\gamma,h^k}(\tau), \ \ k = 1, \ldots. \tag{10.71}$$

D'après (2.70), cette équation a une solution telle que, pour tout $k \geq 1$,

$$\|h^{k+1} - h^k\| \leq a_2 \|x_{\gamma,h^k}(\tau)\|. \tag{10.72}$$

Notons u^k et x^k les commandes et états associés formés par l'algorithme. Pour que celui-ci soit bien défini pour tout k il faut, pour tout k, vérifier que $\|h^k\| \leq a_1$; c'est le cas pour $k = 0$. Le lemme 2.17 montre que, étant donné $\varepsilon_1 > 0$, pour $\gamma > 0$ assez petit et réduisant a_1 si nécessaire, on a pour tout h et h' dans S_1 :

$$\left\| x_{\gamma,h'}(\tau) - x_{\gamma,h}(\tau) - \sum_{i=1}^{r} (h'_i - h_i) y_i(\tau) \right\| \leq \varepsilon_1 \|h' - h\|. \tag{10.73}$$

Donc tant que $h^k \in S_1$, pour $k \geq 1$, on a avec (2.72) et (2.73)

$$\|h^{k+1} - h^k\| \leq a_2 \|x_{\gamma,h^k}(\tau)\| \leq \varepsilon_1 a_2 \|h^k - h^{k-1}\|, \tag{10.74}$$

et donc prenant $\varepsilon_1 = \frac{1}{2} a_2$, tant que $h^{k+1} \in S_1$, pour $k \geq 1$

$$\|h^{k+1}\| \leq \sum_{i=0}^{k} \|h^{i+1} - h^i\| \leq 2\|h^1 - h^0\| \leq 2a_2 \|x(\tau)\|. \tag{10.75}$$

Si on prend $\tau < T$ tel que $2a_2\|x(\tau)\| < a_1$, on obtient par récurrence que $\|h^k\| \leq a_1$, donc la suite est bien définie ; de plus $\|h^k - h^{k-1}\| \to 0$, donc avec (2.74) $\|x_{\gamma,h^k}(\tau)\| \to 0$. Posant $h^\infty := \lim_k h^k$, on obtient $x_{\gamma,h^\infty} = 0$ comme il fallait le montrer. ■

Traitons enfin le cas général, en commençant par un lemme préliminaire.

Lemme 10.25 *Si C est d'intérieur vide, moyennant si nécessaire un changement d'origine et de la base de $\mathbb{R}^n$ on peut supposer qu'il est de la forme*

$$C = \{x \in \mathbb{R}^n;\ x_i = 0,\ i = 1, \ldots, q;\ \ (x_{q+1}, \ldots, x_n) \in \tilde{C}\}, \tag{10.76}$$

avec $\tilde{C}$ partie convexe de $\mathbb{R}^{n-q}$ d'intérieur non vide.

Démonstration. Le résultat est vrai si C est d'intérieur non vide. Sinon, comme C est convexe, ceci implique qu'il est contenu dans un hyperplan que par changement d'origine et de base on peut supposer de la forme $x_1 = 0$. Posons $C_1 := \{x' \in \mathbb{R}^{n-1};\ (0, x') \in C\}$. Procédant de même pour C_1, on arrive par récurrence au résultat cherché. ■

Démonstration du théorème 2.3. Soit u solution du problème (2.8). En raison du lemme 2.21, il suffit de montrer que $0 \notin \operatorname{int}\left(x(T) + \hat{\mathcal{C}}_T(u) - C\right)$. Procédons par l'absurde : supposons donc (2.65) satisfait. D'après le lemme 2.25, on peut supposer que $x(T) = 0$ et que C est de la forme (2.76). Notons $\tilde{y}$ le vecteur formé des composantes $q+1$ à n de $y \in \mathbb{R}^n$. Nous allons montrer qu'il existe une variation $y \in \hat{C}_T(u)$ telle que

$$y_i = 0,\ i = 1 \text{ à } q;\ \ \tilde{y} \in \operatorname{int} \tilde{C}. \tag{10.77}$$

En effet, posons $K := \hat{C}_T(u) - \{0\}_{\mathbb{R}^q} \times \operatorname{int} \tilde{C}$. Alors (2.77) équivaut à $0 \in K$; comme K est un cône convexe, si $0 \notin K$, alors $0 \in \partial K$. D'après le lemme 1.13, il existe donc

une forme linéaire (non nulle) séparant 0 de K, donc $\hat{C}_T(u)$ de $\{0\}_{I\!R^q} \times \operatorname{int} \tilde{C}$, et donc $\hat{C}_T(u)$ de C, ce qui contredit (2.65). La contradiction ainsi obtenue montre que (2.77) est satisfait.

Soit v_γ la perturbation en aiguille associée à y et x_γ l'état correspondant. Alors $(x_\gamma)_i(T) = o(\gamma)$, $i = 1, \ldots, q$, et $(\tilde{x}_\gamma)_i(T) = 2\gamma y + o(\gamma)$.

Il existe donc $\alpha > 0$ tel que, pour tout $\varepsilon > 0$, si $\gamma > 0$ est assez petit, on a $|x_{\gamma,i}(T)| \leq \frac{1}{2}\varepsilon\gamma$, $i \leq q$, et $\tilde{x}_\gamma(T) + 2\alpha\gamma B(0,1) \subset \tilde{C}$. Pour $\tau < T$ assez proche de T, on aura donc

$$|x_{\gamma,i}(\tau)| \leq \varepsilon\gamma, \quad i \leq q; \quad \tilde{x}_\gamma(\tau) + \alpha\gamma B_{n-q}(0,1) \subset \tilde{C}. \tag{10.78}$$

On procède alors comme dans le cas $C = \{0\}$ pour effectuer une correction assurant $x_i(\tau) = 0$, $i = 1$ à q. Comme cette correction modifie l'état à l'instant τ d'une quantité $O(\varepsilon\gamma)$, ceci assure (pour $\varepsilon > 0$ assez petit) $\tilde{x}(\tau) \in \tilde{C}$ et donc $x(\tau) \in C$ ce qui donne la contradiction recherchée à (2.65). ■

10.5 Notes

On trouvera d'autres approches du principe du maximum dans l'école russe : Alexéev, V. Tikhomirov et Fomine [3], Ioffe and Tihomirov [37]. Pour les extensions au cadre non différentiable on consultera Clarke [21], Frankowska [30].

Chapitre 11

Commande optimale : l'approche HJB

11.1 Cadre

Dans ce chapitre nous étudions une classe de problèmes de commande optimale généralisant les problèmes de transfert en temps minimal. Cette classe est paramétrée par x, la condition initiale sur l'état. Nous montrerons que la valeur du problème est solution, en un sens généralisé, d'une équation aux dérivées partielles en la variable x, dite équation de Hamilton-Jacobi-Bellman (HJB). La commande optimale s'obtient alors en minimisant un hamiltonien faisant intervenir le gradient de la fonction valeur.

La classe de problèmes de commande optimale est la suivante :

$$(P_x) \qquad \begin{cases} \text{Min } \mathcal{V}(x,u,T) := \displaystyle\int_0^T \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t; \\ \\ \dot{y}_{x,u}(t) = f(y_{x,u}(t),u(t)), \quad t \in [0,+\infty[, \quad y_{x,u}(0) = x; \\ \\ y_{x,u}(T) \in C; \quad u(t) \in U, \quad \text{p.p.} \quad t \in [0,+\infty[. \end{cases}$$

Ici la *cible* C est une partie fermée (qui peut être vide ou non convexe) de $\mathbb{R}^n$, u est la *commande*, et doit appartenir à presque chaque instant à l'ensemble U, compact (convexe ou non) de $\mathbb{R}^m$; T est appelé *temps de transfert* de l'état initial x à la cible avec la commande u ; il vaut par définition $+\infty$ si celle-ci n'est jamais atteinte ; $y_{x,u}$ est l'état, $\lambda \geq 0$ est un *coefficient d'actualisation*, $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ est la *dynamique*, et $\ell : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ est le *coût distribué*. Nous faisons les hypothèses suivantes sur f et ℓ :

$$\begin{cases} f & \text{est lipschitzienne,} \\ \ell & \text{est lipschitzienne et bornée.} \end{cases} \tag{11.1}$$

On notera L_f et L_ℓ les constantes de Lipschitz. Ces hypothèses assurent que l'équation d'état admet, pour une commande $u \in L^\infty(0,T,U)$ donnée, une solution unique, et que le *critère* $\mathcal{V}(x,u,T)$ est bien défini si T est fini ou si $\lambda > 0$.

On dit que (u,T) est *admissible* si $u(t) \in U$ p.p. t, $\int_0^T \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t$ est bien définie, et si de plus T est fini, alors $y_{x,u}(T) \in C$. On appelle *valeur* du problème (P_x) la quantité

$$V(x) := \inf\{\mathcal{V}(x,u,T);\ \ (u,T)\ \text{ admissibles}\}. \tag{11.2}$$

Si une commande (u,T) admissible atteint l'infimum, on l'appelle *commande optimale* et on dit qu'elle est *solution* du problème (P_x).

Remarque 11.1 La fonction valeur est, si $\lambda > 0$, une fonction bornée, car de

$$|\mathcal{V}(x,u)| \le \int_0^\infty |\ell(y_{x,u}(t),u(t))|e^{-\lambda t}\mathrm{d}t \le \lambda^{-1}\|\ell\|_\infty, \tag{11.3}$$

on déduit que

$$|V(x)| \le \lambda^{-1}\|\ell\|_\infty. \tag{11.4}$$

De plus V est positive si ℓ l'est. Dans ce dernier cas, $V(x) = 0$ si $x \in C$. Si de plus $\ell(x,u)$ est strictement positif pour tout $(x,u) \in I\!R^n \times U$, et si (u,T) est solution de (P_x), alors T est le premier instant où l'état atteint la cible.

Remarque 11.2 Nous retrouvons le cas particulier des problèmes de transfert en temps minimal dans le cas où ℓ vaut identiquement 1. En effet, le critère à minimiser vaut alors

$$\mathcal{V}(x,u,T) = \int_0^T e^{-\lambda t}\mathrm{d}t = \begin{cases} \lambda^{-1}(1-e^{-\lambda T}) & \text{si } \lambda > 0, \\ T & \text{si } \lambda = 0. \end{cases} \tag{11.5}$$

Minimiser ce critère équivaut bien à minimiser le temps de transfert. En particulier, si $\lambda = 0$, $V(x)$ est égal au temps minimal de transfert $T(x)$.

Un coefficient d'actualisation strictement positif permet de donner une valeur finie (égale à λ^{-1} dans le cas de problèmes de transfert en temps minimal) au critère si la cible n'est pas atteinte, ce qui facilite l'analyse mathématique ainsi que la discussion des procédés d'approximation numérique. Pour cette raison, nous *supposerons dans la suite* $\lambda > 0$.

11.2 Valeur fonction de l'état

11.2.1 Principe de programmation dynamique

Notons l'ensemble des commandes par

$$\mathcal{U} := \{u : [0,\infty[\to I\!R^m \quad \text{mesurable}\,;\quad u(t) \in U, \quad \text{p.p.} \quad t\}\,, \tag{11.6}$$

ou encore (puisque U est borné) $\mathcal{U} = L^\infty(0,\infty,U)$. Notons que, si $x \in I\!R^n \setminus C$, le fait que f soit lipschitzienne et que U soit compact assure que le temps minimal de transfert à C vérifie $T(x) > 0$.

Le théorème ci-dessous énonce le principe de programmation dynamique sous une forme un peu restrictive, mais qui suffit pour l'instant. Une forme plus complète est donnée dans le théorème 4.1.

Théorème 11.3 (Principe de Programmation Dynamique I) *Si $x \in I\!R^n \setminus C$ et $\tau \in]0, T(x)[$, alors la valeur $V(x)$ du problème (P_x) satisfait :*

$$V(x) = \inf_{u\in\mathcal{U}} \left\{ \int_0^\tau \ell(y_{x,u}(t), u(t)) e^{-\lambda t} \mathrm{d}t + e^{-\lambda\tau} V(y_{x,u}(\tau)) \right\}. \tag{11.7}$$

Démonstration. Notons $v^*(x)$ le membre de droite de l'égalité ci-dessus. Rappelons que $\mathcal{V}(x,u,T)$ est le coût associé à l'état initial x et à une commande admissible (u,T). Alors $\tau < T(x) \le T$, donc

$$\begin{aligned}
\mathcal{V}(x,u,T) &= \int_0^\tau \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t + \int_\tau^T \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t,\\
&= \int_0^\tau \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t + e^{-\lambda\tau}\int_0^{T-\tau} \ell(y_{x,u}(t+\tau),u(t+\tau))e^{-\lambda t}\mathrm{d}t,\\
&= \int_0^\tau \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t + e^{-\lambda\tau}\mathcal{V}(y_{x,u}(\tau),u(\cdot+\tau),T-\tau),\\
&\ge \int_0^\tau \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t + e^{-\lambda\tau}V(y_{x,u}(\tau)).
\end{aligned}$$

Minimisant chaque membre par rapport à u, il vient $V(x) \ge v^*(x)$. Pour montrer l'inégalité inverse, fixons $\varepsilon > 0$ et soit $\tilde{u}_\varepsilon$ une solution ε-optimale du problème de minimisation dans (3.7) et $\tilde{y}_\varepsilon$ l'état associé. On a donc

$$v^*(x) \ge \int_0^\tau \ell(\tilde{y}_\varepsilon(t),\tilde{u}_\varepsilon(t))e^{-\lambda t}\mathrm{d}t + V(\tilde{y}_\varepsilon(\tau))e^{-\lambda\tau} - \varepsilon.$$

Soit $(\hat{u}_\varepsilon, T)$ admissible et ε-optimal pour le problème $(P_{\tilde{y}_\varepsilon(\tau)})$ et $\hat{y}_\varepsilon$ l'état associé. Alors

$$\begin{aligned}
V(\tilde{y}_\varepsilon(\tau))e^{-\lambda\tau} + \varepsilon &\ge \int_0^T \ell(\hat{y}_\varepsilon(t),\hat{u}_\varepsilon(t))e^{-\lambda(t+\tau)}\mathrm{d}t &(11.8)\\
&= \int_\tau^{\tau+T} \ell(\hat{y}_\varepsilon(t-\tau),\hat{u}_\varepsilon(t-\tau))e^{-\lambda t}\mathrm{d}t. &(11.9)
\end{aligned}$$

Définissons la commande u_ε par

$$u_\varepsilon(t) = \begin{cases} \tilde{u}_\varepsilon(t) & \text{si} \quad t \in [0,\tau],\\ \hat{u}_\varepsilon(t-\tau) & \text{si} \quad t \in]\tau,\infty], \end{cases} \tag{11.10}$$

et soit y_ε l'état associé. Alors

$$v^*(x) \geq \int_0^{\tau+T} \ell(y_\varepsilon(t), u_\varepsilon(t)) e^{-\lambda t} \mathrm{d}t - 2\varepsilon = \mathcal{V}(x, u_\varepsilon, \tau + T) - 2\varepsilon \geq V(x) - 2\varepsilon.$$

Puisque ε peut être pris arbitrairement petit, ceci entraîne $v^*(x) \geq V(x)$, d'où le théorème. ■

Remarque 11.4 Le choix du poids exponentiel se traduit par une invariance de la valeur par rapport à l'instant initial : c'est la clé de la démonstration ci-dessus.

Remarque 11.5 Le principe de programmation dynamique peut se formuler ainsi : sur un horizon inférieur au temps minimal de transfert, la valeur optimale est égale à l'infimum de la somme du coût de transition entre les états aux instants 0 et τ et de la valeur actualisée en l'état à l'instant τ.

Exemple 11.6 Pour un problème de temps minimal de transfert, $\ell(x, u) = 1$, et le principe de programmation dynamique s'écrit donc :

$$\forall \quad \tau \in]0, T(x)[, \quad V(x) = \lambda^{-1}(1 - e^{-\lambda\tau}) + e^{-\lambda\tau} \inf_{u \in \mathcal{U}} V(y_{x,u}(\tau)). \tag{11.11}$$

11.2.2 Equation de Hamilton-Jacobi-Bellman

En vue de la discrétisation du principe de programmation dynamique, étudions le cas où $\tau \downarrow 0$ dans (3.7). Le lemme technique suivant sera utile à plusieurs reprises.

Lemme 11.7 *Soient $x \in \mathbb{R}^n \setminus C$ et $\tau \in]0, T(x)[$. Alors*

$$\tau\lambda V(x) = \inf_{u \in \mathcal{U}} \left\{ \int_0^\tau \ell(x, u(t)) \mathrm{d}t + V(y_{x,u}(\tau)) - V(x) \right\} + o(\tau). \tag{11.12}$$

Démonstration. Le principe de programmation dynamique peut s'écrire

$$e^{\lambda\tau} V(x) = \inf_{u \in \mathcal{U}} \left\{ \int_0^\tau \ell(y_{x,u}(t), u(t)) e^{\lambda(\tau - t)} \mathrm{d}t + V(y_{x,u}(\tau)) \right\}. \tag{11.13}$$

Puisque f est lipschitzienne, on a $y_{x,u}(t) = x + O(\tau)$, uniformément par rapport $t \in [0, \tau]$ et à la commande. Plus précisément, il existe $c > 0$, tel que, si $\tau > 0$ est assez petit, pour tout $t \in [0, \tau]$, on a

$$\|y_{x,u}(t) - x\| \leq c\tau, \quad \text{pour tout} \quad u \in \mathcal{U}. \tag{11.14}$$

En conséquence,

$$\int_0^\tau \ell(y_{x,u}(t), u(t)) e^{\lambda(\tau - t)} \mathrm{d}t = \int_0^\tau \ell(x, u(t)) \mathrm{d}t + o(\tau), \tag{11.15}$$

là encore uniformément par rapport à la commande. De plus,

$$e^{\lambda\tau}V(x) = (1+\lambda\tau)V(x) + o(\tau). \tag{11.16}$$

Combinant avec (3.13) et (3.15), on obtient

$$\tau\lambda V(x) = \inf_{u\in\mathcal{U}} \left\{ \int_0^\tau \ell(x,u(t))\mathrm{d}t + V(y_{x,u}(\tau)) - V(x) + o(\tau) \right\} + o(\tau). \tag{11.17}$$

avec le premier $o(\tau)$ uniforme par rapport à la commande, et on conclut avec le lemme 1.16. ■

Introduisons le *hamiltonien* $\mathcal{H}$:

$$\mathcal{H}(x,p) \quad := \quad \min_{u\in U}\{\ell(x,u) + p\cdot f(x,u)\}. \tag{11.18}$$

Remarque 11.8 Dans le cas de problèmes en temps optimal, on a introduit en (2.14) le pseudo hamiltonien $H(x,u,p) := p\cdot f(x,u)$ (dans le cas de données autonomes). Dans ce cas $\ell(x,u)=1$, donc $\mathcal{H}(x,p) = 1 + \min_{u\in U} H(x,u,p)$.

Lemme 11.9 *Si V est différentiable en $x\in\mathbb{R}^n\setminus C$, alors*

$$\lambda V(x) = \mathcal{H}(x, DV(x)).$$

Démonstration. Puisque f est lipschitzienne, utilisant (3.14), il vient

$$y_{x,u}(\tau) = x + \int_0^\tau f(y_{x,u}(t),u(t))\mathrm{d}t = x + \int_0^\tau f(x,u(t))\mathrm{d}t + o(\tau), \tag{11.19}$$

avec $o(\tau)/\tau \to 0$ quand $\tau\downarrow 0$, uniformément par rapport à la commande. Comme V est différentiable en x, on a

$$V(y_{x,u}(\tau)) = V(x) + \int_0^\tau DV(x)\cdot f(x,u(t))\mathrm{d}t + o(\tau), \tag{11.20}$$

avec encore un $o(\tau)$ uniforme. Combinant avec les lemmes 1.16 et 3.7, il vient

$$\tau\lambda V(x) = \inf_{u\in\mathcal{U}} \left\{ \int_0^\tau \left[\ell(x,u(t)) + DV(x)f(x,u(t))\right]\mathrm{d}t \right\} + o(\tau). \tag{11.21}$$

L'infimum ci-dessus est atteint en minimisant séparément pour chaque t ; en conséquence,

$$\tau\lambda V(x) = \tau\mathcal{H}(x,DV(x)) + o(\tau), \tag{11.22}$$

d'où la conclusion en divisant par $\tau\downarrow 0$. ■

On appellera *équation de Hamilton-Jacobi-Bellman* (HJB), pour la famille de problèmes de commande optimale (P_x), l'équation aux dérivées partielles non linéaire du premier ordre sur $\mathbb{R}^n\setminus C$ avec conditions aux limites sur C :

$$\left\{ \begin{array}{lllll} \text{(i)} & \lambda v(x) & = & \mathcal{H}(x,Dv(x)), & x\in\mathbb{R}^n\setminus C, \\ \text{(ii)} & v(x) & = & 0, & x\in C, \end{array} \right. \tag{11.23}$$

dans laquelle l'inconnue est la fonction $v:\mathbb{R}^n\to\mathbb{R}$.

Remarque 11.10 L'étude de cette équation aux dérivées partielles présente plusieurs difficultés :

(i) $V(x)$ n'est en général pas différentiable sur $\mathbb{R}^n \setminus C$. Il faut donc donner un sens à (3.23)(i) aux points où $V(x)$ n'est pas différentiable.

(ii) $V(x)$ n'est pas nécessairement continue sur C (voir l'exemple 3.11). Là encore il faut donner un sens à la condition aux limites.

(iii) Il peut y avoir plusieurs solutions continues sur $\mathbb{R}^n$, et différentiable sur $\mathbb{R}^n \setminus C$, de (3.23) (exemple 3.12).

Exemple 11.11 Soit le problème de transfert à 0, en dimension 1, avec la dynamique $\dot{x} = u$, $0 \leq u \leq 1$. Considérons la formulation actualisée avec $\lambda = 1$.

On sait que $V(x) = 1 - e^{-T(x)}$. Or $T(x)$ vaut $-x$ si $x \leq 0$, et $+\infty$ sinon ; donc

$$V(x) = \begin{cases} 1 - e^x & \text{si } x \leq 0, \\ 1 & \text{sinon.} \end{cases} \tag{11.24}$$

La valeur est donc discontinue en 0.

Exemple 11.12 Soit le problème de transfert à 0, en dimension 1, avec la dynamique $\dot{x} = u$, $-1 \leq u \leq 1$. Considérons la formulation actualisée avec $\lambda = 1$. Alors $T(x) = |x|$, et donc $V(x) = 1 - e^{-|x|}$. La valeur est continue, et différentiable en tout point différent de la cible 0. Le hamiltonien a pour expression

$$\mathcal{H}(x, p) = \min_{u \in [-1,1]} \{1 + up\} = 1 - |p|, \tag{11.25}$$

et l'équation HJB s'écrit donc

$$\begin{cases} v(x) &= 1 - |Dv(x)|, \quad x \neq 0, \\ v(0) &= 0. \end{cases} \tag{11.26}$$

La valeur est bien solution de cette équation. Mais les fonctions $w_1(x) = 1 - e^x$ et $w_2(x) = 1 - e^{-x}$ sont d'autres solutions continues et différentiables en tout point différent de 0 (elles sont même différentiables en 0). Notons cependant que ces solutions "parasites" sont non bornées alors que $V(x)$ l'est.

11.2.3 Continuité uniforme de la valeur

Une fonction est d'autant plus facile à approcher numériquement qu'elle est régulière. Montrons que, *si la cible est vide* (On note alors $\mathcal{V}(x, u)$ le critère) la fonction V est hölderienne.

Lemme 11.13 *Si $C = \emptyset$, la fonction valeur $V(x)$ est hölderienne et bornée.*

Démonstration. Nous savons par (3.3)-(3.4) que V est bornée ; montrons qu'elle est uniformément continue. Puisque f est lipschitzien, la quantité

$$\lambda_0 := \sup_{\substack{u\in U\\ x\neq x'}} \frac{(f(x',u)-f(x,u))\cdot(x'-x)}{|x'-x|^2} \tag{11.27}$$

est finie. Montrons que deux trajectoires associées à la même commande u satisfont la relation

$$|y_{x'}(t)-y_{x,u}(t)| \leq |x'-x|e^{\lambda_0 t}. \tag{11.28}$$

En effet, posant $z(t) := y_{x'}(t)-y_{x,u}(t)$, il vient

$$\tfrac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|z(t)|^2 = z(t)\cdot\dot{z}(t) \leq \lambda_0|z(t)|^2,$$

et donc $|z(t)|^2 \leq e^{2\lambda_0 t}|x'-x|^2$, d'où (3.28). Par ailleurs, (1.27) implique

$$|V(x')-V(x)| \leq \sup_{u\in\mathcal{U}}\left\{\int_0^\infty |\ell(y_{x'}(t),u(t))-\ell(y_{x,u}(t),u(t))|e^{-\lambda t}\mathrm{d}t\right\}.$$

Soit $T>0$. Notons

$$\begin{aligned}\Delta_1 &:= \sup_{u\in\mathcal{U}}\int_0^T |\ell(y_{x'}(t),u(t))-\ell(y_{x,u}(t),u(t))|e^{-\lambda t}\mathrm{d}t,\\ \Delta_2 &:= \sup_{u\in\mathcal{U}}\int_T^\infty |\ell(y_{x'}(t),u(t))-\ell(y_{x,u}(t),u(t))|e^{-\lambda t}\mathrm{d}t.\end{aligned}$$

Alors $|V(x')-V(x)| \leq \Delta_1+\Delta_2$. Supposant sans perte de généralité $\lambda_0>\lambda$ (il suffit que λ_0 majore le membre de droite de (3.3)), nous obtenons avec (3.28)

$$\begin{aligned}\Delta_1 &\leq L_\ell\int_0^T |x'-x|e^{(\lambda_0-\lambda)t}\mathrm{d}t = L_\ell\frac{e^{(\lambda_0-\lambda)T}-1}{\lambda_0-\lambda}|x'-x|,\\ \Delta_2 &\leq 2\int_T^\infty \|\ell\|_\infty e^{-\lambda t}\mathrm{d}t = \frac{2}{\lambda}e^{-\lambda T}\|\ell\|_\infty.\end{aligned}$$

Soit x' tel que $|x'-x|<1$. Choisissons $T>0$ tel que $e^{-T}=|x'-x|^{\frac{1}{\lambda_0}}$ (c'est possible !). Alors les quantités Δ_1 et Δ_2 se majorent ainsi :

$$\begin{aligned}\Delta_1 &\leq \frac{L_\ell}{\lambda_0-\lambda}|x'-x|\left(|x'-x|^{\frac{\lambda}{\lambda_0}-1}-1\right) \leq \frac{L_\ell}{\lambda_0-\lambda}|x'-x|^{\frac{\lambda}{\lambda_0}},\\ \Delta_2 &\leq \frac{2}{\lambda}\|\ell\|_\infty|x'-x|^{\frac{\lambda}{\lambda_0}},\end{aligned}$$

et donc

$$|V(x')-V(x)| \leq \left(\frac{L_\ell}{\lambda_0-\lambda}+\frac{2}{\lambda}\|\ell\|_\infty\right)|x'-x|^{\frac{\lambda}{\lambda_0}},$$

d'où la conclusion. ■

11.3 Commande optimale

Sous les hypothèses faites au début du chapitre, il n'existe pas en général de commande optimale (comme le montre l'exemple 2.1). Nous allons cependant, sous des hypothèses fortes, établir dans cette section comment obtenir la commande optimale à partir de la connaissance de la fonction valeur V.

Théorème 11.14 *Supposons la fonction valeur continûment différentiable sur $\mathbb{R}^n \setminus C$, et continue en tout point de C. Soit $x \in \mathbb{R}^n \setminus C$. Alors la commande u est optimale si et seulement si, p.p. $s \in [0,T]$, où T est le premier instant (éventuellement infini) pour lequel la cible est atteinte avec la commande u, cette commande minimise le hamiltonien au sens suivant :*

$$\mathcal{H}(y_{x,u}(s), DV(y_{x,u}(s)) = \ell(y_{x,u}(s), u(s)) + f(y_{x,u}(s), u(s)) \cdot DV(y_{x,u}(s)). \quad (11.29)$$

Démonstration. Soient (u,T) une commande admissible, et $s \in]0,T[$; $y_{x,u}(s)$ n'appartient pas à C, donc V est dérivable en $y_{x,u}(s)$. Le lemme 3.9 et la définition du hamiltonien impliquent

$$\lambda V(y_{x,u}(s)) - f(y_{x,u}(s), u(s)) \cdot DV(y_{x,u}(s)) \leq \ell(y_{x,u}(s), u(s)), \quad (11.30)$$

avec égalité ssi (3.29) est satisfait pour presque tout $s \in [0,T]$.

Soit $\tau \in]0,T[$. La régularité de V permet d'écrire, compte-tenu de (3.30) :

$$\begin{aligned} V(x) - e^{-\lambda\tau} V(y_{x,u}(\tau)) &= \int_0^\tau \frac{\mathrm{d}}{\mathrm{d}t} \left[-e^{-\lambda t} V(y_{x,u}(t))\right] \mathrm{d}t \\ &= \int_0^\tau \left[\lambda V(y_{x,u}(t)) - f(y_{x,u}(t), u(t)) \cdot DV(y_{x,u}(t))\right] e^{-\lambda t} \mathrm{d}t \\ &\leq \int_0^\tau \ell(y_{x,u}(t), u(t)) e^{-\lambda t} \mathrm{d}t, \end{aligned} \quad (11.31)$$

avec égalité ssi (3.29) est satisfait pour presque tout $t \in [0,\tau]$.

Faisons maintenant tendre τ vers T. Si T est fini, de $V(y_{x,u}(T)) = 0$ on déduit que u est optimal ssi (3.29) est satisfait. Si $T = +\infty$, on a $e^{-\lambda\tau} V(y_{x,u}(\tau)) \to 0$ puisque V est bornée, d'où la même conclusion. ■

Remarque 11.15 Le résultat précédent a plusieurs extensions utiles, par exemple au cas où la fonction valeur V est seulement dérivable en tout $y_{x,u}(s)$, $s \in]0,T[$, sauf peut-être en un nombre fini d'entre eux.

Le théorème précédent donne le moyen de vérifier si une commande *fonction du temps* est optimale. Voyons maintenant le résultat principal de la section, qui montre comment construire la commande optimale *en fonction de l'état (forme feedback)* :

Théorème 11.16 *Supposons* (i) *la fonction valeur continûment différentiable sur* $\mathbb{R}^n \setminus C$*, de dérivée localement lipschitzienne, et continue en tout point de* C*,* (ii) *le minimum dans la définition du hamiltonien* (3.18) *atteint en un point unique* $\Upsilon(x,p)$*, quand* $p = DV(x)$*, la fonction* Υ *étant localement lipschitzienne.*

Alors la commande ci-dessous, sous forme feedback, est optimale :

$$u(x) = \Upsilon(x, DV(x)). \tag{11.32}$$

Démonstration. L'équation différentielle

$$\dot{y}_{x,u}(t) = f(y_{x,u}(t), \Upsilon(y_{x,u}(t), DV(y_{x,u}(t)))) \tag{11.33}$$

a un second membre borné et localement lipschitzien, donc a une solution unique ; l'optimalité de la commande découle du théorème 3.14. ■

Exemple 11.17 reprenons le problème de l'exemple 3.12. On a $V(x) = 1 - e^{-|x|}$, et $V'(x) = -e^x$ si $x < 0$, $V'(x) = e^{-x}$ si $x > 0$. La commande réalisant le maximum dans la définition du hamiltonien est donc $u(x) = 1$ si $x < 0$, $u(x) = -1$ si $x > 0$. Chacun des deux théorèmes précédents peut être appliqué à ce problème.

Remarque 11.18 La vérification de l'hypothèse (ii) du théorème 3.16 se ramène à une analyse de stabilité de la solution d'un problème d'optimisation en dimension finie, voir [16, Section 4.4.1]. Cette hypothèse se vérifie par exemple dans le cas (assez restrictif) où U est convexe fermé, f est affine par rapport à la commande, et ℓ est uniformément fortement convexe par rapport à la commande, pour tout x.

Remarque 11.19 La fonction valeur n'est en général pas continûment différentiable, même sous les hypothèses fortes de la remarque 3.18. Les théorèmes 3.14 et 3.16 ne peuvent donc être appliqués que dans un nombre limité de cas.
On retiendra néanmoins la règle heuristique suivante. Soit $x \in \mathbb{R}^n \setminus C$. Alors un "candidat sérieux", pour être commande optimale en x, est l'argument de la minimisation dans la définition du hamiltonien, évaluant celui-ci en $(x, DV(x))$.

11.4 Solution de viscosité

Cette section présente une notion permettant de donner un sens à l'équation (3.23), dite HJB :

$$\left\{ \begin{array}{lllll} \text{(i)} & \lambda v(x) & = & \mathcal{H}(x, Dv(x)), & x \in \mathbb{R}^n \setminus C, \\ \text{(ii)} & v(x) & = & 0, & x \in C, \end{array} \right. \tag{11.34}$$

avec C partie fermée de $\mathbb{R}^n$, *même quand la solution n'est pas différentiable.* On pourra passer les preuves en première lecture. Nous limiterons l'étude aux solutions continues sur $\mathbb{R}^n$. Le problème principal est de donner un sens à (3.34)(i), de manière à ce que la valeur soit l'unique solution de (3.34).

11.4.1 Notion de solutions de viscosité

Notons dans la suite

$$\Omega := I\!R^n \setminus C. \tag{11.35}$$

On peut définir une notion de solution généralisée de (3.34) grâce à l'observation suivante.

Lemme 11.20 *Soit Φ une fonction différentiable en $x \in \Omega$, telle que $V - \Phi$ a un maximum (resp. minimum) local en x. Alors*

$$\lambda V(x) - \mathcal{H}(x, D\Phi(x)) \quad \leq 0 \quad (\text{resp.} \geq 0). \tag{11.36}$$

Démonstration. Il suffit de donner la démonstration dans le cas où $V - \Phi$ a un maximum local en x. Alors, pour tout x' dans un voisinage $\mathcal{N}$ de x, on a

$$V(x') - V(x) \leq \Phi(x') - \Phi(x). \tag{11.37}$$

Pour $\tau > 0$ assez petit, puisque f est bornée, $y_{x,u}(\tau) \in \mathcal{N}$, quelle que soit la commande appliquée. Combinant (3.37) et le lemme 3.7, il vient

$$\tau\lambda V(x) \leq \inf_{u\in\mathcal{U}} \left\{ \int_0^\tau \ell(x, u(t))\mathrm{d}t + \Phi(y_{x,u}(\tau)) - \Phi(x) \right\} + o(\tau). \tag{11.38}$$

On procède alors comme dans la démonstration du lemme 3.9, en adaptant (3.20), (3.21) et (3.22) (changements d'égalités en inégalités, remplacement de V par Φ dans les membres de droite). ■

Formalisons ce qui précède en introduisant un vocabulaire adapté.

Définition 11.21 Une fonction $v : I\!R^n \to I\!R$ est dite *sous* (resp. *sur*) *solution* au *sens de viscosité* de (3.34)(i) si, pour tout $x_0 \in \Omega$, et $\Phi : I\!R^n \to I\!R$ de classe C^1, telle que x_0 est point de maximum (resp. minimum) local de $v - \Phi$, alors

$$\lambda v(x_0) - \mathcal{H}(x_0, D\Phi(x_0)) \quad \leq 0 \quad (\text{resp.} \geq 0). \tag{11.39}$$

On dit que v est *solution au sens de viscosité* de (3.34)(i) si elle est à la fois sur et sous solution au sens de viscosité.

Théorème 11.22 *La fonction valeur V est solution au sens de viscosité de* (3.34)(i).

Le théorème est conséquence immédiate du lemme 3.20. Ce dernier est apparemment plus fort, car il suppose seulement la fonction Φ dérivable en x. Nous allons voir que les deux énoncés sont équivalents, en introduisant un concept important.

Définition 11.23 Soit v une fonction $\mathbb{R}^n \to \mathbb{R}$. On dit que $p \in \mathbb{R}^n$ est une *sous* (resp. *sur*) *dérivée* (ou sous, sur gradient) de v en x, si

$$v(x') - v(x) - p \cdot (x' - x) \geq o(\|x' - x\|) \quad (\text{resp. } \leq o(\|x' - x\|)) \,. \tag{11.40}$$

On notera $D^- v(x)$ (resp. $D^+ v(x)$) l'ensemble des sous gradients (resp. sur gradients) de v en x.

Exemple 11.24 La fonction valeur absolue $v : \mathbb{R} \to \mathbb{R}$, $v(x) := |x|$, est telle que $D^- v(0) = [-1, 1]$ et $D^+ v(0) = \emptyset$.

Remarque 11.25 (i) Si v est dérivable en un point x, alors

$$D^- v(x) = D^+ v(x) = \{Dv(x)\}. \tag{11.41}$$

(ii) Si en un point x on a $D^- v(x) \neq \emptyset$ et $D^+ v(x) \neq \emptyset$, alors v est dérivable en x et (3.41) est satisfait.

Soit p un sur gradient de v en x. Posons

$$\Phi(x') = \max(v(x'), v(x) + p \cdot (x' - x)).$$

Alors $v - \Phi$ atteint un maximum local en x et, par définition du sur gradient, la fonction Φ a pour dérivée p en x. Réciproquement, si une fonction Φ dérivable en x est telle que $v - \Phi$ atteint un maximum local en x, il est clair que $D\Phi(x)$ est un sur gradient de v en x. Nous avons montré que

$$D^+ v(x) = \{p;\ \exists \Phi : \mathbb{R}^n \to \mathbb{R};\ p = D\Phi(x);\ v - \Phi \text{ a un maximum local en } x\}.$$

On peut montrer (voir par exemple Barles [10, Section 2.2]) que $D^+ v(x)$ est aussi l'ensemble des gradients de fonctions continûment dérivables, telles que $v - \Phi$ atteint un maximum local en x. Bien entendu on a un résultat similaire pour les sous gradients. Les conditions du lemme 3.20 et du théorème 3.22 coïncident donc. Ceci implique le résultat suivant :

Lemme 11.26 *Soit $x \in \Omega$ et $v : \mathbb{R}^n \to \mathbb{R}$. Les énoncés suivants sont équivalents :*
(i) *On a $\lambda v(x) \leq \mathcal{H}(x, D\Phi(x))$, pour toute fonction Φ dérivable en x telle que $v - \Phi$ atteint un maximum local en x.*
(ii) *On a $\lambda v(x) \leq \mathcal{H}(x, D\Phi(x))$, pour toute fonction Φ continûment dérivable telle que $v - \Phi$ atteint un maximum local en x.*
(ii) *On a $\lambda v(x) \leq \mathcal{H}(x, p)$, pour tout $p \in D^+ v(x)$.*

Nous laissons le lecteur énoncer le résultat correspondant concernant les sous gradients.

Remarque 11.27 (i) Soit v une sous solution de 3.36(i) au sens de viscosité, et $x \in \Omega$ tel que v soit dérivable en x. Combinant le lemme précédent et la remarque 3.25, il vient $\lambda v(x) \leq \mathcal{H}(x, Dv(x))$. De même pour les sur solutions.
(ii) Soit v dérivable sur Ω. Combinant le point (i) et le lemme précédent, on voit que v est sous solution de 3.36(i) au sens classique ssi elle est sous solution de viscosité. De même pour les sur solutions.

11.4.2 Théorème de comparaison

Nous avons noté que la fonction valeur V n'est pas toujours continue. Dans tous les cas, cette solution est solution de l'équation HJB (3.34) au sens de viscosité.

Le résultat principal de cette section (théorème 3.31) implique, si la cible C est vide, l'unicité (autrement dit, l'existence d'au plus une) d'une solution hölderienne et bornée de l'équation HJB. Sachant que V est hölderienne dans ce cas, on obtient donc l'existence et l'unicité de la solution, dans la classe des fonctions hölderiennes et bornées.

Pour l'étude de convergence des schémas numériques, nous avons besoin d'un résultat un peu plus fort que l'unicité : des résultats de comparaison entre les sous-solutions semi continues supérieurement (s.c.s.) et les sur solutions semi continues inférieurement (s.c.i.).

Définition 11.28 On dit que la fonction $v : \mathbb{R}^n \to \mathbb{R}$ est semi continue supérieurement (s.c.s.) (resp. semi continu inférieurement (s.c.i.)) si pour tout $x \in \mathbb{R}^n$ on a

$$v(x) \geq \limsup_{x' \to x} v(x'), \quad \left(\text{resp. } v(x) \leq \liminf_{x' \to x} v(x')\right).$$

Remarque 11.29 On peut exprimer les propriétés précédentes à l'aide de suites convergentes vers x. Ainsi, la fonction $v : \mathbb{R}^n \to \mathbb{R}$ est s.c.s. ssi, pour toute suite x_k convergeant vers x, on a $v(x) \geq \limsup_k v(x_k)$; ou encore, si pour tout point d'adhérence $v^* \in \mathbb{R} \cup \{\pm\infty\}$ de $v(x^k)$, on a $v(x) \geq v^*$. De même pour la semi continuité inférieure.

Définition 11.30 On appelle *principe d'unicité fort* pour l'équation (3.34) tout résultat du type suivant : Soient v (resp. w) une sous solution (resp. sur solution) de (3.34) (assorti éventuellement de conditions de régularité sur v et w satisfaites par la fonction valeur). Alors $\sup v \leq \inf w$.

Compte tenu de la difficulté de ce type de résultat, nous limiterons l'analyse au cas $C = \emptyset$. Pour l'extension aux problèmes avec temps d'arrêt (section 3.5.1), il est utile de considérer une équation aux dérivées partielles générale du premier ordre, notée

$$\overline{\mathbf{H}}(x, v(x), Dv(x)) = 0, \quad \text{pour tout } x \in \mathbb{R}^n. \tag{11.42}$$

On suppose que le "hamiltonien abstrait" $\overline{\mathbf{H}}$ vérifie les relations

$$\begin{aligned} |\overline{\mathbf{H}}(x,v,p') - \overline{\mathbf{H}}(x,v,p)| &\leq c_1\|p'-p\|; && (11.43)\\ |\overline{\mathbf{H}}(x',v,p) - \overline{\mathbf{H}}(x,v,p)| &\leq c_2\|x'-x\|(1+\|p\|); && (11.44)\\ \overline{\mathbf{H}}(x,v',p) - \overline{\mathbf{H}}(x,v,p) &\geq c_3(v'-v), && (11.45)\end{aligned}$$

avec $c_3 > 0$. Dans le cas de l'équation HJB on a

$$\overline{\mathbf{H}}(x,v,p) := \lambda v - \inf_{u\in U}\{\ell(x,u) + p\cdot f(x,u)\}, \tag{11.46}$$

et si la dynamique est bornée, on vérifie (3.43)-(3.45), avec

$$c_1 := \sup\{\|f(x,u)\|;\ (x,u)\in\Omega\times U\}; \quad c_2 := L_\ell + L_f; \quad c_3 := \lambda. \tag{11.47}$$

Une fonction $v : \mathbb{R}^n \to \mathbb{R}$ est dite *sous solution* (resp. *sur solution*) au *sens de viscosité* de (3.42) si, pour tout $\Phi : \mathbb{R}^n \to \mathbb{R}$ de classe C^1, telle que x_0 est point de maximum (resp. minimum) local de $v - \Phi$, on a

$$\overline{\mathbf{H}}(x_0, v(x_0), D\Phi(x_0)) \quad \leq 0 \quad (\text{resp.} \geq 0). \tag{11.48}$$

On dit que v est *solution au sens de viscosité* de (3.42) si elle est à la fois sur et sous solution au sens de viscosité.

Théorème 11.31 (Principe d'unicité fort) *Sous les hypothèses* (3.43)-(3.45)*, si v est une sous solution s.c.s. bornée supérieurement de* (3.42)*, et w est une sur solution s.c.i. bornée inférieurement de* (3.42)*, une de ces deux fonction étant hölderienne, alors $v(x) \leq w(x)$, pour tout $x \in \mathbb{R}^n$.*

Corollaire 11.32 *Si la dynamique est bornée et $C = \emptyset$, la fonction valeur $V(x)$ du problème (P_x) est l'unique solution de viscosité continue et bornée sur $\mathbb{R}^n$ de l'équation HJB* (3.34)*.*

Démonstration. Le lemme 3.13 dit que la fonction valeur $V(x)$ est hölderienne et bornée. D'après le théorème 3.22, $V(x)$ est solution de viscosité dans $\mathbb{R}^n$ de (3.34)(i). Soit v une autre solution de viscosité continue et bornée. Le théorème 3.31 implique $v \leq V$ et $V \leq v$, d'où $v = V$. ■

Il reste à démontrer le théorème 3.31. La démonstration est quelque peu technique et le lecteur intéressé principalement par les méthodes numériques peut la sauter en première lecture. Donnons cependant un résultat de comparaison élémentaire (mais sous des hypothèses trop fortes) qui donnera une idée du principe de la démonstration.

Proposition 11.33 *On suppose que $C = \emptyset$. Soient v et w une sous et sur solution de* (3.48) *respectivement. Supposons le maximum de $v - w$ atteint en un point x_0 où v et w sont différentiables. Alors $v(x) \leq w(x)$, pour tout $x \in \mathbb{R}^n$.*

Démonstration. Puisque $v - w$ atteint son maximum en x_0, $Dv(x_0) = Dw(x_0)$. Cette valeur commune p_0 étant sur et sous gradient de v et w en x_0, on a

$$\overline{\mathbf{H}}(x_0, v(x_0), p_0) \leq 0 \leq \overline{\mathbf{H}}(x_0, w(x_0), p_0), \tag{11.49}$$

et on conclut avec (3.45). ■

Démonstration du théorème 3.31. Supposons v hölderienne, l'autre cas se traitant d'une manière similaire. L'idée essentielle de la démonstration est le dédoublement des variables : Pour tout $\varepsilon > 0$, posons

$$\varphi(x, y) := v(x) - w(y) - \tfrac{1}{2}\varepsilon^{-2}\|x - y\|^2, \quad \text{pour tout } (x, y) \in I\!R^n \times I\!R^n. \tag{11.50}$$

Le rôle du dernier terme est d'obtenir des points x et y proches quand on considère des solutions approchées du problème de maximisation de φ. Soit $\delta \in]0, 1[$. On a

$$\sup \varphi \leq \sup v - \inf w < +\infty, \tag{11.51}$$

donc il existe (x_1, y_1) in $I\!R^{2n}$ tel que

$$\varphi(x_1, y_1) > \sup \varphi - \delta. \tag{11.52}$$

Il existe aussi une fonction ξ, de classe C^∞ à support compact, telle que

$$\xi(x_1, y_1) = 1, \qquad 0 \leq \xi \leq 1, \quad \sup_{x,y} \|D\xi(x, y)\| \leq 1. \tag{11.53}$$

Posons

$$\psi(x, y) = \varphi(x, y) + \delta\xi(x, y), \qquad \text{pour tout} \quad (x, y) \in I\!R^{2n}. \tag{11.54}$$

Si (x, y) n'est pas dans le support de ξ, on a

$$\psi(x, y) = \varphi(x, y) \leq \sup \varphi < \psi(x_1, y_1). \tag{11.55}$$

Une suite maximisante pour ψ est donc, à partir d'un certain rang, incluse dans le support de ξ qui est compact ; or ψ est s.c.s., donc atteint son maximum en un point (x_o, y_o). Autrement dit,

$$\psi(x_o, y_o) \geq \psi(x, y) \quad \text{pour tout } (x, y) \in I\!R^{2n}. \tag{11.56}$$

En particulier, la fonction $x \to v(x) - \frac{1}{2}\varepsilon^{-2}\|x - y_0\|^2 + \delta\xi(x, y_0)$ atteint un maximum local en x_0. Par définition d'une sous solution de viscosité, on a donc

$$\overline{\mathbf{H}}(x_0, v(x_0), \varepsilon^{-2}(x_o - y_o) - \delta D_x\xi(x_o, y_o)) \leq 0. \tag{11.57}$$

De même, $y \to w(y) + \frac{1}{2}\varepsilon^{-2}\|x_0 - y\|^2 - \delta\xi(x_0, y)$ atteint un minimum local en x_0, donc par définition d'une sur solution de viscosité, on a

$$\overline{\mathbf{H}}(y_0, w(y_0), \varepsilon^{-2}(x_o - y_o) + \delta D_y\xi(x_o, y_o)) \geq 0. \tag{11.58}$$

Utilisant (3.43), (3.53) et (3.57)-(3.58), il vient

$$\overline{\mathbf{H}}(x_0, v(x_0), \varepsilon^{-2}(x_o - y_o)) \leq \delta c_1; \quad \overline{\mathbf{H}}(y_0, w(y_0), \varepsilon^{-2}(x_o - y_o)) \geq -\delta c_1. \tag{11.59}$$

Soustrayant ces relations, nous obtenons avec (3.44) et (3.45),

$$\begin{aligned} c_3(v(x_0) - w(y_0)) &\leq \overline{\mathbf{H}}(x_0, v(x_0), \varepsilon^{-2}(x_o - y_o)) - \overline{\mathbf{H}}(x_0, w(y_0), \varepsilon^{-2}(x_o - y_o)) \\ &\leq \overline{\mathbf{H}}(y_0, w(y_0), \varepsilon^{-2}(x_o - y_o)) - \overline{\mathbf{H}}(x_0, w(y_0), \varepsilon^{-2}(x_o - y_o)) + 2\delta c_1 \\ &\leq c_2 \varepsilon^{-2} \|x_o - y_o\|^2 + c_2 \|x_o - y_o\| + 2\delta c_1. \end{aligned} \tag{11.60}$$

Estimons maintenant les membres de cette inégalité. On a, avec (3.56),

$$\sup \varphi - \delta \leq \psi(x_o, y_o) = v(x_0) - w(y_0) - \tfrac{1}{2}\varepsilon^{-2}\|x_o - y_o\|^2, \tag{11.61}$$

et donc

$$\tfrac{1}{2}\varepsilon^{-2}\|x_o - y_o\|^2 \leq \sup v - \inf w + \delta - \sup \varphi. \tag{11.62}$$

Ceci implique $\|x_o - y_o\| \to 0$ quand $\varepsilon \downarrow 0$. Notons c_v la constante de Hölder de v. Prenant ε assez petit, comme v est hölderienne, il vient $v(x_0) - v(y_0) \leq c_v \|x_0 - y_0\|^\gamma$. Choisissant $x = y = y_0$ dans (3.56), il vient après simplification et usage de (3.53),

$$\begin{aligned} \tfrac{1}{2}\varepsilon^{-2}\|x_o - y_o\|^2 &\leq v(x_0) - v(y_0) + \delta(\xi(x_0, y_0) - \xi(y_0, y_0)) \\ &\leq c_v \|x_0 - y_0\|^\gamma + \delta \|x_0 - y_0\|. \end{aligned} \tag{11.63}$$

On peut sans perte de généralité supposer γ dans $]0, 1[$, donc quand $\|x_0 - y_0\|$ est assez petit

$$c_v \|x_0 - y_0\|^\gamma + \delta \|x_0 - y_0\| \leq \tfrac{1}{2} K |x_o - y_o|^\gamma \tag{11.64}$$

pour une certaine constante K indépendante de ε et δ. Avec (3.63), nous obtenons $\varepsilon^{-2}\|x_o - y_o\|^2 \leq K\|x_o - y_o\|^\gamma$, soit $\|x_o - y_o\| \leq K\varepsilon^{\frac{2}{2-\gamma}}$. Combinant avec (3.60), il vient

$$\lambda(v(x_0) - w(y_0)) \leq K' \varepsilon^{\frac{2\gamma}{2-\gamma}} + 2\delta c_1, \tag{11.65}$$

pour un certain K' indépendant de ε et δ. Or

$$\sup(v - w) \leq \sup \varphi \leq \varphi(x_0, y_0) + \delta \leq v(x_0) - w(y_0) + \delta. \tag{11.66}$$

Il vient donc $\sup(v - w) \leq O(\varepsilon^{\frac{2\gamma}{2-\gamma}} + \delta)$. Faisant tendre ε et δ vers 0, nous obtenons la conclusion. ■

Remarque 11.34 La preuve ci-dessus a l'intérêt d'être très proche de celle de l'estimation d'erreur du schéma de discrétisation : voir la section 4.3.2.

11.5 Temps d'arrêt et commande impulsionnelle

Les résultats principaux de cette section concernent les problèmes de commande impulsionnelle. Afin de préparer les outils nécessaires à leur étude, nous étudions d'abord les problèmes avec décision d'arrêt, qui ont leur propre intérêt.

11.5.1 Problèmes avec temps d'arrêt

Nous considérons un problème de commande optimale dans lequel on peut s'arrêter à tout instant en payant un coût φ actualisé :

$$(P_x) \qquad \begin{cases} \text{Min } \mathcal{V}(x,u,\theta) := \displaystyle\int_0^\theta \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t + e^{-\lambda\theta}\varphi(y_{x,u}(\theta)) \\ \dot{y}_{x,u}(t) = f(y_{x,u}(t),u(t)), \quad t\in[0,\theta[, \quad y_{x,u}(0)=x; \\ u(t)\in U \text{ p.p. } t\in[0,\theta[, \end{cases}$$

avec $\theta \geq 0$ et $\varphi : \mathbb{R}^n \to \mathbb{R}$ supposée lipschitzienne et bornée, ainsi que f et ℓ. On note $a \wedge b := \min(a,b)$, et χ_s vaut 1 si s est vrai, et 0 sinon. La démonstration du théorème ci-dessous ne présente pas de difficulté.

Théorème 11.35 (Principe de Programmation Dynamique)
La fonction valeur $V(x)$ satisfait, pour tout $\tau > 0$:

$$V(x) = \inf_{(u,\theta)} \Big(\int_0^{\tau\wedge\theta} \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t + \chi_{\tau<\theta}e^{-\lambda\tau}V(y_{x,u}(\tau)) + \chi_{\tau\geq\theta}e^{-\lambda\theta}\varphi(y_{x,u}(\theta))\Big), \tag{11.67}$$

où le minimum s'entend sous les contraintes $u(t)\in U$, p.p. $t\in[0,\tau]$, et $\theta\geq 0$.

Soit $\mathcal{H}(x,p)$ toujours défini par (3.18). L'équation HJB de ce problème est dite inéquation variationnelle, par analogie avec les problèmes de contact en mécanique :

$$\max[\lambda v - \mathcal{H}(x,Dv), v-\varphi(x)] = 0, \qquad \text{pour tout } x\in\mathbb{R}^n. \tag{11.68}$$

Théorème 11.36 *La fonction valeur V du problème ci-dessus est bornée, hölderienne, et c'est une solution au sens de viscosité de* (3.68)*, au sens où, pour tout $x\in\mathbb{R}^n$:*

$$\max[\lambda V(x) - \mathcal{H}(x,p), V(x)-\varphi(x)] \leq 0 \quad (\text{resp. } \geq 0), \tag{11.69}$$

pour tout $p\in D^+v(x)$ (resp. $p\in D^-v(x)$).

Démonstration. La démonstration du caractère borné et hölderien de V est similaire à celle du lemme 3.13. Montrons que V est solution de viscosité. Il est clair que $V(x)\leq\varphi(x)$ pour tout x, puisqu'une impulsion à l'instant initial est possible. Distinguons deux cas.
a) Si $V(x) < \varphi(x)$, puisque V et φ sont continues, il existe $\varepsilon > 0$ tel que $V(x') + \varepsilon < \varphi(x')$, pour tout x' appartenant à un voisinage $\mathcal{N}$ de x. Puisque f est bornée, on déduit que pour τ assez petit, toute stratégie optimale à ε près ne comporte pas d'impulsion pour $t\in[0,\tau]$. Le principe de programmation dynamique (3.7) est donc valable pour τ assez petit. La démonstration du lemme 3.20 s'applique

donc ; elle montre que (3.36) est satisfaite en x si $V - \Phi$ a un maximum (resp. minimum) local en x. On en déduit (3.69) en combinant avec le lemme 3.26.
b) Si $V(x) = \varphi(x)$, le second cas de (3.69) est trivialement satisfait. Reste à montrer que si $p \in D^+v(x)$, alors $\lambda V(x) - \mathcal{H}(x,p) \leq 0$. Puisque les stratégies sans impulsions sont possibles, on a

$$V(x) \leq \inf_{u \in \mathcal{U}} \left\{ \int_0^\tau \ell(y_{x,u}(t), u(t)) e^{-\lambda t} \mathrm{d}t + V(y_{x,u}(\tau)) e^{-\lambda \tau} \right\}. \tag{11.70}$$

Il suffit alors de reprendre les calculs des lemmes 3.7 et 3.20, en tenant compte de l'inégalité dans (3.70), pour vérifier que (3.36) est satisfaite, si Φ une fonction différentiable en x, telle que $V - \Phi$ a un maximum (resp. minimum) local en x. On conclut avec le lemme 3.26. ■

Théorème 11.37 (Unicité forte) *Soient v une sous solution s.c.s. de* (3.68) *bornée supérieurement, w une sur solution s.c.i. de* (3.68) *bornée inférieurement. Si une de ces deux fonctions est hölderienne, alors $v(x) \leq w(x)$, pour tout $x \in \mathbb{R}^n$.*

Démonstration. Il suffit d'appliquer le théorème 3.31 ; la vérification des hypothèses (3.43)-(3.45) se fait sans difficultés. ■

Il résulte des résultats précédents que V est l'unique soluion au sens de viscosité de (3.68), dans la classe des fonctions hölderiennes et bornées.

11.5.2 Commande impulsionnelle

Dans de nombreux problèmes de commande optimale, on a la possibilité de faire changer l'état de manière discontinue, en payant un prix associé. Un exemple typique est celui de la gestion de stock, dans lequel une commande a un coût fixe (déplacement du camion) et un coût proportionnel à la quantité livrée (n'excédant pas la capacité du camion). La modification de l'état peut ne s'effectuer qu'après un certain délai (temps de livraison).

Nous allons nous limiter ici à la discussion de problèmes de commande optimale impusionnelle sans délai. La dynamique du système est régie par les relations suivantes :

$$\begin{array}{lcll} \dot{y}_{x,u}(t) & = & f(y_{x,u}(t), u(t)), & t \in]\theta_i, \theta_{i+1}[, \\ y_{x,u}(\theta_i^+) & = & y_{x,u}(\theta_i^-) + \xi_i, & i = 1, \ldots, N, \\ y_{x,u}(0) & = & x. & \end{array} \tag{11.71}$$

La *dynamique* $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ est supposée *lipschitzienne* et *bornée*, ainsi que l'*ensemble des commandes* $U \subset \mathbb{R}^m$, supposé compact, et le coefficient d'actualisation $\lambda > 0$. On convient de noter L_f la constante de Lipschitz de f, et de même pour les autres fonctions. La suite $\{\theta_i\}$, $i = 1, \ldots, N$, de *temps d'arrêt* positifs, est finie (on pose alors $\theta_{N+1} = +\infty$) ou non (on a alors $N = +\infty$), croissante et sans points d'accumulation, et $\theta_0 = 0$. Les impulsions ξ_i appartiennent à $\Xi \subset \mathbb{R}^n$. Les suites θ et ξ font partie de la commande. Ainsi *l'état* $y_{x,u}(t)$ appartient à $\mathbb{R}^n$ et la commande, ou

contrôle, $u(t)$ appartient à $\mathbb{R}^m$. Le critère à minimiser se décompose en une intégrale d'un *coût distribué* et une somme de *coûts de transition* :

$$\mathcal{V}(x,u,\theta,\xi) := \int_0^\infty \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t + \sum_{i=1}^{N}(c_0+c(\xi_i))e^{-\lambda\theta_i}. \tag{11.72}$$

Le coût distribué $\ell : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ est supposé *lipschitzien* et *borné*. Le coût de transition est $c_0 + c(\xi_i)$. La constante $c_0 > 0$ représente un coût fixe, et la fonction continue $c : \mathbb{R}^n \to \mathbb{R}_+$ est telle que $c(0) = 0$ et

$$c(\xi_1+\xi_2) \le c(\xi_1)+c(\xi_2), \quad \forall\, \xi_1,\xi_2 \in \mathbb{R}^n. \tag{11.73}$$

La stricte positivité de c_0 donne une borne sur le nombre d'impulsions d'une stratégie sous optimale sur un intervalle de temps fini, et la relation précédente implique qu'il n'est pas restrictif d'imposer que les instants θ_i soient tous différents. Le problème à résoudre est

$$(P_x) \qquad \min_{(u,\theta,\xi)} \mathcal{V}(x,u,\theta,\xi) \qquad \text{soumis à (3.71)}; \quad u(t)\in U, \ \text{ p.p. } t\in[0,+\infty[.$$

La *valeur* de ce problème (infimum du critère sur les commandes admissibles) est notée $V(x)$. Nous allons dans un premier temps établir un résultat de régularité de la fonction valeur ainsi que le principe de programmation dynamique pour un problème sans impulsion.

Proposition 11.38 *La fonction valeur V est hölderienne et bornée.*

Démonstration. a) Montrons que V est bornée. Soit la commande constante $u(t) = u_0$, où $u_0 \in U$, sans impulsion. Alors

$$V(x) \le \int_0^\infty \ell(y_{x,u}(t),u_0)e^{-\lambda t}\mathrm{d}t \le \lambda^{-1}\|\ell\|_\infty.$$

D'autre part, puisque le coût de transition est positif, on a pour tout commande (u,θ,ξ)

$$V(x) \ge \int_0^\infty \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t \ge -\lambda^{-1}\|\ell\|_\infty,$$

et donc $\|V\|_\infty \le \lambda^{-1}\|\ell\|_\infty$.
b) Montrons que V est hölderienne. On a

$$V(x') - V(x) \le \sup_{(u,\theta,\xi)} \left\{\mathcal{V}(x',u,\theta,\xi) - \mathcal{V}(x,u,\theta,\xi)\right\},$$

donc après simplification des coûts de transition,

$$V(x') - V(x) \le \sup_{(u,\theta,\xi)} \left\{\int_0^\infty [\ell(y_{x'}(t),u(t)) - \ell(y_{x,u}(t),u(t))]e^{-\lambda t}\mathrm{d}t\right\}.$$

Soient $y_{x'}$ et $y_{x,u}$ deux trajectoires associées à la même commande (u, θ, ξ). Comme dans le cas sans impulsion, on a $|y_{x'} - y_{x,u}| \leq |x' - x|e^{\lambda_0 t}$, où λ_0 est défini par (3.3). On peut alors finir la démonstration de manière analogue à celle du lemme 3.13. ■

Théorème 11.39 (Principe de Programmation Dynamique)
La fonction valeur $V(x)$ satisfait, pour tout $\tau > 0$:

$$V(x) = \inf_{(u,\theta,\xi)} \left(\int_0^\tau \ell(y_{x,u}(t), u(t)) e^{-\lambda t} \mathrm{d}t + \sum_{i=1}^{N} (c_0 + c(\xi_i)) e^{-\lambda \theta_i} + e^{-\lambda \tau} V(y_{x,u}(\tau^-)) \right), \tag{11.74}$$

où le minimum s'entend sous les contrainte $u(t) \in U$, p.p. $t \in [0, \tau]$, les θ_i sont strictement croissants, et $\theta_N < \tau$.

Démonstration. La démonstration est similaire à celle du théorème 3.3. ■

Définissons l'opérateur qui à une fonction $w(\cdot)$ associe la valeur optimale après impulsion, noté

$$Mw(x) := \inf_{\xi \in \mathbb{R}^n} \{w(x + \xi) + c_0 + c(\xi)\}. \tag{11.75}$$

Lemme 11.40 *L'opérateur M associe à une fonction bornée une autre fonction bornée, est non expansif*[1] *pour la norme uniforme, et conserve les constantes de Hölder.*

Démonstration. Soient w et w' fonctions bornées dans $\mathbb{R}^n$. Puisque $c(\cdot) \geq 0$ et $c(0) = 0$, on a

$$c_0 - \|w\|_\infty \leq Mw(x) \leq c_0 + w(x) \leq c_0 + \|w\|_\infty,$$

donc M associe à une fonction bornée une autre fonction bornée. Avec (1.27), il vient $|Mw'(x) - Mw(x)| \leq |w'(x) - w(x)|$. Il en résulte que M est non expansif pour la norme uniforme. En particulier, on a pour tout $y \in \mathbb{R}^n$, $|Mw(x + y) - Mw(x)| \leq |w(x + y) - w(x)|$ ce qui montre que M préserve les constantes de Hölder. ■

Théorème 11.41 *La fonction valeur V est solution au sens de viscosité de l'équation*

$$\max[\lambda V(x) - \mathcal{H}(x, DV(x)), V(x) - MV(x)] = 0, \tag{11.76}$$

au sens où

$$\max[\lambda V(x) - \mathcal{H}(x, p), V(x) - MV(x)] \leq 0 \quad (\text{resp. } \geq 0), \tag{11.77}$$

pour tout $p \in D^+v(x)$ (resp. $p \in D^-v(x)$).

Démonstration. La démonstration se réduit à celle du théorème 3.36, en identifiant la décision d'impulsion à une décision d'arrêt de coût $\varphi(x) := MV(x)$. ■

Pour le résultat d'unicité on se reportera à Barles [10, Section 3.2.2].

[1] C'est à dire lipschitzien de constante 1.

11.6 Notes

La référence classique sur la programmation dynamique est R. Bellman [12]. Une présentation simple, avec de nombreux exemples est donnée dans D. Bertsekas [13].

L'approche par solution de viscosité est dûe à Crandall et Lions [23]. Barles [10] fournit une introduction à ce sujet. Notons aussi l'ouvrage de Bardi et Capuzzo-Dolcetta [9].

Chapitre 12

Résolution numérique de l'équation HJB

Ce chapitre discute la résolution numérique du problème (P_x) du chapitre 3, en discrétisant l'équation HJB. Nous supposerons dans ce chapitre f et ℓ lipschitziennes et bornées, $\lambda > 0$, U compact non vide, et C fermé (supposé vide dans certains énoncés). Introduisons deux espaces de fonctions, l'ensemble $B(\mathbb{R}^n)$ l'ensemble des fonctions bornées $\mathbb{R}^n \to \mathbb{R}^n$, muni de la norme

$$\|v\|_\infty := \sup_{x\in\mathbb{R}^n} |v(x)| \tag{12.1}$$

qui en fait un espace de Banach, (à ne pas confondre avec $\mathcal{L}^\infty(\mathbb{R}^n)$, l'espace des fonction définies presque partout, essentiellement bornées) et l'espace

$$BUC(\mathbb{R}^n) := \{ \text{ Fonctions bornées, uniformément continues : } \mathbb{R}^n \to \mathbb{R}\}\,. \tag{12.2}$$

On pose

$$a_+ := \max(a,0); \quad a_- := \min(a,0). \tag{12.3}$$

12.1 Motivation : problème continu

Le but de cette section est d'analyser une variante du principe de programmation dynamique qui se formule comme un opérateur de point fixe contractant, dit *itération sur les valeurs*. L'algorithme convergent qui en découle n'est pas implémentable sur ordinateur, puisqu'il s'applique au problème continu. Cependant, on obtiendra des algorithmes effectifs après discrétisation de l'espace d'état.

Rappelons l'expression du principe de programmation dynamique (théorème 3.3) : si $x \in \mathbb{R}^n \setminus C$, et $\tau \in]0, T(x)[$, alors

$$V(x) := \inf_{u\in\mathcal{U}} \left\{ \int_0^\tau \ell(y_{x,u}(t), u(t)) e^{-\lambda t} \mathrm{d}t + V(y_{x,u}(\tau)) e^{-\lambda\tau} \right\}. \tag{12.4}$$

Nous allons voir une variante de cette formulation qui permet de définir un opérateur dans tout l'espace. On note $t_1 \wedge t_2 := \min(t_1, t_2)$.

Théorème 12.1 (Principe de Programmation Dynamique II) *Soit $x \in \mathbb{R}^n \setminus C$ et $\tau > 0$. Alors $V(x) = \mathcal{M}^\tau V(x)$, où*

$$\mathcal{M}^\tau v(x) := \inf_{(u,T)} \left\{ \int_0^{\tau \wedge T} \ell(y_{x,u}(t), u(t)) e^{-\lambda t} \mathrm{d}t + \chi_{\tau<T} v(y_{x,u}(\tau)) e^{-\lambda \tau} \right\}, \qquad (12.5)$$

l'infimum portant sur les couples (u, T) admissibles.

Démonstration. La démonstration est similaire à celle du théorème 3.3. ■

Proposition 12.2 *Pour tout $\tau > 0$, l'opérateur $\mathcal{M}^\tau$ est monotone croissant de $B(\mathbb{R}^n)$ dans lui même, et c'est une contraction de rapport $e^{-\lambda\tau}$.*

Démonstration. Sachant que ℓ est borné, et donc

$$\left| \int_0^{\tau \wedge T} \ell(y_{x,u}(t), u(t)) e^{-\lambda t} \mathrm{d}t \right| \leq \lambda^{-1} \|\ell\|_\infty, \qquad (12.6)$$

il est clair que $\mathcal{M}^\tau$ applique $B(\mathbb{R}^n)$ dans lui-même. La monotonie de $\mathcal{M}^\tau$ est immédiate. Soient v et v' deux fonctions de $B(\mathbb{R}^n)$. Par une majoration similaire à (1.27), il vient

$$|\mathcal{M}^\tau v'(x) - \mathcal{M}^\tau v(x)| \leq \sup_{(u,T)} e^{-\lambda\tau} \chi_{\tau<T} \, |v'(y_{x,u}(\tau)) - v(y_{x,u}(\tau))| \leq e^{-\lambda\tau} \|v' - v\|_\infty,$$

d'où la conclusion. ■

On déduit du résultat précédent l'"algorithme" (en espace d'état continu) d'*itérations sur les valeurs* ci-dessous.

Corollaire 12.3 *On peut calculer V par l'algorithme de point fixe suivant : fixer $\tau > 0$, et former la suite $V_{k+1} = \mathcal{M}^\tau V_k$, en partant de $V_0 \in B(\mathbb{R}^n)$ quelconque. Cette suite vérifie*

$$\|V_{k+1} - V\|_\infty \leq e^{-k\lambda\tau} \|V_k - V\|_\infty. \qquad (12.7)$$

Nous allons maintenant formuler des schémas numériques de discrétisation de l'équation HJB. Ces schémas se reformulent comme des points fixes d'opérateurs contractants qui s'interprètent comme des discrétisations de l'opérateur $\mathcal{M}^\tau$.

12.2 Schémas décentrés et extensions

12.2.1 Dimension d'espace $n = 1$

Nous allons discrétiser l'équation HJB en remplaçant la dérivée en espace par une différence finie. Soit $\Delta x > 0$ le pas d'espace. On note $x_j := j\Delta x$. L'espace discret est $\{x_j, j \in \mathbb{Z}\} = \Delta x\mathbb{Z}$.

On pose $\Omega := I\!R^n \setminus C$, et on notera $C_{\Delta x}$ et $\Omega_{\Delta x}$ les discrétisations des ensembles C et Ω, respectivement ; ces deux ensembles forment une partition de $\Delta x\mathbb{Z}$. Nous supposerons que

$$C_{\Delta x} \text{ converge vers } C, \text{ au sens de la } \textit{distance de Hausdorff}. \tag{12.8}$$

On rappelle que, si C_1 et C_2 sont deux parties de $I\!R^n$, leur distance de Hausdorff est

$$\mathrm{dist}(C_1, C_2) := \max\left(\sup_{c_1 \in C_1} \mathrm{dist}(c_1, C_2), \sup_{c_2 \in C_2} \mathrm{dist}(c_2, C_1) \right). \tag{12.9}$$

On désire approcher $V(x_j)$ par la quantité v_j. Notons

$$D^d v_j = \frac{v_{j+1} - v_j}{\Delta x}, \qquad D^g v_j = \frac{v_j - v_{j-1}}{\Delta x}, \qquad D^0 v_j = \frac{v_{j+1} - v_{j-1}}{2\Delta x}, \tag{12.10}$$

les *différences divisées à droite, à gauche et centrées*, respectivement. Laquelle faut-il prendre pour discrétiser l'équation HJB ?

L'idée essentielle est de s'appuyer sur le principe de programmation dynamique, qui relie les valeurs de V en x et en les points voisins dans la direction de $f(x, u)$. *Il convient donc de décentrer à droite si $f(x, u)$ est positive, et à gauche sinon*[1]. On obtient ainsi le *schéma décentré*

$$\left\{ \begin{array}{ll} \lambda v_j & = \inf\limits_{u \in U} \left\{ \ell(x_j, u) + f(x_j, u)_+ \dfrac{v_{j+1} - v_j}{\Delta x} + |f(x_j, u)_-| \dfrac{v_{j-1} - v_j}{\Delta x} \right\}, \quad j \in \Omega_{\Delta x}, \\ v_j & = 0, \quad j \in C_{\Delta x}. \end{array} \right. \tag{12.11}$$

Exemple 12.4 Soit le problème de transfert en temps minimal à 0, avec la dynamique $\dot{x} = u$, et la contrainte $-1 \leq u \leq 1$. Il est naturel de prendre $C_{\Delta x} = \{0\}$. La fonction à minimiser dans (4.11) est linéaire par morceaux : elle atteint son minimum en 0, -1 ou 1. Le schéma décentré s'écrit donc, pour $j \neq 0$,

$$\lambda v_j = 1 + \min\left\{ 0, \frac{v_{j-1} - v_j}{\Delta x}, \frac{v_{j+1} - v_j}{\Delta x} \right\}, \tag{12.12}$$

ou encore

$$(1 + \lambda\Delta x) v_j = \Delta x + \min\left\{ v_{j-1}, v_j, v_{j+1} \right\}, \tag{12.13}$$

On cherche (à l'image du prblème continu) une solution paire, croissante avec $|j|$, donc pour $j \geq 1$, $(1 + \lambda\Delta x)v_j = \Delta x + v_{j-1}$, d'où $(1 + \lambda\Delta x)(v_j - 1/\lambda) = v_{j-1} - 1/\lambda$, et finalement $v_j = \lambda^{-1}(1 - (1 + \lambda\Delta x)^{-|j|})$, pour tout $j \in \mathbb{Z}$.

[1] Ce qui traduit le fait que la prise de décision optimale nécessite d'envisager les conséquences de ses actes.

12.2.2 Forme de point fixe contractant

Nous allons réécrire le schéma (4.11) sous une forme de point fixe contractant, ce qui permettra de vérifier qu'il a une solution unique. Cette réécriture fait apparaître un pas de temps $\Delta t > 0$ *fictif*. Multipliant (4.11) par $\Delta t > 0$, ajoutant v_j à chaque membre, et divisant par $(1 + \lambda\Delta t)$, il vient, pour tout $j \in \Omega_{\Delta x}$:

$$\begin{aligned} v_j = \quad & (1+\lambda\Delta t)^{-1} \inf_{u\in U} \left\{ \Delta t \ell(x_j,u) + \left(1 - \frac{\Delta t}{\Delta x}|f(x_j,u)|\right) v_j \right. \\ & \left. + \frac{\Delta t}{\Delta x}|f(x_j,u)_-| v_{j-1} + \frac{\Delta t}{\Delta x} f(x_j,u)_+ v_{j+1} \right\}. \end{aligned} \tag{12.14}$$

Nous allons vérifier que, pour Δt assez petit, (4.14) est une équation de point fixe monotone et contractant. Notons $N(f)$ la norme infinie de f restreinte à $\mathbb{R}^n \times U$,

$$N(f) := \sup_x \sup_{u\in U} |f(x,u)|, \tag{12.15}$$

et considérons la *condition de stabilité*

$$\frac{\Delta t}{\Delta x} N(f) \le 1. \tag{12.16}$$

Remarque 12.5 Si (4.16) est satisfait, la combinaison linéaire de v_{j-1}, v_j, et v_{j+1} apparaissant dans (4.14) est tout simplement une formule d'*interpolation linéaire* de la valeur de v au point $x_j + \Delta t f(x_j, u)$. Ceci permet d'interpréter (4.14) comme une discrétisation du principe de programmation dynamique (4.5), le pas Δt correspondant à τ.

Proposition 12.6 (i) *Le schéma* (4.11) *possède une solution unique, telle que*

$$\|v\|_\infty \le \lambda^{-1} \|\ell\|_\infty. \tag{12.17}$$

(ii) *Si* Δt *vérifie la condition de stabilité* (4.16)*, alors* (4.14) *est une équation de point fixe contractant (sur l'ensemble des fonctions nulles sur* $C_{\Delta x}$*) pour la norme uniforme*

$$\|v_j\|_\infty := \sup\{|v_j|, \quad j \in \mathbb{Z}\}, \tag{12.18}$$

de rapport de contraction $(1 + \lambda\Delta t)^{-1}$.

Démonstration. Soit $\mathcal{N}^{\Delta t}$ l'opérateur de point fixe du membre de droite de (4.14). Notons

$$\tilde f(x_j, u) := \frac{\Delta t}{\Delta x} f(x_j, u)$$

qui représente une mise à l'échelle de la dynamique. Utilisant (1.27), et le fait que (4.16) implique $1 - |\tilde f(x_j, u)| \ge 0$, il vient

$$\begin{aligned} |(\mathcal{N}^{\Delta t} v')_j - (\mathcal{N}^{\Delta t} v)_j| \le \quad & (1+\lambda\Delta t)^{-1} \sup_{u\in\mathcal{U}} \Big\{ (1 - |\tilde f(x_j,u)|)|v'_j - v_j| \\ & + |\tilde f(x_j,u)_-||v'_{j-1} - v_{j-1}| + \tilde f(x_j,u)_+ |v'_{j+1} - v_{j+1}| \Big\}. \end{aligned} \tag{12.19}$$

Majorant $|v'_i - v_i|$, pour $i = j-1, j, j+1$, par $\|v' - v\|_\infty$ on obtient

$$|(\mathcal{N}^{\Delta t} v')_j - (\mathcal{N}^{\Delta t} v)_j| \leq (1 + \lambda \Delta t)^{-1} \|v' - v\|_\infty \tag{12.20}$$

d'où (ii). L'existence et l'unicité sont conséquence directe de (ii). Enfin soit v la solution de (4.14) ; utilisant (4.14), pour tout $j \in \mathbb{Z}$, il vient

$$|v_j| \leq (1 + \lambda \Delta t)^{-1} \left[\Delta t \|\ell\|_\infty + \|v\|_\infty \right],$$

d'où (4.17). ■

Remarque 12.7 Rien n'empêche de considérer l'opérateur obtenu en prenant dans (4.14) un pas de temps Δt_j dépendant de l'indice d'espace ; cela peut être avantageux d'un point de vue numérique. La condition de stabilité devient

$$\frac{\Delta t_j}{\Delta x} \sup_{u \in U} |f(x_j, u)| \leq 1, \quad \text{pour tout} \quad j \in \mathbb{Z}, \tag{12.21}$$

et le rapport de contraction est $(1 + \lambda \inf_j \Delta t_j)^{-1}$.

Remarque 12.8 (i) On appelle CFL (Courant-Friedrich-Levy) la quantité $\frac{\Delta t}{\Delta x} N(f)$. La condition de stabilité (4.16) dit que le CFL ne doit pas dépasser 1.
(ii) La condition de stabilité assure que, pendant le pas de temps Δt, le système dynamique varie au plus de Δx. Autrement dit, l'information se propage au moins aussi vite dans le schéma numérique que dans le problème d'origine.

Remarque 12.9 Les expressions (4.14) permettent, si la condition de stabilité (4.16) est satisfaite, d'interpréter le schéma décentré comme le principe de programmation dynamique pour le problème de commande optimale d'une chaîne de Markov : voir la remarque 5.19.

Remarque 12.10 Le coefficient de contraction, assurant la convergence de l'algorithme, est $(1 + \lambda \Delta t)^{-1}$. Compte-tenu de la condition de stabilité, on voit que la constante optimale, obtenue pour $CFL = 1$, vaut $(1 + \lambda \Delta x \, N(f)^{-1})^{-1}$. La convergence devient donc très lente quand $\Delta x \downarrow 0$.
D'autres algorithmes sont possibles, en particulier l'itération sur les stratégies (voir la section 5.1).

12.2.3 Dimension d'espace quelconque

Le schéma décentré monodimensionnel peut se généraliser de multiples manières dans le cas où $n > 1$. Donnons seulement la plus naïve.

Soient $h_1, \ldots, h_n$ les pas d'espace, strictement positifs. A $j \in \mathbb{Z}^n$, on associe le point $x_j \in I\!R^n$ de coordonnées $j_i h_i$. Notons $e_1, \ldots, e_n$ la base naturelle de $I\!R^n$. Le

décentrage se fait, pour chaque composante, suivant le signe de $f_i(x_j, u)$; on obtient le schéma suivant :

$$\begin{array}{ll} \lambda v_j = & \inf\limits_{u\in U}\left\{\ell(x_j,u)+\sum\limits_{i=1}^{n}\left(f_i(x_j,u)_+\dfrac{v_{j+e_i}-v_j}{h_i}+|f_i(x_j,u)_-|\dfrac{v_{j-e_i}-v_j}{h_i}\right)\right\}, \\ & j\in\Omega_h; \\ v_j = & 0,\ \ j\in C_h, \end{array} \tag{12.22}$$

où Ω_h et C_h forment une partition de $\mathbb{Z}^n$. Comme dans le cas monodimensionnel, il convient de multiplier (4.11) par un pas de temps fictif qu'on notera h_0, et d'ajouter v_j à chaque membre, ce qui donne

$$\begin{aligned} v_j = \ & (1+\lambda h_0)^{-1}\inf_{u\in U}\left\{h_0\ell(x_j,u)+\left(1-\sum_{i=1}^{n}\frac{h_0}{h_i}|f_i(x_j,u)|\right)v_j\right. \\ & \left.+\sum_{i=1}^{n}\frac{h_0}{h_i}f_i(x_j,u)_+v_{j+e_i}+\sum_{i=1}^{n}\frac{h_0}{h_i}|f_i(x_j,u)_-|v_{j-e_i}\right\}. \end{aligned} \tag{12.23}$$

Proposition 12.11 (i) *Le schéma* (4.22) *possède une solution unique, telle que*

$$\|v\|_\infty \le \lambda^{-1}\|\ell\|_\infty. \tag{12.24}$$

(ii) *Si* h_0 *vérifie la* condition de stabilité

$$h_0\sum_{i=1}^{n}\sup_x\sup_{u\in U}\frac{|f_i(x,u)|}{h_i}\le 1, \tag{12.25}$$

alors (4.23) *est une équation de point fixe contractant (sur l'ensemble des fonctions nulles sur* C_h*) pour la norme uniforme, de rapport de contraction* $(1+\lambda h_0)^{-1}$.

Démonstration. La démonstration est similaire à celle de la proposition 4.6. ■

Exemple 12.12 Soit le système dynamique $\dot{x} = u$, avec $U = [-1, 1]^n$. On a dans ce cas $\sup_x \sup_{u\in U} |f_i(x, u)| = 1$, pour $i = 1, \ldots, n$, et la condition de stabilité se réduit à

$$\frac{1}{h_0}\ge\sum_{i=1}^{n}\frac{1}{h_i}. \tag{12.26}$$

Autrement dit, le pas de temps maximal est dans ce cas la *moyenne harmonique* des pas d'espace.

Remarque 12.13 Le schéma aux différences finies (4.22) fait intervenir le point j et les 2^n points voisins de la grille, obtenus en changeant une seule coordonnée de j de ± 1. Si $n = 2$ on parle d'un schéma à 5 points.

Remarque 12.14 On peut étendre la remarque 4.5 : le schéma, sous la forme (4.23), est très proche du principe de programmation dynamique (4.5). Sous la condition de stabilité (4.25), les poids des $v_{j\pm e_i}$ s'interprètent comme les coordonnées barycentriques du point $x_j + \Delta t f(x_j, u)$.

Remarque 12.15 Si $|f(x,u)|$ peut prendre des valeurs élevées, la condition de stabilité oblige à prendre h_0 très petit. Pour éviter cela, on peut adopter des schémas faisant intervenir des points plus éloignés. Nous en donnons un exemple dans la section suivante.

12.2.4 Discrétisation par triangulation

Donnons maintenant un procédé de discrétisation spatiale qui constitue une alternative intéressante aux méthodes de différences finies. Un *simplexe* de $\mathbb{R}^n$ est un polyèdre formé par l'ensemble des combinaisons convexes de $n+1$ points (appelés sommets) non contenus dans un hyperplan. Autrement dit, un simplexe est de la forme

$$\left\{ \sum_{i=1}^{n+1} \alpha_i x_i; \quad \alpha \geq 0, \quad \sum_{i=1}^{n+1} \alpha_i = 1 \right\},$$

où $x_1, \ldots, x_{n+1}$, sont des points de $\mathbb{R}^n$ non contenus dans un hyperplan. On appelle *face* du simplexe l'ensemble des combinaisons convexes de n des points ; la frontière du simplexe est l'union de ses $n+1$ faces.

Considérons une *triangulation régulière* de $\mathbb{R}^n$ réalisée par une famille de simplexes S_J, $J \in \mathbb{N}$. Autrement dit, l'union de ces simplexes est égale à $\mathbb{R}^n$, et l'intersection de deux simplexes est égale à une face de chacun des deux simplexes. On note $\mathcal{S}$ l'ensemble des simplexes, et $L_{\mathcal{S}}$ l'espace des fonctions linéaires sur chaque simplexes. Les fonctions de $L_{\mathcal{S}}$ sont déterminées par leur valeur aux sommets des simplexe. On a pour une telle fonction v, notant encore h_0 un pas de temps,

$$v(x_i + h_0 f(x_i, u)) = \sum_{j=1}^{n+1} \alpha_i(u) v(x_j), \tag{12.27}$$

où les $\alpha_i(u)$ sont les coefficients de la combinaison convexe représentant le point $x_i + h_0 f(x_i, u)$ dans un des simplexes auquel il appartient (coefficients barycentriques), tels que

$$x_j + h_0 f(x_j, u) = \sum_{j=1}^{n+1} \alpha_i(u) x_j; \quad 0 \leq \alpha_i(u) \leq 1; \quad \sum_{j=1}^{n+1} \alpha_i(u) = 1. \tag{12.28}$$

Un *schéma associé à la triangulation* est obtenu en écrivant une sorte de principe de programmation dynamique discret aux sommets de la triangulation :

$$\left\{ \begin{array}{lcll} v_j & = & (1+\lambda h_0)^{-1} \inf_{u \in U} \left\{ h_0 \ell(x_j, u) + v(x_j + h_0 f(x_j, u)) \right\}, & j \in \Omega_{\mathcal{S}}, \\ v_j & = & 0, & j \in C_{\mathcal{S}}, \end{array} \right. \tag{12.29}$$

où $\Omega_{\mathcal{S}}$ et $C_{\mathcal{S}}$ sont les ensembles de sommets considérés hors de, et dans la cible. L'opérateur $\mathcal{M}^{\mathcal{S}}$ défini par le membre de droite de (4.29) est une contraction de rapport $(1+\lambda h_0)^{-1}$ pour la norme du max, ce qui permet de vérifier que (4.29) a un point fixe unique uniformément borné par $\lambda^{-1}\|\ell\|_\infty$.

Remarque 12.16 Ce schéma permet de raffiner la discrétisation dans une région donnée, et ne comporte pas de condition restrictive sur le pas de temps. En revanche son implémentation est plus complexe ; un point délicat est de reconnaître rapidement dans quel triangle se trouve le point $x_j + h_0 f(x_j, u)$.

12.3 Convergence des schémas et essais numériques

Nous supposons dans l'ensemble de la section que la cible est vide. On sait alors que la valeur est hölderienne (lemme 3.13). Nous donnons deux résultats de convergence des schémas de différences finies. Celui de la section 4.3.1, dont la démonstration est simple, établit la convergence uniforme sur les compacts. Celui de la section 4.3.2 fournit une estimation d'erreur au prix de calculs plus élaborés.

12.3.1 Un argument élémentaire de convergence

L'énoncé ci-dessous se limite au cas $n = 1$, mais la preuve s'étend facilement au schéma aux différences finies pour n quelconque, ainsi qu'à la discrétisation par triangulation. Notons V la fonction valeur, et $v^{\Delta x}$ la solution obtenue pour un pas d'espace Δx. La démonstration utilise de façon essentielle les limites inférieure et supérieure

$$\bar{v}(x) := \limsup_{\substack{j\Delta x \to x \\ \Delta x \downarrow 0}} v_j^{\Delta x}, \qquad \underline{v} := \liminf_{\substack{j\Delta x \to x \\ \Delta x \downarrow 0}} v_j^{\Delta x} \,. \tag{12.30}$$

Théorème 12.17 (Convergence du schéma décentré) *Si $C = \emptyset$, alors :*
(i) *Les fonctions $\bar{v}$ et $\underline{v}$ sont égales à la fonction valeur V du problème "standard" de commande optimale en horizon infini (P_x).*
(ii) *La convergence des valeurs discrètes est uniforme sur tout compact.*

Démonstration. Nous savons que les solutions discrètes sont uniformément bornées par $\lambda^{-1}\|\ell\|_\infty$. Les fonctions $\bar{v}$ et $\underline{v}$ sont donc bornées. Notons bien que ces fonctions sont définies en chaque point, et non presque partout. De la définition de $\bar{v}$ et $\underline{v}$, on déduit aisément que $\bar{v}$ est semi continu supérieurement (s.c.s.), et $\underline{v}$ est semi continu inférieurement (s.c.i.).

La définition de $\bar{v}$ et $\underline{v}$ implique $\bar{v} \geq \underline{v}$. Il suffit alors de montrer que $\bar{v}$ est sous solution, et que $\underline{v}$ est sur solution. En effet, d'après le principe d'unicité forte (théorème 3.31), ceci implique $\bar{v} \leq V \leq \underline{v}$, d'où l'égalité de ces trois fonctions. La convergence des valeurs discrètes uniforme sur tout compact se vérifie alors facilement avec une preuve par l'absurde. On se contentera de montrer que $\bar{v}$ est sous solution, le fait que $\underline{v}$ soit sur solution se démontrant de manière analogue.

Soit x_0 un point de maximum local de $\bar v - \Phi$. Il existe donc $r > 0$ tel que $\bar v - \Phi$ atteint en x_0 son maximum sur $\bar B(x_0, r)$ (la boule fermée de centre x_0 et rayon r). Ajoutant $\|x - x_0\|^2 + \bar v(x_0) - \Phi(x_0)$ à Φ si nécessaire, on peut supposer que

$$\bar v(x_0) = \Phi(x_0) \quad \text{et} \quad \bar v(x) < \Phi(x) \ \text{ si } \ x \neq x_0, \ \ x \in \bar B(x_0, r). \tag{12.31}$$

Par définition de $\bar v(x_0)$, il existe des suites $\Delta x_k \downarrow 0$ et $j_k \in \mathbb{Z}$ telles que

$$j_k \Delta x_k \to x_0 \quad \text{et} \quad \bar v(x_0) = \lim v_{j_k}^{\Delta x_k}.$$

Soit $i_k \in \mathbb{Z}$ tel que $i_k \Delta x_k \in \bar B(x_0, r)$, et tel que

$$v_{j'}^{\Delta x_k} - \Phi(x_{j'}) \le v_{i_k}^{\Delta x_k} - \Phi(i_k \Delta x_k), \ \ j' \neq i_k; \ \ j' \Delta x_k \in \bar B(x_0, r). \tag{12.32}$$

Extrayant si nécessaire une sous suite, on peut supposer que $i_k \Delta x_k \to \bar x$, et nécessairement $|\bar x - x_0| \le r$. Par définition de $\bar v$, on a :

$$\bar v(x_0) - \Phi(x_0) = \lim_k \left(v_{j_k}^{\Delta x_k} - \Phi(j_k \Delta x_k) \right) \le \limsup_k \left(v_{i_k}^{\Delta x_k} - \Phi(i_k \Delta x_k) \right) \le \bar v(\bar x) - \Phi(\bar x).$$

Ceci, joint à (4.31), montre que $\bar x = x_0$ et aussi

$$\bar v(x_0) = \lim_k v_{i_k}^{\Delta x_k}. \tag{12.33}$$

Combinant (4.11) et (4.32), il vient

$$\begin{aligned} \lambda v_{i_k}^{\Delta x_k} \le \ \inf_{u \in U} \ \Big\{ & \ell(x_{i_k}, u) + f(x_{i_k}, u)_+ \frac{\Phi((i_k+1)\Delta x_k) - \Phi(i_k \Delta x_k)}{\Delta x_k} \\ & + |f(x_{i_k}, u)_-| \frac{\Phi((i_k-1)\Delta x_k) - \Phi(i_k \Delta x_k)}{\Delta x_k} \Big\}. \end{aligned}$$

Puisque $i_k \Delta x_k \to \bar x = x_0$, passant à la limite quand $\Delta x_k \downarrow 0$, on obtient avec (4.33) :

$$\lambda \bar v(x_0) + \mathcal{H}(x_0, D\Phi(x_0)) \le 0, \tag{12.34}$$

ce qui prouve que $\bar v$ est sous solution. ■

12.3.2 Estimation d'erreur

Dans cette section on suppose que la cible est vide, et on donne une estimation de l'erreur de discrétisation. On note par $v(\cdot)$ la fonction telle que $v(x_j) = v_j$ où $\{v_j; j \in \mathbb{Z}^n\}$ est la solution du schéma avec les pas $h_1, \ldots, h_n$. On remarquera le lien entre la démonstration ci-dessous et celle du théorème 3.31[2].

[2]La démonstration du théorème étant technique, on pourra l'admettre en première lecture.

Théorème 12.18 *Soit $\gamma \in]0,1[$ une constante de Hölder de V. Alors il existe $C > 0$ tel que, pour tout $(h_1, \dots, h_n) \in (\mathbb{R}_+^*)^n$, on a*

$$|V(x) - v(x)| \leq C \left(\max_{1 \leq i \leq n} h_i \right)^{\gamma/2}, \quad \text{pour tout } x \in \Pi_i(h_i \mathbb{Z}). \tag{12.35}$$

Démonstration. Soit $0 < \varepsilon < 1$; posons

$$\beta_\varepsilon(x) := -\varepsilon^{-2}|x|^2, \quad x \in \mathbb{R}^n. \tag{12.36}$$

ainsi que

$$\varphi(x,y) := v(x) - V(y) + \beta_\varepsilon(x-y), \quad (x,y) \in \mathbb{R}_h^n \times \mathbb{R}^n.$$

Soit $\delta \in (0,1)$, et notons $\mathbb{R}_h^n = \{(j_1 h_1, \dots, j_n h_n); j \in \mathbb{Z}^n\}$. Puisque V et v sont bornées, il existe (x_1, y_1) dans $\mathbb{R}_h^n \times \mathbb{R}^n$ tel que

$$\varphi(x_1, y_1) > \sup \varphi - \delta. \tag{12.37}$$

Soit $\xi \in C^\infty(\mathbb{R}^n \times \mathbb{R}^n)$ à support compact, tel que

$$\xi(x_1, y_1) = 1, \qquad 0 \leq \xi \leq 1, \quad \sup_{x,y} \|D\xi(x,y)\| \leq 1, \tag{12.38}$$

et posons

$$\psi(x,y) = \varphi(x,y) + \delta\xi(x,y), \quad (x,y) \in \mathbb{R}_h^n \times \mathbb{R}^n. \tag{12.39}$$

Alors ψ atteint son maximum sur $\mathbb{R}_h^n \times \mathbb{R}^n$ en un point (x_o, y_o) du support de ξ. Autrement dit,

$$\psi(x_o, y_o) \geq \psi(x,y), \quad \text{pour tout } (x,y) \in \mathbb{R}_h^n \times \mathbb{R}^n. \tag{12.40}$$

En particulier, $y \to -\psi(x_o, y)$ atteint son minimum en y_o. Par définition d'une solution de viscosité, il existe $u^* \in U$ tel que

$$\lambda V(y_o) + f(y_o, u^*) \cdot (D\beta_\varepsilon(x_o - y_o) - \delta D_y \xi(x_o, y_o)) - \ell(y_o, u^*) \geq 0. \tag{12.41}$$

Puisque x_o appartient à $\mathbb{R}_h^n$, il existe $j \in \mathbb{Z}^n$ tel que $x_o = x_j$. On a avec (4.22)

$$\lambda v_j \leq \ell(x_j, u^*) + \sum_{i=1}^n \left(f_i(x_j, u^*)_+ \frac{v_{j+e_i} - v_j}{h_i} + |f_i(x_j, u^*)_-| \frac{v_j - v_{j-e_i}}{h_i} \right). \tag{12.42}$$

Utilisant (4.40) avec $x = x_0 \pm h_i e_i$ et $y = y_o$, nous obtenons

$$\begin{aligned} v_{j \pm e_i} - v_j &\leq \beta_\varepsilon(x_0 - y_0) + \delta\xi(x_0, y_0) \\ &\qquad - \beta_\varepsilon(x_0 \pm h_i e_i - y_0) - \delta\xi(x_0 \pm h_i e_i, y_0) \\ &\leq -\beta_\varepsilon'(x_0 - y_0)(\pm h_i e_i) + \varepsilon^{-2} h_i^2 + \delta h_i. \end{aligned}$$

Multiplions cette inégalité (dans laquelle $\pm = -$) par $f_i(x_j, u^*)_+/h_i$; et (avec $\pm = +$)) par $|f_i(x_j, u^*)_-|/h_i$; ajoutons ces inégalités à (4.42) ; il vient

$$\lambda v_j \leq \ell(x_j, u^*) - \beta'_\varepsilon(x_0 - y_0) f(x_j, u^*) + O\left(\varepsilon^{-2} \max_i h_i + \delta\right). \tag{12.43}$$

Soustrayant (4.41) de l'inégalité précédente, nous obtenons

$$\begin{aligned} \lambda(v_j - V(y_0)) \quad &\leq \quad (\ell(x_0, u^*) - \ell(y_0, u^*)) \\ &\quad +\beta'_\varepsilon(x_0 - y_0)(f(y_0, u^*) - f(x_0, u^*)) + O\left(\varepsilon^{-2} \max_i h_i + \delta\right). \end{aligned}$$

Combinant avec les relations

$$\ell(x_0, u^*) - \ell(y_0, u^*) \quad = \quad O(|x_0 - y_0|), \tag{12.44}$$

$$f(x_0, u^*) - f(y_o, u^*) \quad = \quad O(|x_0 - y_0|), \tag{12.45}$$

et prenant $\delta = O(\max_i h_i)$, il vient pour un certain $C > 0$

$$v(x_o) - V(y_o) \leq C\left[|x_o - y_o| + \frac{|x_o - y_o|^2}{\varepsilon^2} + \frac{\max_i h_i}{\varepsilon^2}\right]. \tag{12.46}$$

De $ab \leq \frac{1}{2}(a^2 + b^2)$ on déduit que

$$|x_o - y_o| = \varepsilon \frac{|x_o - y_o|}{\varepsilon} \leq \tfrac{1}{2}(\varepsilon^2 + \frac{|x_o - y_o|^2}{\varepsilon^2}).$$

Avec (4.46), nous obtenons

$$v(x_o) - V(y_o) \leq 2C\left[\varepsilon^2 + \frac{|x_o - y_o|^2}{\varepsilon^2} + \frac{\max_i h_i}{\varepsilon^2}\right]. \tag{12.47}$$

Or

$$\sup \varphi - \delta \leq v(x_0) - V(y_0) - \frac{|x_o - y_o|^2}{\varepsilon^2},$$

donc

$$\frac{|x_o - y_o|^2}{\varepsilon^2} \leq \sup v - \inf V - \sup \varphi$$

ce qui prouve que $|x_0 - y_0| \to 0$ quand $\varepsilon \downarrow 0$. Prenant $x = y = x_0$ dans (4.40), et utilisant le fait que V est hölderienne de constante γ, il vient

$$\frac{1}{\varepsilon^2}|x_o - y_o|^2 \leq V(x_0) - V(y_0) + \delta|x_o - y_o| \leq K|x_o - y_o|^\gamma,$$

pour un certain K indépendant de ε et h. De là $|x_o - y_o| \leq K\varepsilon^{\frac{2}{2-\gamma}}$, et avec (4.47),

$$v(x_o) - V(y_o) \leq K\left[\varepsilon^{\frac{2\gamma}{2-\gamma}} + \frac{\max_i h_i}{\varepsilon^2}\right].$$

Prenant $\varepsilon = (\max_i h_i)^{(2-\gamma)/4}$, on obtient

$$v(x_o) - V(y_o) \leq K \left(\max_i h_i\right)^{\gamma/2}. \tag{12.48}$$

Enfin puisque $\delta = O(\max_i h_i)$, il vient

$$\sup(v - V) \leq \sup \varphi \leq v(x_0) - V(y_0) + O(\max_i h_i) \leq O\left(\max_i h_i\right)^{\gamma/2} \tag{12.49}$$

d'où l'inégalité recherchée.

L'inégalité inverse se prouve de manière similaire, en maximisant la fonction

$$\varphi(x, y) := V(y) - v(x) + \beta_\varepsilon(x - y), \quad (x, y) \in I\!R_h^n \times I\!R^n.$$

Pour $\delta \in (0, 1)$, on a encore l'existence de (x_1, y_1) dans $I\!R_h^n \times I\!R^n$ satisfaisant (4.37). Définissant ξ et ψ par (4.38) et (4.39) on obtient (4.40). Puisque x_o appartient à $I\!R_h^n$, il existe $j \in \mathbb{Z}^n$ tel que $x_o = x_j$. On poursuit de la même manière en faisant intervenir la commande $u^* \in U$ réalisant le minimum dans l'expression (4.22) du schéma. ■

Remarque 12.19 Si λ est assez grand, une variante de la démonstration du lemme 3.13 permet de montrer que V est lipschitzien. Dans ce cas on a une estimation d'erreur sur V de l'ordre de $O((\max_i h_i)^{1/2})$.

Remarque 12.20 On trouvera la discussion d'autres schémas numériques dans l'annexe du livre [9], due à M. Falcone.

12.3.3 Equation eikonale

On considère le système dynamique suivant :

$$\dot{x} = F(x)u, \tag{12.50}$$

où $F : I\!R^n \to I\!R_+$ représente la vitesse du milieu. La commande u doit rester dans la boule unité pour la norme euclidienne. Pour $\lambda = 0$, l'équation HJB associée, dite équation eikonale, est de la forme

$$\begin{cases} 1 - F(x)\|Dv(x)\| = 0 & \text{dans} \quad \Omega, \\ V(x) = 0, \quad x \in C. \end{cases} \tag{12.51}$$

Dans l'exemple numérique, on a pris $C = \{0\}$, et $F(x) = 1$ pour tout x, de sorte que le temps de transfert est la distance euclidienne à 0.

Sur la figure 4.1, on a représenté les lignes de niveau de la différence entre solution calculée et valeur exacte, en limitant le domaine à $[0, 0.1] \times [0, 0.1]$. La grille est de taille 25×25, et on a effectué 100 itérations sur les valeurs avec $\lambda = 0$. Comme on peut s'y attendre, on observe que les erreurs sont plus importantes dans les coins, en raison de l'accumulation d'erreurs inhérente à l'algorithme.

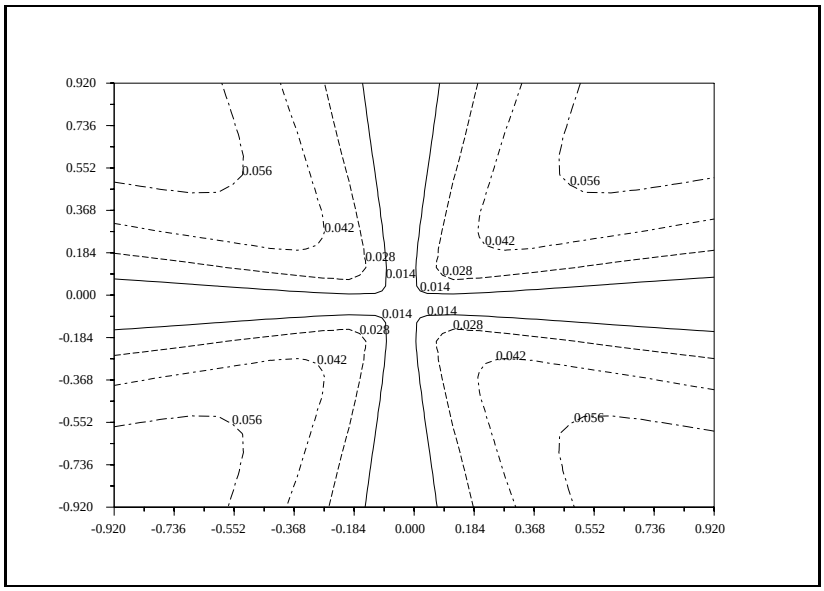

FIG. 12.1 – Equation eikonale : erreur sur le temps minimal

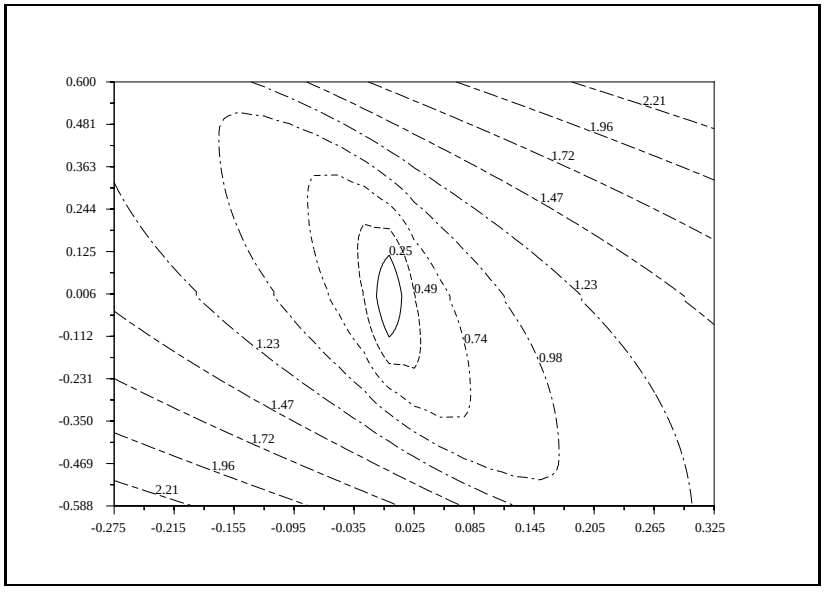

FIG. 12.2 – Problème d'alunissage : isovaleurs du temps minimal

12.3.4 Problème d'alunissage

Nous reprenons le problème d'alunissage discuté en section 1.2. Le problème discret est résolu sur le domaine $(z, \dot{z}) \in [-1, 1] \times [-2, 2]$. On prend 80 points de discrétisation pour z et on impose $\Delta t = \Delta x$. La condition de stabilité impose alors de prendre 320 points de discrétisation pour la vitesse. On fixe une condition aux limites artificielle égale à 100 sur le bord.

Les isovaleurs du temps minimal sont représentées en figure 4.2. La figure ne reprend que la partie centrale du domaine, pour éviter les effets dus au caractère borné du domaine.

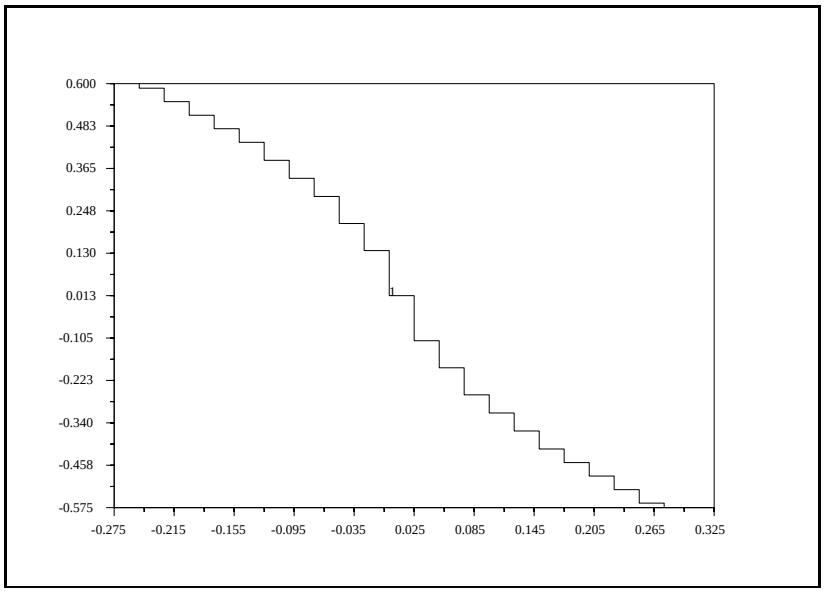

FIG. 12.3 – Problème d'alunissage : lieu de changement de signe

La figure 4.3 représente le lieu de changement de signe de l'estimation numérique de $\partial V/\partial \dot{z}$, qui en raison du théorème 3.14 détermine la statégie de feedback. Elle se relie bien aux isovaleurs de la figure 4.3. On la comparera à la figure 1.1 qui donne le lieu de changement de signe de la commande optimale.

12.4 Notes

La démonstration de convergence du théorème 4.17 reprend G. Barles and P. E. Souganidis [11]. L'estimation d'erreur du théorème 4.18 suit Capuzzo-Dolcetta et Ishii [26]. Une estimation analogue, dans le cas parabolique, se trouve dans M. G. Crandall and P.-L. Lions [22].

Chapitre 13

Commande optimale stochastique

13.1 Chaînes de Markov commandées

13.1.1 Quelques exemples

Un exemple classique de commande de chaînes de Markov est la *gestion de stock* : les achats des clients arrivent de manière aléatoire, et la commande consiste à réapprovisionner, avec paiement de pénalités pour tout achat non honoré. Autre exemple, la maintenance d'un parc d'outils de production. L'état du système est l'ensemble des outils en état de fonctionnement, et la commande consiste à effectuer les réparations des outils en panne.

Enfin les problèmes de commande optimale (déterministes ou stochastiques) en espace continu (et temps continu ou discret) résolus en discrétisant l'équation HJB reviennent, comme on le verra, à résoudre un problème de commande d'une chaînes de Markov. En particulier, les problèmes d'évaluation d'options financière, d'identification de volatilité implicite, et de gestion de portefeuille sont de cette nature.

13.1.2 Chaînes de Markov et valeurs associées

Considérons un *système dynamique* dont l'état peut prendre un nombre fini ou dénombrable de valeurs, soit $1, \ldots, m$, avec m fini ou non. Il est utile de traiter le cas $m = \infty$ pour discuter le problème de discrétisation de systèmes continus.

On note x^k la valeur de l'état au temps k, où $k \in I\!N$. On suppose connue la probabilité M_{ij}^k de transition de l'état i au temps k, à l'état j au temps $k+1$. Autrement dit, notant $\mathcal{P}$ la loi de probabilité, on a

$$\mathcal{P}(x^{k+1} = j | x^k = i) = M_{ij}^k. \tag{13.1}$$

On supposera cette loi *markovienne*, c'est à dire

$$\mathcal{P}(x^{k+1} = j | x^k = i, x^{k-1} = i_{k-1}, \ldots, x^0 = i_0) = M^k_{ij}. \tag{13.2}$$

Ceci signifie que si on connaît la valeur de l'état au temps k, la connaissance des états passés n'apporte rien pour la prédiction du futur.

La "matrice" $M^k = \{M^k_{ij}\}$, où i et j varient de 1 à m, est le tableau (fini ou non) de valeur M^k_{ij} en ligne i et colonne j. Tous ses éléments sont positifs ou nuls, et la somme des éléments d'une ligne vaut 1. Une telle matrice est dite *stochastique*.

Si $m = \infty$, l'extension naturelle du calcul matriciel : produit de deux matrices, produit d'une matrice avec un vecteur (vertical) à droite ou (horizontal) à gauche, et produit de deux matrices, demande quelques précautions : il faut que les quantités en jeu soient sommables. Plus précisément, soient ℓ^1 et ℓ^∞, respectivement, l'espace des suites sommables et bornées, dont les éléments sont indicés de 1 à m, et représentés comme des vecteurs horizontaux (pour ℓ^1) et verticaux (pour ℓ^∞). Si $x \in \ell^1$ et $v \in \ell^\infty$, et si M est une matrice stochastique, on peut définir leur produit $xM \in \ell^1$ et $Mv \in \ell^\infty$ par

$$(xM)_j := \sum_{i=1}^m x_i M_{ij}; \quad (Mv)_i := \sum_{j=1}^m M_{ij} v_j.$$

On a en effet $\|xM\|_1 \leq \|x\|_1$ et $\|Mv\|_\infty \leq \|v\|_\infty$. Si M^1 et M^2 sont deux matrices stochastiques, on peut définir leur produit $M^1 M^2$ par

$$(M^1 M^2)_{ij} := \sum_{k=1}^m M^1_{ik} M^2_{kj}.$$

Il est facile de vérifier que le produit de deux matrices stochastiques est une matrice stochastique. On interprètera

$$\left\{ p \in \ell^1; \quad p_i \geq 0, \;\; i = 1, \ldots, m; \quad \sum_{i=1}^m p_i = 1 \right\}$$

comme l'*espace des probabilités* pour l'état du système à un temps donné, et ℓ^∞ comme un *espace de valeurs*. Cette dernière terminologie sera plus claire dans la suite.

Si l'état x^k du système à l'instant k est connu, la loi de probabilité de x^{k+1} est la ligne de M^k d'indice x^k. Si on dispose seulement d'une loi de probabilité pour x^k, notée $p^k = (p^k_1, \ldots, p^k_m)$, et considérée comme un vecteur horizontal, alors la loi de probabilité de x^{k+1} vérifie l'*équation de Kolmogorov avant*

$$p^{k+1} := \mathcal{P}(x^{k+1} | p^k) = \sum_i p^k_i M^k_{i,\cdot} = p^k M^k, \tag{13.3}$$

d'où on déduit par récurrence, si la probabilité initiale est p^0,

$$\mathcal{P}(x^{k+1} | p^0) = p^0 M^0 M^1 \ldots M^k. \tag{13.4}$$

Associons maintenant à ce processus la *fonction coût* $\{c_i^k\}$, $i = 1, \ldots, m$, $k \in \mathbb{N}$. On suppose que $c^k := \{c_i^k\}_{i=1,\ldots,m}$ appartient à ℓ^∞, ce qui veut dire que les coûts sont uniformément bornés en espace, et que c^k est représenté comme un vecteur vertical. Soit φ une application $\{1, \ldots, m\} \to \ell^\infty$, appelée coût final. Définissons la *fonction valeur* du problème avec état initial i et instant initial k comme

$$V_i^k := \mathbb{E}\left(\sum_{\ell=k}^{N-1} c_{x^\ell}^\ell + \varphi(x^N) \mid x^k = i\right). \tag{13.5}$$

Ici $N > 0$ est l'*horizon*, et $\mathbb{E}$ représente l'*espérance mathématique*.

Proposition 13.1 *Pour tout $k = 0, \ldots, N$, la fonction valeur V^k est bien définie et appartient à ℓ^∞. De plus, la suite $\{V^k\}$ est solution de l'équation de récurrence de Kolmogorov arrière*

$$\begin{cases} V^k = c^k + M^k V^{k+1}, \quad k = 0, \ldots, N-1, \\ V^N = \varphi. \end{cases} \tag{13.6}$$

Démonstration. Si x^k a la valeur i, alors d'après l'équation de Kolmogorov avant

$$V_i^k = c_i^k + \sum_{j=1}^m M_{ij}^k V_j^{k+1},$$

d'où le résultat. ■

Considérons maintenant un problème avec $c^k = c \in \ell^\infty$ et $M^k = M$ indépendants du temps, *horizon infini*, et taux d'actualisation $\beta \in]0, 1[$. La valeur de ce problème, c'est à dire

$$V_i := \mathbb{E}\left(\sum_{k=0}^\infty \beta^{k+1} c_{x^k} | x^0 = i\right), \tag{13.7}$$

est bien définie et appartient à ℓ^∞. En raison de l'équation de Kolmogorov avant, elle est solution de l'équation

$$V = \beta(c + MV). \tag{13.8}$$

Comme M est lipschitzienne de constante 1, cette équation est celle d'un opérateur de point fixe strictement contractant et a donc une solution unique.

13.1.3 Quelques lemmes

Commençons par le rappel du théorème de point fixe de Banach-Picard.

Lemme 13.2 *Soient X un espace de Banach et C une partie fermée de X. Soit T un opérateur contractant de C vers lui même. Autrement dit, il existe $c \in [0, 1[$ tel que, si $x^i \in C$, $i = 1, 2$, alors $Tx^i \in C$, $i = 1, 2$, et*

$$\|Tx^2 - Tx^1\| \leq c\|x^2 - x^1\|. \tag{13.9}$$

Alors T a un unique point fixe $x^* \in C$ (c.a.d. l'équation $Tx = x$ a pour solution unique x^*). De plus, quel que soit $x^0 \in C$, la suite $\{x^k\}$ telle que $x^{k+1} = Tx^k$ converge vers x^*, et

$$\|x^k - x^*\| \leq c^k \|x^0 - x^*\|. \tag{13.10}$$

Voici un autre lemme, qui sera utile à plusieurs reprises.

Lemme 13.3 *Soit M une matrice stochastique, $\beta \in]0,1[$, $\varepsilon > 0$ et $w \in \ell^\infty$ tels que $w \leq \varepsilon \mathbf{1} + \beta M w$. Alors $w \leq (1-\beta)^{-1}\varepsilon \mathbf{1}$.*

Démonstration. On a $Mw \leq (\sup w)\mathbf{1}$ puisque M est une matrice stochastique, et donc $w \leq (\varepsilon + \beta \sup w)\mathbf{1}$. En conséquence, $\sup w \leq \varepsilon + \beta \sup w$, d'où la conclusion. ■

13.1.4 Principe de Programmation dynamique

Considérons maintenant une chaîne de Markov dont les probabilités de transition $M_{ij}(u)$ dépendent d'une variable de commande $u \in U_i$, où U_i est un ensemble quelconque dépendant de l'état i (certains résultats supposeront U_i *métrique compact*). Donnons nous des coûts dépendant de u et de l'état, soit $c_i^k(u) : U_i \to I\!R$, telle que

$$\sup_{k,i,u} |c_i^k(u)| < \infty. \tag{13.11}$$

On considère le problème de minimisation du critère sur horizon fini

$$V_i^k(u) := I\!E\left(\sum_{\ell=k}^{N-1} c_{x^\ell}^{\ell}(u^\ell) + \varphi(x^N) | x^k = i\right). \tag{13.12}$$

Ici u^k est la valeur de la commande au temps k ; pour donner un sens à ce problème, il faut spécifier l'information dont on dispose au temps k pour choisir la valeur de u^k. Nous allons nous limiter au cas de l'*observation complète*, dans lequel l'état x^k est connu. Ceci permet de choisir u^k fonction de l'état x^k, et bien sûr du temps k. Autrement dit, on choisit une stratégie de *retour d'état* (feedback). Posons

$$\mathcal{U} := \Pi_i U_i. \tag{13.13}$$

On notera u_i la commande adoptée (au temps k) par la stratégie de retour d'état $u \in \mathcal{U}$ si l'état est i, et $M(u)$ la "matrice" de terme générique $M_{ij}(u_i)$. On considère donc le problème de calcul d'un retour d'état optimal

$$V_i^k := \inf_{u \in \mathcal{U}} V_i^k(u), \quad i = 1, \ldots, m. \quad k = 1, \ldots, N. \tag{13.14}$$

Proposition 13.4 *La fonction valeur V^k, solution du problème* (5.14) *avec observation complète, est solution du principe de programmation dynamique*

$$\left\{ \begin{array}{l} V_i^k = \inf\limits_{u\in U_i} \left\{ c_i^k(u) + \sum\limits_j M_{ij}^k(u) V_j^{k+1} \right\}, \quad i = 1,\dots,m, \quad k = 0,\dots,N-1, \\ V^N = \varphi. \end{array} \right. \tag{13.15}$$

De plus, l'ensemble $\bar{U}_i^k$ (éventuellement vide) des commandes optimales à l'instant k lorsque $x^k = i$ est

$$\bar{U}_i^k = \operatorname*{argmin}_{u\in U_i} \left\{ c_i^k(u) + \sum_j M_{ij}^k(u) V_j^{k+1} \right\}. \tag{13.16}$$

Démonstration. On raisonne par récurrence. Il est clair que $V^N = \varphi$. Fixons $k < N$ et $i \in \{1,\dots,m\}$. Si $x^k = i$, d'après l'équation de Kolmogorov arrière, le choix de la commande u à l'instant k donne la valeur $c_i^k(u) + \sum_j M_{ij}^k(u) V_j^{k+1}$. On obtient donc V_i^k en prenant l'infimum de cette quantité, et une commande est optimale si elle appartient à l'argument du minimum. De plus la quantité

$$\|V^k\|_\infty \le \sup_u \|c^k(u)\| + \|V^{k+1}\|_\infty$$

est bien bornée. ■

13.1.5 Problèmes à horizon infini

Dans cette section, nous supposons la fonction coût $c(u)$ indépendante du temps, de même que la matrice de transition. Nous supposons aussi le coût actualisé avec un coefficient $\beta \in]0,1[$. Le théorème suivant caractérise les stratégies optimales, et montre en particulier qu'on peut se limiter aux stratégies de retour d'état stationnaires (la commande ne dépend que de l'état mais pas du temps).

Théorème 13.5 (i) *Dans le cas de l'observation complète, la fonction valeur définie par*

$$V_i := \inf_{u\in\mathcal{U}} I\!E \left\{ \sum_{k=0}^{\infty} \beta^{k+1} c_{x^k}(u_k) | x^0 = i \right\}, \quad i = 1,\dots,m, \tag{13.17}$$

où $\beta \in]0,1[$, est solution unique de l'équation de programmation dynamique : trouver $v \in I\!R^m$ tel que

$$v_i = \beta \inf_{u\in U_i} \left\{ c_i(u) + \sum_j M_{ij}(u) v_j \right\}, \quad i = 1,\dots,m. \tag{13.18}$$

(ii) *Soit $\varepsilon \geq 0$ et $u \in \mathcal{U}$ une stratégie et v la valeur associée, solution de $v = \beta(c(u) + M(u)V)$. Supposons que, pour tout i,*

$$v_i \leq \beta \inf_{\tilde{u} \in U_i} \left(c_i(\tilde{u}) + \sum_j M_{ij}(\tilde{u}) v_j \right) + \varepsilon. \tag{13.19}$$

Posons $\varepsilon' := (1-\beta)^{-1}\varepsilon$. Alors la stratégie u est ε' sous optimale, dans le sens où la valeur associée satisfait

$$v \leq V + \varepsilon' \mathbf{1}. \tag{13.20}$$

(iii) *Supposons, pour tout i et j, U_i* métrique compact *et les fonctions $c_i(u)$ et $M_{ij}(u)$ continues. Alors il existe (au moins) une stratégie optimale.*

Démonstration. a) Montrons d'abord que (5.18) possède une solution unique. Cette équation est de la forme $v = Tv$, avec

$$(Tw)_i := \beta \inf_{u \in U_i} \left\{ c_i(u) + \sum_j M_{ij}(u) w \right\}. \tag{13.21}$$

Comme $\|Tw\|_\infty \leq \beta(\|c\|_\infty + \|w\|_\infty)$, T est un opérateur de ℓ^∞ dans lui même. Avec (1.27) et étant donnés w et w' dans ℓ^∞, utilisant le fait que la somme des éléments d'une ligne de $M(u)$ vaut 1, il vient :

$$|(Tw')_i - (Tw)_i| \leq \beta \sup_{u \in U_i} \sum_{j=1}^m |M_{ij}(u)(w'-w)_j| \leq \beta \|w'-w\|_\infty.$$

En conséquence, T est une contraction de rapport β dans ℓ^∞. Il découle alors du lemme 5.2 que l'équation (5.18) a une solution unique notée v^*.
b) Soit $u \in \mathcal{U}$ une stratégie et v la valeur associée. Utilisant $v^* \leq \beta(c(u) + M(u)v^*)$, il vient $v^* - v \leq \beta M(u)(v^* - v)$. Le lemme 5.3 assure que $v^* \leq v$. Comme ceci est vrai pour toute stratégie, on a aussi $v^* \leq V$.
c) Si (5.19) est satisfait, utilisant (1.27) il vient

$$v_i - v_i^* \leq \varepsilon + \sup_{\tilde{u} \in U_i} \beta M_{ij}(\tilde{u})(v_j - v_j^*) \leq \varepsilon + \beta \sup(v - v^*). \tag{13.22}$$

Passant au supremum en i, on déduit $\sup(v - v^*) \leq \varepsilon'$. Comme $v^* \leq V$ on en déduit (5.20), d'où (ii).
d) Montrons que $V \leq v^*$, et donc $V = v^*$. Soient w^k une suite de valeurs satisfaisant $w^{k+1} = Tw^k$, et $\varepsilon > 0$. Il existe une suite u^k de stratégies telle que $w^{k+1} \geq c(u^k) + M(u^k)w^k - \frac{1}{2}\varepsilon\mathbf{1}$. Comme w^k converge uniformément vers v^*, après un certain rang, $v^* \geq c(u^k) + M(u^k)v^* - \varepsilon\mathbf{1}$. La valeur v^k associé à la stratégie u^k vérifiant $v^k = c(u^k) + M(u^k)v^k$, on a $v^k - v^* \leq M(u^k)(v^k - v^*) + \varepsilon\mathbf{1}$. Le lemme 5.3 implique

$\sup(v^k - v^*) \leq \varepsilon'$, donc $\sup(V - v^*) \leq \limsup_k \sup(v^k - v^*) \leq \varepsilon'$. Ceci étant vrai pour tout $\varepsilon > 0$, on obtient $V \leq v^*$.
e) Montrons (iii). D'après le point (ii), l'existence d'une stratégie optimale équivaut à la possibilité d'atteindre, pour tout état i, l'infimum dans (5.18). Pour i fixé, notons $\{u^q\}$ une suite minimisante. Puisque U_i est métrique compact pour tout i, extrayant une sous-suite si nécessaire, on peut supposer que la suite converge vers $\bar{u} \in U_i$. A tout $\varepsilon \in]0,1[$, on peut associer une partition (I,J) de $\{1,\dots,m\}$, telle que I est de cardinal fini et $\sum_{i\in I} M_{ij}(\bar{u}) \geq 1 - \frac{1}{2}\varepsilon$. Puisque I est fini, pour q assez grand, on a $\sum_{i\in I} M_{ij}(u^q) \geq 1 - \varepsilon$, et donc $\sum_{i\in J} M_{ij}(u^q) \leq \varepsilon$. De là

$$\begin{aligned}
\Delta &:= \left| \limsup_q (c_i(u_i^q) + \sum_j M_{ij}(u_i^q)V_j - c_i(\bar{u}_i) - \sum_j M_{ij}(\bar{u}_i)V_j) \right| \\
&= \left| \limsup_q \sum_{j\in J} (M_{ij}(u^q)V_j - M_{ij}(\bar{u})V_j) \right| \\
&\leq \limsup_q \sum_{j\in J} |M_{ij}(u^q) - M_{ij}(\bar{u})| \|V\|_\infty \leq 2\varepsilon \|V\|_\infty.
\end{aligned}$$

Ceci prouve que $(c(\bar{u}) + M(\bar{u})V)_i = \inf_{u\in U} (c(u) + M(u)V)_i$, d'où (iii).

■

13.1.6 Algorithmes numériques

Dans le cas de problèmes avec horizon infini, on peut mettre en œuvre un algorithme itératif de calcul de v à partir du principe de programmation dynamique. La méthode la plus simple est l'*itérations sur les valeurs*

$$v_i^{q+1} = \beta \inf_{u\in\mathcal{U}} \left\{ c_i(u) + \sum_j M_{ij}(u) v_j^q \right\}, \quad i = 1,\dots,m, \quad q \in I\!N. \tag{13.23}$$

Ici v^q (à ne pas confondre avec la notation v^k employée dans le cas de l'horizon fini) représente la suite formée par l'algorithme.

Proposition 13.6 *L'algorithme d'itération sur les valeurs converge vers la solution unique v^* de* (5.18)*, et on a*

$$\|v^q - v^*\|_\infty \leq \beta^q \|v^0 - v^*\|_\infty. \tag{13.24}$$

Démonstration. Soit T l'opérateur construit en (5.21). Nous avons montré (démonstration du théorème 5.5) que T est contractant de rapport β dans la norme du max. L'algorithme d'itération sur les valeurs s'écrit $v^q = Tv^{q-1}$. On conclut avec le lemme 5.2. ■

Dans le cas assez fréquent où β est proche de 1, l'algorithme d'itération sur les valeurs peut être très lent. Une alternative intéressante est l'algorithme d'*itérations sur les stratégies*, ou *algorithme de Howard*. On fera l'hypothèse suivante :

$$\left\{ \begin{array}{l} \text{U est métrique compact} \\ \text{Les fonctions } c_i(u) \text{ et } M_{ij}(u) \text{ sont continues pour tout } i \text{ et } j. \end{array} \right. \tag{13.25}$$

Chaque itération de l'algorithme comporte deux étapes :

– Etant donné une stratégie $u^q \in \mathcal{U}$, calculer la valeur v^q associée, solution de l'équation linéaire

$$v^q = \beta(c(u^q) + M(u^q)v^q). \tag{13.26}$$

– Calculer u^{q+1} solution de

$$u_i^{q+1} \in \arg\min_{u \in U_i} \left\{ c_i(u) + \sum_j M_{ij}(u) v_j^q \right\}, \quad i = 1, \ldots, m. \tag{13.27}$$

Proposition 13.7 *On suppose* (5.25). *Alors l'algorithme d'itérations sur les politiques, initialisé avec une stratégie $u^0 \in \mathcal{U}$ quelconque, a les propriétés suivantes :*
(i) *Il est bien défini,*
(ii) *La suite v^q décroît,*
(iii) *Elle vérifie $\|v^{q+1} - v^*\| \leq \beta \|v^q - v^*\|$, où v^* est la fonction valeur, unique solution du principe de programmation dynamique* (5.18).

Démonstration. (i) Vérifions que l'algorithme est bien défini. Le système linéaire (5.26) a une solution unique, car c'est l'équation de point fixe d'un opérateur contractant (lemme 5.2). Utilisant les arguments de la démonstration du théorème 5.5, on vérifie que le minimum dans la seconde étape est atteint en raison de (5.25).

Par ailleurs, la suite v^q est bornée dans ℓ^∞ car la relation

$$\|v^q\|_\infty \leq \beta(\|c(u^q)\|_\infty + \|M(u^q)v^q\|_\infty) \leq \beta(\|c(u^q)\|_\infty + \|v^q\|_\infty)$$

donne l'estimation

$$\|v^q\|_\infty \leq (1-\beta)^{-1}\beta\|c\|_\infty. \tag{13.28}$$

(ii) Les relations (5.26) et (5.27) impliquent

$$\begin{array}{rcl} \beta^{-1}(v^{q+1} - v^q) & = & c(u^{q+1}) + M(u^{q+1})v^{q+1} - c(u^q) - M(u^q)v^q, \\ & \leq & c(u^{q+1}) + M(u^{q+1})v^{q+1} - c(u^{q+1}) - M(u^{q+1})v^q, \\ & = & M(u^{q+1})(v^{q+1} - v^q), \end{array}$$

et donc $v^{q+1} - v^q \leq 0$ d'après le lemme 5.3.

(iii) Notons $\bar{v}^{q+1}$ la valeur calculée à partir de v^q, par l'itération sur les valeurs. On sait que $\|\bar{v}^{q+1} - v^*\| \leq \beta\|v^q - v^*\|$. Puisque $v^* \leq v^{q+1}$, il suffit d'établir que $v^{q+1} \leq \bar{v}^{q+1}$. Or

$$\begin{array}{rcl} \beta^{-1}(v^{q+1} - \bar{v}^{q+1}) & = & c(u^{q+1}) + M(u^{q+1})v^{q+1} - (c(u^{q+1}) - M(u^{q+1})v^q), \\ & = & M(u^{q+1})(v^{q+1} - v^q). \end{array}$$

D'après le point (ii), $v^{q+1} \leq v^q$; donc $v^{q+1} \leq \bar{v}^{q+1}$. ■

Remarque 13.8 La démonstration précédente montre que l'itération sur les stratégies converge au moins aussi vite que l'itération sur les valeurs.

13.1.7 Problèmes de temps de sortie

Soit Ω une partie de $\{1, \ldots, m\}$, et considérons une chaîne de Markov (sans commande) de matrice de transition M. Soit τ le premier instant de sortie de Ω :

$$\tau := \min\{k \in I\!N;\ x^k \notin \Omega\}. \tag{13.29}$$

Bien entendu, τ est une variable aléatoire. On considère la fonction valeur, où $i \in \{1, \ldots, m\}$:

$$V_i := I\!E\left(\sum_{k=0}^{\tau-1} \beta^{k+1} c_{x^k} + \beta^\tau \varphi_{x^\tau} | x^0 = i\right). \tag{13.30}$$

Proposition 13.9 *On suppose c et φ dans ℓ^∞. Alors l'espérance ci-dessus est bien définie, la fonction valeur du problème de temps de sortie appartient aussi à ℓ^∞, et est solution unique de l'équation*

$$\begin{cases} v_i = \beta\left(c_i + \sum_j M_{ij} v_j\right), & i \in \Omega, \\ v_i = \varphi_i, & i \notin \Omega. \end{cases} \tag{13.31}$$

Démonstration. Elle est similaire à celle des propositions précédentes. ■

Considérons maintenant le cas de la chaîne de Markov commandée de probabilité de transition $M_{ij}(u)$, avec $u \in U_i$, ensemble métrique compact, et les fonctions $c_i(u)$ et $M_{ij}(u)$ continues. On considère le problème de minimisation du critère avec temps de sortie

$$V_i := \inf_{u \in \mathcal{U}} I\!E\left\{\sum_{k=0}^{\tau-1} \beta^{k+1} c(u)_{x^k} + \beta^\tau \varphi_{x^\tau} | x^0 = i\right\}, \tag{13.32}$$

dans le cas de l'observation complète.

Remarque 13.10 Si c est le vecteur de coordonnées toutes égales à 1, et si φ est nul, alors le critère s'interprète comme une mesure du temps de sortie. Le problème est alors dit à temps minimal.

Proposition 13.11 *On suppose* $\sup_{u\in U}|c_i(u)|$ *fini et* φ *borné. Alors la fonction valeur du problème avec temps de sortie est solution unique de l'équation de la programmation dynamique*

$$\begin{cases} v_i = \beta \inf\limits_{u\in U_i}\left\{c_i(u)+\sum\limits_j M_{ij}(u)v_j\right\}, & i\in\Omega, \\ v_i = \varphi_i, & i\notin\Omega. \end{cases} \tag{13.33}$$

Démonstration. Elle est similaire à celle des propositions précédentes. ■

L'extension des algorithmes d'itérations sur les valeurs et sur les stratégies à la situation étudiée ici ne présente pas de difficulté.

13.1.8 Problèmes avec décision d'arrêt

Nous étudions un problème de commande similaire à celui de la sous-section précédente, ajoutant la possibilité d'arrêt à tout instant, avec un coût d'arrêt $\psi\in\mathbb{R}^m$.

Soit Ω une partie de $\{1,\dots,m\}$, et soient une chaîne de Markov commandée de matrice de transition $M_{ij}(u)$, avec $u\in U$, ensemble métrique compact, et les fonctions $c(u)$ et $M_{ij}(u)$ continues. On note τ le premier instant de sortie de Ω, et θ l'instant de décision d'arrêt. Posons

$$\chi_{\theta<\tau} = \begin{cases} 1 & \text{si } \theta<\tau, \\ 0 & \text{sinon }, \end{cases}$$

et adoptons une convention similaire pour $\chi_{\theta\geq\tau}$. On considère le problème de minimisation du critère avec temps d'arrêt

$$V_i := \inf_{u\in\mathcal{U}} \mathbb{E}\left\{\sum_{k=0}^{\theta\wedge\tau-1}\beta^{k+1}c(u)_{x^k}+\beta^\theta\chi_{\theta<\tau}\psi_{x^\theta}+\beta^\tau\chi_{\theta\geq\tau}\varphi_{x^\tau}|x^0=i\right\}, \tag{13.34}$$

dans le cas de l'observation complète.

Remarque 13.12 Le cadre de cette section recouvre plusieurs situations intéressantes : (i) ensemble Ω égal à l'espace d'état, (ii) U_i réduit à un point pour tout i : la seule décision est d'arrêter ou non, (iii) stratégie optimale pouvant être de ne jamais arrêter le jeu.

Théorème 13.13 *On suppose* $\sup_{u\in U}|c_i(u)|$ *fini, et* ψ *et* φ *bornées. Alors la fonction valeur* V *du problème de temps d'arrêt est solution unique du système*

$$\begin{cases} \text{(i)} & v_i = \min\left(\beta\inf\limits_{u\in U_i}\left\{c_i(u)+\sum\limits_j M_{ij}(u)v_j\right\},\psi_i\right), & i\in\Omega, \\ \text{(ii)} & v_i=\varphi_i, & i\notin\Omega. \end{cases} \tag{13.35}$$

Démonstration. La démonstration est similaire à celle des sections précédentes ; contentons-nous de démontrer que l'équation 5.35 a une solution unique v^*. Définissons l'opérateur T de $\mathbb{R}^m$ dans lui même par

$$\begin{cases} (Tv)_i = \min\left(\beta \inf\limits_{u\in U_i}\left\{c_i(u) + \sum\limits_j M_{ij}(u)v_j\right\}, \psi_i\right), & i \in \Omega, \\ (Tv)_i = \varphi_i, & i \notin \Omega, \end{cases} \tag{13.36}$$

alors pour la norme du max, T est une contraction stricte de rapport β, et a donc un unique point fixe v^*. Ceci établit l'existence et l'unicité de la solution de (5.35). ■

Les arguments qui précèdent assurent la convergence de l'*algorithme d'itérations sur les valeurs*, qui s'écrit, en reprenant les notations de (5.36),

$$v^{q+1} = T(v^q), \tag{13.37}$$

ou encore

$$\begin{cases} v_i^{q+1} = \min\left(\beta \inf\limits_{u\in U_i}\left\{c_i(u) + \sum\limits_j M_{ij}(u)v_j^q\right\}, \psi_i\right), & i \in \Omega, \\ v_i^{q+1} = \varphi_i, & i \notin \Omega. \end{cases} \tag{13.38}$$

En ce qui concerne l'*algorithme d'itérations sur les stratégies*, on peut écrire un algorithme de principe sous la forme suivante :

1. Choisir arbitrairement la stratégie initiale $u^0 \in \mathcal{U}$.
   Poser $q := 0$.
2. Etant donné une stratégie $u^q \in \mathcal{U}$, calculer v^q solution de

$$\begin{cases} v_i^q = \min\left(\beta\left\{c_i(u_i^q) + \sum\limits_j M_{ij}(u_i^q)v_j^q\right\}, \psi_i\right), & i \in \Omega, \\ v_i^q = \varphi_i, & i \notin \Omega. \end{cases} \tag{13.39}$$

3. Calculer u^{q+1} solution, pour tout i, de

$$u_i^{q+1} \in \arg\min_{u\in U_i}\left\{c_i(u) + \sum_j M_{ij}(u)v_j^q\right\}. \tag{13.40}$$

4. $q := q + 1$, aller en 1.

Nous admettons la proposition suivante, dont la démonstration, extension de celle de la proposition 5.7, utilise (1.28).

Proposition 13.14 *L'algorithme ci-dessus, initialisé avec une stratégie $u^0 \in \mathcal{U}$ quelconque, est bien défini, et forme une suite de valeurs v^q décroissante, et qui vérifie $\|v^{q+1} - V\| \leq \beta\|v^q - V\|$, où V est solution unique de* (5.35).

13.1.9 Un algorithme implémentable

L'algorithme d'itérations sur les stratégies que nous venons de présenter nécessite à chaque itération la résolution de l'équation non linéaire (5.39), ce qui peut être très coûteux. Nous allons formuler un autre algorithme, itérant sur les stratégies, dans lequel on ne résout qu'une équation linéaire à chaque itération. L'idée est, étant donné une stratégie u^q, de calculer v^q solution de l'équation linéaire

$$\left\{ \begin{array}{ll} v_i^q = \beta \left(c_i(u_i^q) + \sum_j M_{ij}(u_i^q) v_j^q \right), & i \in I^q, \\ v_i^q = \psi_i, & i \in \Omega \setminus I^q, \\ v_i^q = \varphi_i, & i \notin \Omega. \end{array} \right. \tag{13.41}$$

L'ensemble I^q, inclus dans Ω, est une prédiction des états i pour lesquels la contrainte $v_i \leq \psi_i$ n'est pas active à l'optimum. Cette prédiction est mise à jour à chaque itération. Ceci conduit à l'algorithme suivant :

1. Choisir arbitrairement la stratégie initiale $u^0 \in \mathcal{U}$.
   Calculer $\hat{v}^0$ solution de l'équation linéaire

$$\left\{ \begin{array}{ll} \hat{v}_i^0 = \beta \left(c_i(u_i^0) + \sum_j M_{ij}(u_i^0) \hat{v}_j^0 \right), & i \in \Omega, \\ \hat{v}_i^0 = \varphi_i, & i \notin \Omega. \end{array} \right. \tag{13.42}$$

   Calculer v^0 comme suit :

$$\left\{ \begin{array}{ll} v_i^0 = \min(\hat{v}_i^0, \psi_i), & i \in \Omega, \\ \hat{v}_i^0 = \varphi_i, & i \notin \Omega. \end{array} \right. \tag{13.43}$$

   Poser $q := 0$ et

$$I^0 := \{ i \in \Omega; \;\; v_i^0 < \psi_i \}. \tag{13.44}$$

2. Faire $q := q + 1$. Calculer u^q solution de

$$u_i^q \in \arg\min_{u \in U_i} \left\{ c_i(u) + \sum_j M_{ij}(u) v_j^{q-1} \right\}, \quad i = 1, \ldots, m. \tag{13.45}$$

3. Poser

$$I^q := I^{q-1} \cup \left\{ i \in \Omega; \;\; \beta \left(c_i(u_i^q) + \sum_j M_{ij}(u_i^q) v_j^{q-1} \right) < \psi_i \right\}. \tag{13.46}$$

4. Calculer v^q, solution de l'équation linéaire (5.41).
   Aller en 2.

Proposition 13.15 *L'algorithme ci-dessus forme une suite de valeurs v^q décroissant vers la solution unique V de* (5.35).

Démonstration. a) Montrons la décroissance de v^q. S'il n'en est pas ainsi, soient $q \in I\!N$ et $i \in \Omega$ tels que $v_i^{q+1} - v_i^q > 0$. Etant donné $\varepsilon > 0$, on peut supposer que $(v^{q+1} - v^q)_i \geq \sup_j (v^{q+1} - v^q)_j - \varepsilon$. Par ailleurs, $i \in I^{q+1}$ (sinon v_i^{q+1} et v_i^q sont égaux à ψ_i). Donc

$$v_i^{q+1} = \beta \left(c_i(u_i^{q+1}) + \sum_j M_{ij}(u_i^{q+1}) v_j^{q+1} \right). \tag{13.47}$$

Posons $w := v^{q+1} - v^q$, et distinguons deux cas. Si $i \in I^q$, alors

$$v_i^q = \beta \left(c_i(u_i^q) + \sum_j M_{ij}(u_i^q) v_j^q \right), \tag{13.48}$$

et donc avec (5.45)

$$\begin{array}{rcl} w_i & = & \beta \left(c_i(u_i^{q+1}) + \sum_j M_{ij}(u_i^{q+1}) v_j^{q+1} - c_i(u_i^q) - \sum_j M_{ij}(u_i^q) v_j^q \right), \\ & \leq & \beta \left(\sum_j M_{ij}(u^{q+1}) w_j \right) \leq \beta(w_i + \varepsilon), \end{array} \tag{13.49}$$

ce qui donne la contradiction recherchée pour $\varepsilon > 0$ assez petit.

Si, au contraire, $i \notin I^q$, alors $v_i^q = \psi_i$ et, par définition de I^{q+1}, on a

$$\beta \left(c_i(u_i^{q+1}) + \sum_j M_{ij}(u_i^{q+1}) v_j^q \right) < \psi_i = v_i^q. \tag{13.50}$$

Donc

$$\begin{array}{rcl} w_i & = & \beta \left(c_i(u_i^{q+1}) + \sum_j M_{ij}(u_i^{q+1}) v_j^{q+1} \right)_i - \psi_i, \\ & \leq & \beta \left(c_i(u_i^{q+1}) + \sum_j M_{ij}(u_i^{q+1}) v_j^{q+1} - c_i(u_i^{q+1}) - \sum_j M_{ij}(u_i^{q+1}) v_j^q \right)_i, \end{array} \tag{13.51}$$

ce qui permet de conclure de la même manière.

b) On peut montrer, par des arguments déjà employés, que la suite v^q est bornée. Puisqu'elle est décroissante, elle converge vers une valeur $\hat{v}$. De même, I^q étant croissant, converge vers un certain I^*. Enfin par compacité on a la convergence de u^q vers

$\hat{u} \in \mathcal{U}$ pour une sous suite. Passant à la limite dans (5.41)[1], il vient

$$\begin{cases} \hat{v}_i = \beta \left(c_i(\hat{u}_i) + \sum_j M_{ij}(\hat{u}_i)\hat{v}_j \right), & i \in I^*, \\ \hat{v}_i = \psi_i, & i \in \Omega \setminus I^*, \\ \hat{v}_i = \varphi_i, & i \notin \Omega. \end{cases} \tag{13.52}$$

De plus la décroissance de v^q implique

$$\hat{v}_i \leq \psi_i, \quad i \in I^*, \tag{13.53}$$

et le passage à la limite dans (5.46) donne

$$\beta \left(c_i(\hat{u}_i) + \sum_j M_{ij}(\hat{u}_i)\hat{v}_j \right) \geq \psi_i, \quad i \in \Omega \setminus I^*. \tag{13.54}$$

Les trois relations ci-dessus impliquent que $\hat{v}$ est solution de (5.35), donc est égale à la fonction valeur V. ■

Remarque 13.16 L'algorithme présenté dans cette section peut s'avérer lent si la mise à jour de l'ensemble I^q n'est pas assez efficace. On peut y remédier, soit en introduisant quelques itérations sur les valeurs (peu coûteuses, comparées à la résolution du système (5.42)), soit en s'inspirant des algorithmes de résolution de problèmes de complémentarité linéaire, par exemple ceux basés sur les points intérieurs.

13.2 Problèmes en temps et espace continus

13.2.1 Position du problème

Etudions le problème de commande optimale stochastique

$$(P_x) \qquad \begin{cases} \text{Min } I\!E \displaystyle\int_0^\infty \ell(y(t), u(t)) e^{-\lambda t} \mathrm{d}t; \\ \mathrm{d}y(t) = f(y(t), u(t))\mathrm{d}t + \sigma(y(t), u(t))\mathrm{d}w, \quad u(t) \in U, \quad t \in [0, \infty[, \\ y_0 = x. \end{cases}$$

Dans ce problème nous retrouvons les ingrédients du problème de commande optimale déterministe : le taux d'actualisation $\lambda > 0$, les fonctions $\ell : I\!R^n \times I\!R^m \to I\!R$ et $f : I\!R^n \times I\!R^m \to I\!R^n$, tandis qu'apparaissent $\sigma(\cdot, \cdot)$, application de $I\!R^n \times I\!R^m$ vers l'espace des matrices de taille $n \times r$, et w, brownien standard de dimension r. On suppose dans la suite ℓ, f et σ lipschitziens et bornés.

[1] Par des arguments similaires à ceux employés dans la démonstration du théorème 5.5(iii).

Rappelons qu'un mouvement brownien standard (scalaire) sur l'intervalle de temps $\mathbb{R}_+$ est une variable aléatoire $\mathbb{R}_+ \to \mathbb{R}$ telle que (i) ses accroissements sont indépendants, (ii) $w(0)$ est gaussien de moyenne nulle, et (iii) si $0 \leq s \leq t < \infty$, alors $w(t) - w(s)$ est gaussien de moyenne nulle et variance $t - s$. Un brownien standard de dimension r est un vecteur aléatoire dont les composantes sont des mouvement brownien standard scalaires indépendants.

L'étude de ce problème comporte deux phases : l'analyse mathématique, qui conduit à une équation HJB avec un opérateur différentiel du second ordre, et l'analyse numérique de cette équation HJB. Nous allons commencer par présenter une version en temps discret du problème, qui permettra une dérivation formelle de l'équation HJB.

13.2.2 Problème discrétisé en temps

Soit $h_0 > 0$ le pas de temps. Considérons le problème de commande optimale stochastique en temps discret et espace continu :

$$(P_x^{h_0}) \qquad \left\{ \begin{array}{l} \text{Min } \mathbb{E}\left\{ h_0 \sum_{k=0}^{\infty} (1+\lambda h_0)^{-k-1} \ell(y_k, u_k) \right\}; \\ \\ y_{k+1} = y_k + h_0 f(y_k, u_k) + \sqrt{h_0}\, \sigma(y_k, u_k)\, \delta w_k, \quad u_k \in U, \quad k \in \mathbb{N}; \\ \\ y_0 = x. \end{array} \right.$$

Ici $\delta w_k \in \mathbb{R}^r$ est un vecteur aléatoire dont les coordonnées sont des tirages indépendants de ± 1 avec probabilités égales, donc de moyenne nulle et variance unité. Le terme $\sqrt{h_0}$ fait que, pour h_0 assez petit, si la $i$ème ligne de $\sigma(y_k, u_k)$ n'est pas nulle, alors l'essentiel de la variation de la $i$ème composante de l'état est due au bruit. Par ailleurs si $0 \leq s \leq t < \infty$, $s = k_0 h_0$ et $t = k_1 h_0$, alors $\sum_{k=k_0}^{k_1-1} \delta w_k$ est une variable de moyenne nulle et variance $t - s$, ce qui est cohérent avec le problème continu.

A la différence du cas déterministe, il faut préciser quelle information est disponible quand on prend la décision u^k à l'instant k. Par exemple, si les tirages sont connus d'avance, on se retrouve dans une situation déterministe. En général le tirage δw_k n'est pas déterminé jusqu'à l'instant $k+1$; l'information sur ce tirage et sur l'état y_k peut être totale, partielle ou nulle. Il y a donc une variété de situations possibles.

Dans la suite nous supposerons que la décision u_k se fait en connaissant l'état y_k, mais pas les tirages δw_i, pour $i \geq k$: c'est le cas dit de l'*observation complète*. Compte tenu de l'invariance en temps du problème, ceci conduit à chercher une commande sous forme de retour d'état. Autrement dit l'ensemble $\mathcal{U}$ des commandes admissibles est celui des applications $u = u(y)$ de $\mathbb{R}^n$ vers U. A $u \in \mathcal{U}$ est associé un coût $\mathcal{V}^{h_0}(x, u)$ vérifiant la relation suivante (noter que l'espérance ci-dessous se réduit à la somme de deux termes)

$$\mathcal{V}^{h_0}(x,u) = (1+\lambda h_0)^{-1} \left(h_0 \ell(x,u) + \mathbb{E}\left(V(x + h_0 f(x,u) + \sqrt{h_0}\sigma(x,u)\delta w_0) \right) \right). \tag{13.55}$$

On pose

$$V^{h_0}(x) := \inf_{u\in U} \mathcal{V}^{h_0}(x,u). \tag{13.56}$$

Le *principe de programmation dynamique* s'écrit

$$V^{h_0}(x) = (1+\lambda h_0)^{-1} \inf_{u\in U} \left\{ h_0 \ell(x,u) + I\!E\left(V(x + h_0 f(x,u) + \sqrt{h_0}\sigma(x,u)\delta w_0)\right)\right\}. \tag{13.57}$$

Supposons V^{h_0} de classe C^2, et de dérivée seconde uniformément bornées sur $I\!R^n$, uniformément par rapport à h_0 assez petit. Alors

$$\begin{array}{rcl} \Delta & := & V^{h_0}(x + h_0 f(x,u) + \sqrt{h_0}\sigma(x,u)\delta w_0), \\ & = & V^{h_0}(x) + h_0 DV^{h_0}(x) f(x,u) + \sqrt{h_0} DV^{h_0}(x)\sigma(x,u)\delta w_0 \\ & & + \frac{1}{2} h_0 D^2 V^{h_0}(x)(\sigma(x,u)\delta w_0, \sigma(x,u)\delta w_0) + o(h_0). \end{array} \tag{13.58}$$

Si A est une matrice $n\times n$ et $z\in I\!R^n$, on a $z^\top A z = \mathrm{trace}\, A z z^\top$. Utilisant cette relation, il vient

$$D^2 V^{h_0}(x)(\sigma(x,u)\delta w_0, \sigma(x,u)\delta w_0) = \mathrm{trace}\left(D^2 V^{h_0}(x)\sigma(x,u)\delta w_0 \delta w_0^\top \sigma(x,u)^\top\right). \tag{13.59}$$

Posons

$$a(x,u) := \tfrac{1}{2}\sigma(x,u)\sigma(x,u)^\top. \tag{13.60}$$

La matrice $n\times n$ $a(x,u)$ est symétrique et semi définie positive. Puisque w est de moyenne nulle et variance unité, on a, avec la relation précédente :

$$I\!E(\Delta) = V^{h_0}(x) + h_0 DV^{h_0}(x) f(x,u) + h_0 \ \mathrm{trace}\left(D^2 V^{h_0}(x) a(x,u)\right) + o(h_0). \tag{13.61}$$

Noter que

$$\mathrm{trace}\left(D^2 V^{h_0}(x) a(x,u)\right) = \sum_{i,j=1}^n a_{ij}(x,u) D^2_{x_i x_j} V^{h_0}(x). \tag{13.62}$$

Introduisons le hamiltonien $\mathcal{H}^\sigma$:

$$\mathcal{H}^\sigma(x,p,Q) := \inf_{u\in U}\{\ell(x,u) + p\cdot f(x,u) + \mathrm{trace}(a(x,u)Q)\}. \tag{13.63}$$

Ici $p\in I\!R^n$ et Q est une matrice symétrique $n\times n$. L'exposant σ fait réference au terme du deuxième ordre qui fait la différence avec le cas déterministe, voir (3.18).

Combinant avec le principe de programmation dynamique (5.57), il vient :

$$\lambda V^{h_0}(x) = \mathcal{H}^\sigma(x, DV^{h_0}(x), D^2 V^{h_0}(x)) + o(1). \tag{13.64}$$

Passant à la limite quand $h_0 \downarrow 0$, on obtient formellement l'équation HJB du problème en temps continu :

$$\lambda V(x) = \mathcal{H}^\sigma(x, DV(x), D^2 V(x)), \tag{13.65}$$

ou encore

$$\lambda V(x) = \inf_{u\in U}\left\{ \ell(x,u) + f(x,u)\cdot DV(x) + \mathrm{trace}(a(x,u)D^2 V(x))\right\}. \tag{13.66}$$

Lorsque $\sigma(x,u)$ est identiquement nul, on retrouve bien l'équation HJB (3.23) du cas déterministe (avec ici $C=\emptyset$).

Dans le cas d'un problème avec horizon fini T et sans terme d'actualisation, une discussion analogue à celle de l'horizon infini permet d'obtenir une équation de Hamilton-Jacobi-Bellman du problème continu, dont est solution la fonction valeur en temps rétrograde

$$W(x,s) := V(x, T-s).$$

Cette équation s'écrit :

$$\left\{ \begin{array}{ll} D_t W(x,t) = \mathcal{H}^\sigma(x, D_x W(x,t), D^2_{xx} W(x,t)), & (x,t) \in I\!R^n \times]0,T[, \\ W(x,0) = \varphi(x), & \forall\, x \in I\!R^n, \end{array} \right. \tag{13.67}$$

ou encore

$$\begin{array}{rl} D_t W(x,t) = & \inf_{u\in U} \left\{ \ell(x,u) + f(x,u)\cdot DW(x,t) + \mathrm{trace}(a(x,u) D^2 W(x,t)) \right\}, \\ & (x,t) \in I\!R^n \times]0,T[, \\ W(x,0) = & \varphi(x), \quad \forall\, x \in I\!R^n. \end{array} \tag{13.68}$$

Nous allons étudier la résolution numérique de cette équation par des schémas aux différences finies, en commençant par le cas d'un état scalaire.

13.2.3 Schémas monotones : dimension 1

On note h_0, h_1, etc les pas de discrétisation en temps et suivants les variables d'espace x_1, etc. Nous discutons les schémas de résolution de problèmes à horizon infini. Présentons une extension, en dimension un, de l'algorithme décentré, dans lequel on approxime la dérivée seconde en espace (suivant la direction de x_i) par $(D^d w_j^k - D^g w_j^k)/h_i$, soit la différence divisée centrée

$$D^{2,0} w_j^k := \frac{1}{h_i^2}(w_{j+1}^k - 2w_j^k + w_{j-1}^k).$$

Le schéma décentré (analogue de celui du cas déterministe, exposé en section 4.2.1) s'écrit alors

$$\begin{array}{rl} \lambda v_j = & \inf\limits_{u\in U} \left\{ \ell(x_j,u) + f(x_j,u)_+ \dfrac{v_{j+1}-v_j}{h_1} + |f(x_j,u)_-| \dfrac{v_{j-1}-v_j}{h_1} \right. \\ & \left. \qquad + a(x_j,u) \dfrac{v_{j+1} - 2v_j + v_{j-1}}{h_1^2} \right\}. \end{array} \tag{13.69}$$

Introduisons un *pas de temps fictif* $h_0 > 0$, par lequel on multiplie l'équation ci-dessus. Ajoutant v_j à chaque membre, et ordonnant les expressions suivant v_{j-1}, v_{j+1} et v_{j+1}, on obtient l'expression équivalente

$$\begin{array}{l} \lambda v_j := \inf\limits_{u\in U} \left\{ h_0 \ell(x_j,u) + \left(1 - \dfrac{h_0}{h_1} |f(x_j,u)| - 2\dfrac{h_0}{h_1^2} a(x_j,u) \right) v_j \right. \\ \left. + \left(\dfrac{h_0}{h_1} |f(x_j,u)_-| + \dfrac{h_0}{h_1^2} a(x_j,u) \right) v_{j-1} + \left(\dfrac{h_0}{h_1} f(x_j,u)_+ + \dfrac{h_0}{h_1^2} a(x_j,u) \right) v_{j+1} \right\}. \end{array} \tag{13.70}$$

On pose

$$\|f\|_\infty := \sup_{(x,u)\in I\!R\times U} |f(x,u)|; \quad \|a\|_\infty := \sup_{(x,u)\in I\!R\times U} |a(x,u)|. \tag{13.71}$$

Proposition 13.17 (i) *Le schéma* (5.69) *possède une solution unique, telle que*

$$\|v\|_\infty \leq \lambda^{-1}\|\ell\|_\infty. \tag{13.72}$$

(ii) *Si* h_0 *vérifie la* condition de stabilité

$$\frac{h_0}{h_1}\|f\|_\infty + \frac{2h_0}{h_1^2}\|a\|_\infty^2 \leq 1, \tag{13.73}$$

alors (5.70) *est une équation de point fixe contractant pour la norme uniforme, de rapport de contraction* $(1+\lambda h_0)^{-1}$.

Démonstration. La démonstration est semblable à celle de la proposition 4.6. La condition de stabilité assure que, dans la formule (5.70), les poids de v_j et $v_{j\pm1}$ sont positif, ce qui permet d'établir que c'est une équation de point fixe contractant et d'obtenir l'estimation (5.72). ■

Remarque 13.18 Le terme dominant dans la condition de stabilité est lié à f si h_1 est grand par rapport à $2\|a\|_\infty/\|f\|_\infty$ (discrétisation spatiale grossière), et au terme de diffusion si h_1 est grand par rapport à $2\|a\|_\infty/\|f\|_\infty$ (discrétisation spatiale fine). Dans ce dernier cas, le pas de temps maximum respectant la condition de stabilité est de l'ordre de $\frac{1}{2}h_1^2/\|a\|_\infty$, donc beaucoup plus petit que dans le cas déterministe (où il vaut $h_1/\|f\|_\infty$).

Remarque 13.19 La condition de stabilité assure la positivité des poids de v_j et $v_{j\pm1}$ dans (5.70), ce qui permet de reconnaître dans cette expression le principe de programmation dynamique du problème de commande d'une chaîne de Markov dont les probabilités de transition sont précisément les poids de v_j et $v_{j\pm1}$.

Remarque 13.20 L'étude de la convergence de ce schéma est trop complexe pour être traitée ici. On se reportera aux notes de fin de chapitre.

Dans le cas de dimension d'espace supérieure à 1, on sait seulement donner des réponses *partielles* au problème de discrétisation par différence finie de l'équation HJB. Nous allons poser le problème et établir quelques résultats.

13.2.4 Différences finies classiques

Nous abordons l'études des schémas de différences finies en dimension d'espace $n > 1$. Notons D_i les dérivées par rapport à x_i, et adoptons le même type de convention pour les dérivées d'ordre supérieur. Pour approcher D_{ii} on utilise encore la formule centrée

$$D^2_{ii} v_j \approx \frac{v_{j+e_i} - 2v_j + v_{j-e_i}}{h_i^2}.$$

Pour alléger les formules il convient de noter $\delta_{\pm i}$, $\delta_{\pm, i\pm k}$, etc les opérateurs de translation de $\pm$ une coordonnée dans la direction i, k, etc ; ainsi

$$\delta_i v_j = v_{j+e_i}, \quad \delta_{i,-k} v_j = v_{j+e_i-e_k}.$$

Avec cette notation l'approximation de D_{ii} est

$$D^2_{ii} \approx \frac{\delta_i - 2\delta_0 + \delta_{-i}}{h_i^2}.$$

Pour le calcul des dérivées croisées ($i \neq j$), plusieurs choix sont possibles. Par exemple, utilisant le développement, pour Φ régulier,

$$\begin{aligned} \Phi(x + h_i e_i + h_k e_k) = \quad & \Phi(x) + D\Phi(x)(h_i e_i + h_k e_k) + \\ & \tfrac{1}{2} D^2\Phi(x)((h_i e_i + h_k e_k), (h_i e_i + h_k e_k)) + o(h_i^2 + h_k^2), \end{aligned} \tag{13.74}$$

et procédant de même pour $\Phi(x + h_i e_i)$ et $\Phi(x + h_k e_k)$, on déduit le choix

$$D^2_{ik} \quad \approx \quad \frac{\delta_{i,k} + \delta_0 - \delta_i - \delta_k}{h_i h_k},$$

qui fait intervenir les quatre points du "rectangle en haut à droite". On peut écrire une formule similaire faisant intervenir les points du rectangle opposé :

$$D^2_{ik} \quad \approx \quad \frac{\delta_{-i,-k} + \delta_0 - \delta_{-i} - \delta_{-k}}{h_i h_k}.$$

Il est classique de centrer l'estimation en prenant la moyenne des deux, ce qui donne

$$D^2_{ik} \approx \frac{\delta_{i,k} + \delta_{-i,-k} + 2\delta_0 - \delta_i - \delta_k - \delta_{-i} - \delta_{-k}}{2h_i h_k}. \tag{13.75}$$

Mais on peut aussi bien faire intervenir les estimations basées sur les deux autres rectangles :

$$D^2_{ik} \approx \frac{\delta_i + \delta_k + \delta_{-i} + \delta_{-k} - \delta_{i,-k} - \delta_{-i,k} - 2\delta_0}{2h_i h_k}. \tag{13.76}$$

Le point important est que ces deux formules font apparaître les points $\delta_{\pm i,\pm k}$ avec des poids positifs dans le premier cas, et négatifs dans le second. Soit $\hat{D}^{x,u}$ la matrice

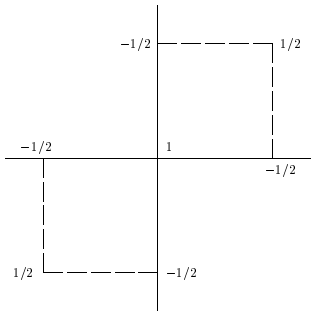

FIG. 13.1 – Poids de l'approximation de D^2_{ij} : cas $a_{ij} > 0$

$n \times n$ d'opérateurs aux différences définie par

$$\hat{D}^{x,u}_{ik} = \begin{cases} \dfrac{\delta_i - 2\delta_0 + \delta_{-i}}{h_i^2} & \text{si } i = k, \\ \dfrac{\delta_{i,k} + \delta_{-i,-k} + 2\delta_0 - \delta_i - \delta_k - \delta_{-i} - \delta_{-k}}{2h_i h_k} & \text{si } a_{ik}(x,u) \geq 0, \\ \dfrac{\delta_i + \delta_k + \delta_{-i} + \delta_{-k} - \delta_{i,-k} - \delta_{-i,k} - 2\delta_0}{2h_i h_k} & \text{sinon.} \end{cases}$$

Pour les termes du premier ordre, on reprend le principe du décentrage exposé dans le cas de la commande optimale déterministe : à (x,u), associons $D^{\eta(x_j,u)} \in I\!R^n$ défini par

$$D^{\eta_i(x,u)} = \begin{cases} \dfrac{v_{j+e_i} - v_j}{h_i} & \text{si} \quad f_i(x,u) \geq 0, \\ \dfrac{v_j - v_{j-e_i}}{h_i} & \text{sinon.} \end{cases} \tag{13.77}$$

Considérons le schéma discret

$$\lambda v_j = \min_{u \in U} \left\{ \ell(x_j,u) + f(x_j,u) \cdot D^{\eta(x_j,u)} v_j + \sum_{i,k=1}^{n} a_{ik}(x_j,u) \hat{D}^{x,u}_{ik} v_j \right\}. \tag{13.78}$$

Multipliant l'équation par un pas de temps fictif h_0, ajoutant v_j à chaque membre,

et réordonnant les expressions, il vient (utilisant la symétrie de a) :

$$\begin{aligned}
\lambda v_j = \min_{u\in U} \{ & h_0 \ell(x_j,u) \\
& + \left(1 - \sum_{i=1}^n \frac{h_0}{h_i}|f_i(x_j,u)| - 2\sum_{i=1}^n \frac{h_0}{h_i^2}|a_{ii}(x_j,u)| + \sum_{i\neq k} \frac{h_0}{h_i h_k}|a_{ik}(x_j,u)| \right) v_j \\
& + \sum_{i=1}^n \left(\frac{h_0}{h_i}|f_i(x_j,u)_-| + \frac{h_0}{h_i^2}a_{ii}(x_j,u) - \sum_{k\neq i} \frac{h_0}{h_i h_k}|a_{ik}(x_j,u)| \right) v_{j-e_i} \\
& + \sum_{i=1}^n \left(\frac{h_0}{h_i} f_i(x_j,u)_+ + \frac{h_0}{h_i^2}a_{ii}(x_j,u) - \sum_{k\neq i} \frac{h_0}{h_i h_k}|a_{ik}(x_j,u)| \right) v_{j+e_i} \\
& + \sum_{i>k} \frac{h_0}{h_i h_k} \left[a_{ik}(x_j,u)_+ (v_{j+e_i+e_k} + v_{j-e_i-e_k}) + |a_{ik}(x_j,u)_-|(v_{j+e_i-e_k} + v_{j-e_i+e_k}) \right] \Bigg\}.
\end{aligned} \tag{13.79}$$

On peut introduire une mise à l'échelle de f et a :

$$f_i^h(x,u) := \frac{f_i(x,u)}{h_i}; \qquad a_{ij}^h(x,u) := \frac{a_{ij}(x,u)}{h_i h_j}; \tag{13.80}$$

d'où l'expression équivalente

$$\begin{aligned}
(1+\lambda h_0) v_j = \min_{u\in U} \{ & h_0\, \ell(x_j,u) \\
& + \left(1 - h_0 \sum_{i=1}^n |f_i^h(x_j,u)| - 2h_0 \sum_{i=1}^n |a_{ii}^h(x_j,u)| + h^0 \sum_{i\neq k} |a_{ik}^h(x_j,u)| \right) v_j \\
& + h_0 \sum_{i=1}^n \left(|f_i^h(x_j,u)_-| + a_{ii}^h(x_j,u) - \sum_{k\neq i} |a_{ik}^h(x_j,u)| \right) v_{j-e_i} \\
& + h_0 \sum_{i=1}^n \left(f_i^h(x_j,u)_+ + a_{ii}^h(x_j,u) - \sum_{k\neq i} |a_{ik}^h(x_j,u)| \right) v_{j+e_i} \\
& + h_0 \sum_{i>k} \left[a_{ik}^h(x_j,u)_+ (v_{j+e_i+e_k} + v_{j-e_i-e_k}) + |a_{ik}^h(x_j,u)_-|(v_{j+e_i-e_k} + v_{j-e_i+e_k}) \right] \Bigg\}.
\end{aligned} \tag{13.81}$$

Proposition 13.21 *On suppose que les pas d'espace $h_1, \ldots, h_n$ sont tels que, pour tout $(x,u) \in \mathbb{R} \times U$, la matrice de terme général $a_{ik}^h(x,u)$ est diagonale dominante. Alors*
(i) *Le schéma* (5.78) *possède une solution unique v, telle que*

$$\|v\|_\infty \leq \lambda^{-1} \|\ell\|_\infty. \tag{13.82}$$

(ii) *Si* h_0 *vérifie la* condition de stabilité

$$h_0 \left[\sum_{i=1}^{n} \frac{|f_i(x_j,u)|}{h_i} + \sum_{i=1}^{n} \left(2\frac{|a_{ii}(x_j,u)|}{h_i^2} - \sum_{k\neq i} \frac{|a_{ik}(x_j,u)|}{h_i h_k} \right) \right] \leq 1, \tag{13.83}$$

alors (5.70) *est une équation de point fixe contractant pour la norme uniforme, de rapport de contraction* $(1+\lambda h_0)^{-1}$.

Démonstration. La démonstration est une extension simple de celle de cas monodimensionnel (proposition 5.17). ■

Quand h tend vers 0 de manière à respecter la condition de diagonale dominante de a^h, on obtient la convergence des valeurs discrètes vers la valeur du problème continu : voir [40]. La théorie des solutions de viscosité joue là encore un rôle important.

Cependant l'hypothèse de diagonale dominance de $a^h(x,u)$ est très restrictive. Nous allons présenter une approche permettant, dans une certaine mesure, de s'en affranchir.

13.2.5 Différences finies généralisées

Dans cette approche, qui généralise la méthode usuelle de différences finies présentée dans la section précédente, le point de départ est l'approximation de la dérivée seconde de la fonction valeur suivant une direction quelconque. Soit $\Phi : \mathbb{R}^n \to \mathbb{R}$ de classe C^2. La dérivée seconde de Φ en $x \in \mathbb{R}^n$ dans la direction $d \in \mathbb{R}^n$ est par définition la quantité

$$D^2\Phi(x)(d,d) = \sum_{i,k=1}^{n} D^2_{x_i x_k}\Phi(x) d_i d_k.$$

Il vient avec la formule de Taylor

$$D^2\Phi(x)(d,d) = \lim_{t\downarrow 0} \frac{\Phi(x+td) - 2\Phi(x) + \Phi(x-td)}{t^2}.$$

En particulier, étant donné $\xi \in \mathbb{Z}^n$, notons

$$\Delta_\xi \Phi := \Phi(x_{j+\xi}) - 2\Phi(x_j) + \Phi(x_{j-\xi}).$$

Il vient, pour tout $j \in \mathbb{Z}^n$,

$$\Delta_\xi \Phi(x_j) = \sum_{i,k=1}^{n} h_i h_k \xi_i \xi_k D^2_{x_i x_k}\Phi(x_j) + o(\|h\|^2). \tag{13.84}$$

Ainsi on peut approcher la courbure de Φ, suivant une direction égale à la différence entre deux points de la grille discrète, par une combinaison des valeurs de Φ en trois

points de la grille. On peut alors se poser le problème d'approcher la partie principale (du second ordre) de l'opérateur différentiel de l'équation HJB par une combinaison de tels termes. Il s'agit de trouver des coefficients $\alpha^u_{j,\xi}$ tels que :

$$\sum_{\xi\in\mathcal{S}} \alpha^u_{j,\xi} \sum_{i,k=1}^{n} h_i h_k \xi_i \xi_k D^2_{x_i x_k}\Phi(x_j) = \sum_{i,k=1}^{n} a_{ik}(x_j,u) D^2_{x_i x_k}\Phi(x_j). \tag{13.85}$$

Ici $\mathcal{S}$ est une partie finie de $\mathbb{Z}^n$, qui représente (à la translation j près) les coordonnées des points entrant dans le schéma. Nous verrons qu'il convient de prendre les coefficients $\alpha^u_{j,\xi}$ positifs pour obtenir la monotonie du schéma. La relation (5.85) est satisfaite pour toute fonction Φ ssi

$$\sum_{\xi\in\mathcal{S}} \alpha^u_{j,\xi} h_i h_k \xi_i \xi_k = a_{ik}(x_j,u), \qquad \text{pour tout} \quad i,\, k = 1 \text{ à } n, \tag{13.86}$$

ou encore

$$\sum_{\xi\in\mathcal{S}} \alpha^u_{j,\xi} \xi\xi^\top = a^h(x_j,u). \tag{13.87}$$

Le schéma correspondant (de discrétisation de l'équation HJB) est

$$\lambda v_j = \inf_{u\in U} \left\{ \ell(x_j,u) + f(x_j,u)\cdot D^{\eta(x_j,u)} v_j + \sum_{\xi\in\mathcal{S}} \alpha^u_{j,\xi} \Delta_\xi v_j \right\}, \quad j\in\mathbb{Z}^n. \tag{13.88}$$

Définition 13.22 On dira que le schéma (5.88) est *fortement consistant* si (5.87) est satisfait.

Nous étudierons dans la section suivante comment vérifier la condition de consistance.

Remarque 13.23 La relation (5.87) permet une estimation de la taille des coefficients. En effet, puisque ξ a des coordonnées entières, la matrice $\xi\xi^\top$ a des éléments diagonaux supérieurs ou égaux à un. Un schéma fortement consistant satisfait donc

$$\sum_{\xi\in\mathcal{S}} \alpha^u_{j,\xi} \le \text{trace } a^h(x_j,u) = O((\inf_i h_i)^{-2}). \tag{13.89}$$

La forme de point fixe correspondante est (comme toujours) obtenue en multipliant la relation (5.88) par un pas de temps fictif h_0, puis en ajoutant v_j à chaque membre, et enfin en divisant par $1+h_0\lambda$. Reprenant la notation f^h définie en (5.80), on obtient

l'expression suivante, à comparer à (4.23) dans le cas déterministe :

$$v_j = (1+\lambda h_0)^{-1} \inf_{u\in U} \left\{ h_0 \ell(x_j,u) + \left(1 - h_0 \sum_{i=1}^n |f_i^h(x_j,u)| - 2h_0 \sum_{\xi\in\mathcal{S}} \alpha_{j,\xi}^u \right) v_j \right.$$
$$\left. + h_0 \sum_{i=1}^n f_i^h(x_j,u)_+ v_{j+e_i} + h_0 \sum_{i=1}^n |f_i^h(x_j,u)_-| v_{j-e_i} + h_0 \sum_{\xi\in\mathcal{S}} \alpha_{j,\xi}^u (v_{j-\xi} + v_{j+\xi}) \right\}. \tag{13.90}$$

Comme dans le cas déterministe, il apparaît que le membre de droite représente une application contractante, de constante $(1+\lambda h_0)^{-1}$, si le coefficient de v_j est positif, ce qui est assuré si la condition de stabilité suivante est satisfaite :

$$h_0 \left(\sum_{i=1}^n \frac{\|f_i\|}{h_i} + 2 \sup_{j\in\mathbb{Z}^n, u\in U} \left(\sum_{\xi\in\mathcal{S}} \alpha_{j,\xi}^u \right) \right) \le 1. \tag{13.91}$$

On peut combiner cette relation avec (5.89) pour en déduire une estimation du pas de temps : $h_0 = O((\inf_i h_i)^{-2})$.

13.2.6 Analyse de la condition de consistance forte

La condition de consistance forte (5.87) revient, puisque les coefficients $\alpha_{j,\xi}^u$ doivent être positifs, à vérifier que $a^h(x_j,u)$ appartient au cône engendré par l'ensemble $\{\xi\xi^\top ; \xi\in\mathcal{S}\}$. Nous allons caractériser ce cône dans quelques situations simples. Pour cela, quelques définitions s'imposent.

Définition 13.24 Soit $q\in\mathbb{N}$, $q>0$. (i) On dit que $C\subset\mathbb{R}^q$ est un *cône* si, pour tout $t>0$ et $c\in C$, on a $tc\in C$. (ii) Soient $c_1,\dots,c_r$ dans $\mathbb{R}^q$. On appelle *cône convexe* C engendré par $c_1,\dots,c_r$ l'ensemble des combinaisons linéaires positives de $c_1,\dots,c_r$. On dit que $c_1,\dots,c_r$ est un générateur de C. (iii) On appelle *générateur minimal* de C un générateur de C ne contenant pas strictement un générateur de C.

Définition 13.25 Soit C un cône convexe fermé de $\mathbb{R}^q$. On appelle *cône polaire* de C l'ensemble

$$C^+ := \{ y\in\mathbb{R}^q ;\ \ y\cdot x \ge 0, \quad \text{pour tout}\ \ x\in C \}. \tag{13.92}$$

C'est un cône convexe fermé.

Voici un résultat important d'analyse convexe, que nous admettrons (voir par exemple [48]).

Proposition 13.26 *Soit C un cône convexe fermé. Alors* (i) *il coïncide avec son cône bipolaire $(C^+)^+$,* (ii) *Si C a un générateur fini, il en est de même pour C^+.*

Il résulte de cette proposition que, si C est un cône convexe fermé de générateur fini, et il existe donc un générateur fini $c_1^*, \ldots, c_{r'}^*$ du cône polaire, alors C est caractérisé par les inégalités linéaires en nombre fini

$$C = \{x \in I\!R^q;\ \ c_i^* \cdot x \geq 0,\ \ i = 1, \ldots, r'\}. \tag{13.93}$$

On notera $C(\mathcal{S})$ le cône engendré par les $\{\xi\xi^\top,\ \ \xi \in \mathcal{S}\}$. Considérons le cas où $\mathcal{S}$ est de la forme $\mathcal{S}_p^n$, avec

$$\mathcal{S}_p^n := \left\{ \xi \in \{-1, 0, 1\}^n; \sum_{i=1}^n |\xi_i| \leq p \right\}. \tag{13.94}$$

Autrement dit, on considère les transitions vers les points dont les coordonnées diffèrent d'au plus 1 (les voisins immédiats), avec au plus p coordonnées différentes.

Proposition 13.27 *On a les caractérisations suivantes :*
(i) *Pour tout $n > 0$, $C(\mathcal{S}_1^n)$ est l'ensemble des matrices diagonales semi définies positives.*
(ii) *Pour tout $n > 0$, $C(\mathcal{S}_2^n)$ est l'ensemble des matrices symétriques à diagonale dominante :*

$$\mathcal{C}(\mathcal{S}_2^n) = \left\{ A \in \mathcal{M}^{n\times n};\ \ A = A^\top;\ \ A_{ii} \geq \sum_{j\neq i} |A_{ij}| \right\}. \tag{13.95}$$

(iii) *$A \in C(\mathcal{S}_3^3)$ si et seulement si,elle est symétrique et, pour tout i, j dans $1, \ldots, n$ et p, q dans $\{0, 1\}$:*

$$\left\{ \begin{array}{lcl} A_{ii} & \geq & |A_{ij}|, \\ A_{ii} + A_{jj} & \geq & (-1)^p A_{ik} + (-1)^q A_{jk} + 2(-1)^{p+q+1} A_{ij}. \end{array} \right. \tag{13.96}$$

Démonstration. Si ξ a une seule coordonnée non nulle j, alors la matrice $\xi\xi^\top$ a pour seule coordonnée non nulle (j, j). On en déduit facilement (i). Si ξ a deux coordonnées non nulles, de valeur ± 1, il est clair que $\xi\xi^\top$ est diagonale dominante. Comme l'ensemble des matrices diagonale dominantes est un cône, ceci prouve que $C(\mathcal{S}_2^n)$ est inclus dans l'ensemble des matrices diagonale dominantes. Inversement, on peut décomposer toute matrice diagonale dominante A en combinaison linéaire positive de matrices du type $\xi\xi^\top$ où ξ a au plus deux coordonnées non nulles de la manière suivante :

$$\begin{array}{rcl} A & = & \displaystyle\sum_{\substack{i\neq j \\ A_{ij}>0}} A_{ij}(e_i + e_j)(e_i + e_j)^\top - \sum_{\substack{i\neq j \\ A_{ij}<0}} A_{ij}(e_i - e_j)(e_i - e_j)^\top + \\ & & \displaystyle\sum_i \left(A_{ii} - \sum_{j\neq i} |A_{ij}| \right) e_i e_i^\top, \end{array} \tag{13.97}$$

d'où (ii). Enfin le point (iii) résulte de l'analyse de [15]. ■

Remarque 13.28 Une des questions ouvertes est le calcul rapide des coefficients $\alpha^u_{j,\xi}$. Pour la dimension 2 le problème est traité dans [14].

13.3 Notes

La commande optimale de chaînes de Markov est discutée dans J.P. Quadrat [49]. E. Altman [5] étudie les problèmes avec contraintes en espérance. Le cas ergodique fait l'objet d'un chapitre de H.J. Kushner et P.G. Dupuis [40].

W. H. Fleming et R. Rishel [28] donnent une introduction générale à la théorie de la commande optimale déterministe et stochastique. L'approche par solutions de viscosité est introduite dans P.L. Lions [45] ; on en trouvera une synthèse dans W.H. Fleming et H.M. Soner [29]. J.L. Lions et A. Bensoussan [44] présentent l'approche de la commande stochastique par les techniques variationnelles d'équations aux dérivés partielles.

Les méthodes numériques pour la commande stochastique sont exposées dans H.J. Kushner et P.G. Dupuis [40]. On y trouvera en particulier une discussion d'une méthode d'approximation par chaîne de Markov qui inclut les différences finies généralisées. Pour les problèmes de très grande taille il peut être utile d'employer des méthodes multigrille, voir M. Akian [2]. De nombreuses méthodes numériques, dans un cadre de problèmes de finance, sont exposées dans L.C.G. Rogers et D. Talay [50].

Bibliographie

[1] R.H. Abraham and C.D. Shaw. *Dynamics – The Geometry of Behavior : I-IV*. Aerial Press, Santa Cruz, California, 1981.

[2] M. Akian. Analyse de l'algorithme multigrille FMGH de résolution d'équations d'Hamilton-Jacobi-Bellman. In A. Bensoussan and J.-L. Lions, editors, *Analysis and optimization of systems (Antibes, 1990)*, volume 144 of *Lecture Notes in Control and Information Sciences*, pages 113–122. Springer Verlag, Berlin, 1990.

[3] V. Alexéev, V. Tikhomirov, and S. Fomine. *Commande optimale*. Mir, Moscow, 1982. Edition originale : Mir, Moscou, 1979.

[4] G. Allaire. *Analyse numérique et optimisation*. Ecole Polytechnique, mathématiques appliquées, 2002.

[5] E. Altman. *Constrained Markov decision processes*. Chapman and Hall, Boca Raton, 1999.

[6] V. Arnold. *Equations Différentielles Ordinaires*. Mir Moscou, 1974.

[7] V. Arnold. *Méthodes Mathématiques de la Mécanique Classique*. Mir Moscou, 1976.

[8] V. Arnold. *Chapitres Supplémentaires de la Théorie des Equations Différentielles Ordinaires*. Mir Moscou, 1980.

[9] M. Bardi and I. Capuzzo-Dolcetta. *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations*. Systems and Control : Foundations and Applications. Birkhäuser, Boston, 1997.

[10] G. Barles. *Solutions de viscosité des équations de Hamilton-Jacobi*, volume 17 of *Mathématiques et Applications*. Springer, Paris, 1994.

[11] G. Barles and P. E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Analysis*, 4 :271–283, 1991.

[12] R. Bellman. *Dynamic programming*. Princeton University Press, Princeton, 1961.

[13] D. Bertsekas. *Dynamic programming and optimal control (2 volumes)*. Athena Scientific, Belmont, Massachusetts, 1995.

[14] J. F. Bonnans, E. Ottenwaelter, and H. Zidani. Numerical schemes for the two dimensional second-order HJB equation. *ESAIM :M2AN 38-4, 723-735*, 2004.

[15] J. F. Bonnans and H. Zidani. Consistency of generalized finite difference schemes for the stochastic HJB equation. *SIAM J. Numerical Analysis*, 41 :1008–1021, 2003.

[16] J.F. Bonnans and A. Shapiro. *Perturbation analysis of optimization problems*. Springer-Verlag, New York, 2000.

[17] J.P. Bourguignon. *Calcul Variationnel*. Ecole Polytechnique, 1989.

[18] K.E. Brenan, S.L. Campbell, and L.R. Petzold. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*. North-Holland, Amsterdam, 1989.

[19] H. Brézis. *Analyse fonctionnelle*. Masson, Paris, 1983.

[20] A. E. Bryson and Y.-C. Ho. *Applied optimal control*. Hemisphere Publishing, New-York, 1975.

[21] F.H. Clarke. *Optimization and nonsmooth analysis*. Wiley, New York, 1983.

[22] M. G. Crandall and P.-L. Lions. Two approximations of solutions of Hamilton-Jacobi equations. *Mathematics of Computation*, 43 :1–19, 1984.

[23] M.G. Crandall and P.-L. Lions. Viscosity solutions of Hamilton Jacobi equations. *Bull. American Mathematical Society*, 277 :1–42, 1983.

[24] M. Crouzeix and A.L. Mignot. *Analyse Numérique des Equations Différentielles*. Masson, Paris, 1992.

[25] M. Demazure. *Géométrie, Catastrophes et Bifurcations*. Ecole Polytechnique, 1987.

[26] I. Capuzzo Dolcetta and H. Ishii. Approximate solutions of the Bellman equation of deterministic control theory. *Appl. Math. Optim.*, 11 :161–181, 1984.

[27] P. Faurre, M. Clerget, and F. Germain. *Opérateurs rationnels positifs*. Dunod, Paris, 1979.

[28] W. H. Fleming and R. Rishel. *Deterministic and stochastic optimal control*, volume 1 of *Applications of mathematics*. Springer, New York, 1975.

[29] W.H. Fleming and H.M. Soner. *Controlled Markov processes and viscosity solutions*. Springer, New York, 1993.

[30] H. Frankowska. Value function in optimal control, 2001. Lecture notes, Summer School on Mathematical Control Theory, Trieste.

[31] F.R. Gantmacher. *Théorie des Matrices : tome 1*. Dunod, Paris, 1966.

[32] F.R. Gantmacher. *Théorie des Matrices : tome 2*. Dunod, Paris, 1966.

[33] J.P. Gauthier and I. Kupka. *Deterministic Observation Theory and Applications*. Cambridge University Press, 2001.

[34] C. Godbillon. *Géométrie différentielle et mécanique analytique*. Hermann, Paris, 1969.

[35] J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*. Springer, New York, 1983.

[36] M.W. Hirsch and S. Smale. *Differential Equations, Dynamical Systems and Linear Algebra*. Acamedic Press : New-York, 1974.

[37] A.D. Ioffe and V.M. Tihomirov. *Theory of Extremal Problems*. North-Holland Publishing Company, Amsterdam, 1979. Russian Edition : Nauka, Moscow, 1974.

[38] T. Kailath. *Linear Systems*. Prentice-Hall, Englewood Cliffs, NJ, 1980.

[39] H.K. Khalil. *Nonlinear Systems*. MacMillan, 1992.

[40] H.J. Kushner and P.G. Dupuis. *Numerical methods for stochastic control problems in continuous time*, volume 24 of *Applications of mathematics*. Springer, New York, 2001. Second edition.

[41] J.P. LaSalle and S. Lefschetz. *Stability by Liapounov's Direct Method With Applications*. Academic Press, New York, 1961.

[42] E.B. Lee and L. Markus. *Foundations of optimal control theory*. John Wiley, New York, 1967.

[43] G. Leitmann. *An introduction to optimal control*. Mc Graw Hill, New York, 1966.

[44] J.-L. Lions and A. Bensoussan. *Application des inéquations variationnelles en contrôle stochastique*, volume 6 of *Méthodes mathématiques de l'informatique*. Dunod, Paris, 1978.

[45] P.-L. Lions. Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations. Part 2 : viscosity solutions and uniqueness. *Communications in partial differential equations*, 8 :1229–1276, 1983.

[46] Ph. Martin, R. Murray, and P. Rouchon. Flat systems, equivalence and trajectory generation, 2003. Technical Report http ://www.cds.caltech.edu/reports/.

[47] I. McCausland. *Introduction to optimal control*. J. Wiley, New York, 1969.

[48] G.L. Nemhauser, A.H.G. Rinnoy Kan, and M.J. Todd, editors. *Optimization*, volume 1 of *Handbooks in Operations Research and Management Science*. North-Holland, Amsterdam, 1989.

[49] J.P. Quadrat. *Décision et commande en présence d'incertitude*. Ecole Polytechnique, Palaiseau, 1994. Polycopié de cours.

[50] L. C. G. Rogers and D. Talay, editors. *Numerical methods in finance*. Cambridge University Press, 1997.

[51] W. Rudin. *Real and complex analysis*. Mc Graw-Hill, New York, 1987.

[52] L. Schwartz. *Méthodes mathématiques pour les sciences physiques*. Hermann, Paris, 1965.

[53] E. Sontag. *Mathematical Control Theory*. Springer Verlag, 1990.

[54] R Thom. *Stabilité Structurelle et Morphogénèse*. Inter-Édition, Paris, 1972.

[55] A. Tikhonov, A. Vasil'eva, and A. Sveshnikov. *Differential Equations*. Springer, New York, 1980.

[56] C. Viterbo. *Systèmes dynamiques et équations différentielles*. Ecole Polytechnique, majeure de mathématiques, 2002.

[57] K. Zhou, J.C. Doyle, and K. Glover. *Robust and optimal control*. Prentice Hall, New Jersey, 1996.

Index

arc singulier, 211
autonome, 209

cône polaire, 275
chaîne de Markov, 255
champ de vecteurs, 209
cible, 191
commande
    impulsionnelle, 238
    optimale stochastique, 266
condition
    de stabilité, 246, 269, 273
consistance
    forte, 274
convexité stricte, 197
crochets de Lie, 211

différences finies
    classiques, 270
    généralisées, 273
distance de Hausdorff, 242
dynamique affine, 209

ensemble des états accessibles, 192
Equation HJB, 227
equation HJB
    de la commande stochastique, 268

face de simplexe, 246
feedback, 191
frontière, 194

générateur minimal, 275

hamiltonien, 226

itérations
    sur les stratégies, 261
    sur les valeurs, 241, 242, 260

lieu
    de changement de signe, 191
    singulier, 213
linéarisation non standard, 216

matrices de Hankel, 159

normale extérieure, 198

perturbation
    en aiguille, 217
point
    de Lebesgue, 217
principe
    d'unicité fort, 233
    de programmation dynamique, 225, 237, 239, 242, 258, 259, 262, 268
    du minimum, 198, 210
pseudo-hamiltonien, 196

séparation d'ensembles convexes, 193
schéma décentré, 242
simplexe, 246
solution de viscosité, 231
synthèse, 191

temps
    d'arrêt, 236, 263
    de sortie, 262
transfert en temps minimal, 191

variation
    en aiguille, 217
    finale, 217