L'apprentissage par renforcement. Etude du cas continu.

Rémi Munos
CEMAGREF, LISC, Parc de Tourvoie,
BP 121, 92185 Antony Cedex, FRANCE
Tel : (0)1 40 96 61 79. Fax : (0)1 40 96 61 79
E-mail: Remi.Munos@cemagref.fr

JURY:
Paul Bourgine, Directeur de thèse
Guy Barles, Professeur Université Tours, Rapporteur
Hugues Bersini, Professeur ULB Bruxelles, Rapporteur
Andrew Moore, Professeur CMU USA, Examinateur
Jean Petitot, Directeur d'études EHESS, Examinateur

30 septembre 1997

Table des matières

1 Introduction à l'apprentissage par renforcement **11**

1.1 Qu'est ce que l'Apprentissage par Renforcement ? 12
1.1.1 Introduction 12
1.1.2 Origines des modèles 13
1.2 L'A/R dans le cas discret 14
1.2.1 Les Processus de Décision Markovien 14
1.2.2 La méthode de la Programmation Dynamique 15
1.2.3 L'approche Apprentissage par Renforcement 19
1.2.4 Présentation sommaire du Q-learning 20
1.3 L'A/R dans le cas continu 21
1.3.1 Les approches existantes 21
1.3.2 Les problèmes posés 22
1.3.3 La méthode utilisée dans la thèse 24
1.4 Apports personnels 26
1.5 Plan de la thèse 27

2 Le contrôle optimal **29**

2.1 Le cas déterministe 30
2.1.1 Quelques définitions et propriétés 30
2.1.2 L'équation de Hamilton-Jacobi-Bellman 33
2.1.3 Fonction valeur et contrôle optimal 35
2.1.4 Continuité de la fonction valeur 35
2.2 Introduction aux Solutions de Viscosité 36
2.2.1 Illustration des problèmes rencontrés 36
2.2.2 Définition des solutions de viscosité 39
2.2.3 Quelques propriétés des solutions de viscosité 39
2.3 Approximation par des schémas numériques 41
2.3.1 Introduction 41
2.3.2 Approximation par des méthodes aux différences finies 42
2.3.3 Approximation par des méthodes aux éléments finis . . 46

2.3.4 Convergence des schémas d'approximation 49
2.4 Le cas stochastique 51
2.4.1 La dynamique d'état 51
2.4.2 L'équation de Hamilton-Jacobi-Bellman 53
2.4.3 Approximation par des schémas numériques convergents 55

3 Algorithmes d'apprentissage par renforcement 59
3.1 La méthode générale 60
3.1.1 Théorème de convergence des algorithmes d'A/R . . . 61
3.1.2 Preuve du théorème 63
3.1.3 Comment concevoir des algorithmes convergents ? . . . 66
3.2 Algorithmes indirects 66
3.2.1 Le cas déterministe 67
3.2.2 Le cas stochastique 68
3.3 Algorithmes directs 69
3.3.1 L'algorithme d'A/R aux EF 69
3.3.2 L'algorithme d'A/R aux DF 72

4 Une double dynamique 75

5 Conclusion 77

A Convergence des schémas 79
A.1 Idée principale de la démonstration. 79
A.2 Définition des solutions de viscosité discontinues 79
A.3 V_{sup} et V_{inf} sont des sous- et sur-solutions de viscosité 80
A.4 Preuve du théorème 81

B Convergence algorithme A/R indirect, cas déterministe 83
B.1 Quelques majorations 84
B.1.1 Comparaison des temps $\frac{\delta}{||f(\xi,u)||_1}$ et $\frac{\delta}{\left\|\widetilde{f}(\xi,u)\right\|_1}$ 84
B.1.2 Comparaison des probabilités $p(\xi,\xi',v)$ et $\widetilde{p}(\xi,\xi',v)$ . . 84
B.2 Convergence de l'algorithme 85
B.2.1 Majoration de $\sum_{\xi'}\left[p(\xi,\xi',u)-\widetilde{p}(\xi,\xi',u)\right]V^{\delta}(\xi')$ 86
B.2.2 Convergence de l'algorithme : 87

C Convergence algorithme A/R indirect, cas stochastique 89
C.1 Quelques majorations 90
C.1.1 Majoration de Itô-Taylor 90
C.1.2 Erreur d'approximation du modèle $\widetilde{f}_n$ et $\widetilde{a}_n$ 92

C.1.3 Erreur d'approximation du temps $\widetilde{\tau}_n$ et des probabilités $\widetilde{p}_n$ 95
C.2 Convergence de l'algorithme 95
C.2.1 Majoration de $\sum_{\xi'} [p(\xi, \xi', u) - \widetilde{p}(\xi, \xi', u)] V^\delta(\xi')$ 96
C.2.2 Convergence de l'algorithme : 97

D Convergence algorithme A/R direct, aux éléments finis 99
D.1 Quelques définitions 99
D.2 Quelques propriétés 101
D.2.1 Les points y', y'_1 et η appartiennent à T_{sortie} 101
D.2.2 Comparaison de τ_ξ and τ 103
D.2.3 Majoration de $V^\delta(\eta) - V^\delta(y')$ 104
D.3 Convergence de l'algorithme 106

E Convergence algorithme A/R direct, aux différences finies 109
E.1 Quelques propriétés 109
E.1.1 Comparaison des temps $\frac{\delta}{\|f(\xi,u)\|_1}$ et $\frac{\delta.\tau}{\|y-x\|_1}$ 109
E.1.2 Comparaison des probabilités p_i et $\widetilde{p}_i$: 110
E.2 Convergence de l'algorithme 111
E.2.1 Majoration de $\sum_{i=1}^{d} [p_i - \widetilde{p}_i] V^\delta(\xi^i)$ 112
E.2.2 Preuve de convergence 113

Liste des figures

1.1 *Discrétisation de l'espace d'état O.* La séquence des cases successives X_i traversées par la trajectoire $x(t)$ ne forme pas un PDM. La transition de X vers X_1 ou X_2 selon un contrôle u (se produisant ici pour des trajectoires $x_1(t)$ et $x_2(t)$) dépend des points d'entrée (x_1 et x_2). 25

2.1 L'espace d'état $\overline{O}$. A partir de l'état initial x_0 au temps t_0, le choix d'un contrôle $u(t)$ génère une unique trajectoire $x(t)$ pour $t_0 \leq t \leq \tau$ où τ est le temps de sortie. 31

2.2 La fonction n'est pas differentiable partout 37

2.3 Une solutions généralisée autre que la fonction valeur. 37

2.4 La condition au bord est une inégalité. Ici $V(0) > R(0) = 1$. . 38

2.5 L'espace d'état discrétisé Σ^δ (représenté par les points carrés) et sa frontière $\partial\Sigma^\delta$ (les ronds). 42

2.6 *Interprétation géométrique du schéma aux DF pour le cas horizon temporel infini.* Les probabilitiés de transition $p(\xi, \xi', v)$ du PDM correspondant sont les coordonnées du vecteur $\frac{1}{\delta}\xi\eta$ avec η la projection de ξ sur le segment $(\xi + \delta.e_1, \xi + \delta.e_2)$ parallèlement à $f(\xi, v)$. 45

2.7 Triangulation Σ^δ de l'espace d'état. $V^\delta(x)$ est une combinaison linéaire des $V^\delta(\xi_i)$ pour $x \in$ Simplexe (ξ_0, ξ_1, ξ_2). 47

2.8 Une approximation aux éléments finis. Considérons un sommet ξ de la triangulation et $\eta(\xi, u_1) = \xi + \tau(\xi, u_1).f(\xi, u_1)$. $V^\delta(\eta(\xi, u_1))$ est une combinaison linéaire de $V^\delta(\xi_0), V^\delta(\xi_1)$ et $V^\delta(\xi_2)$ pondérée par les coordonnées barycentriques de $\eta(\xi, u_1)$ dans (ξ_0, ξ_1, ξ_2). 48

2.9 L'espace d'état O, sa discrétisation Σ^δ(les points carrés) et la frontière discrétisée $\partial\Sigma^\delta$ (les ronds). Une trajectoire $x_k(t)$ traverse le voisinage de ξ. 52

3.1 Une trajectoire traversant un simplexe. $\eta(\xi, u)$ est la projection de ξ sur le côté opposé du simplexe. $\frac{y-x}{\lambda_\xi(x)}$est une bonne approximation de $\eta(\xi, u) - \xi$. 70
3.2 Une trajectoire touche la frontière en $y \in \partial O$. 71
3.3 Les grilles G^δ et ∂G^δ recouvrant O sont composées de cases dont les centres ξ forment les ensembles Σ^δ (points carrés) et $\partial\Sigma^\delta$ (les ronds). Deux trajectoires $x^1(t)$ et $x^2(t)$ traversent la case X et ressortent vers les cases X^1 et X^2. 73

D.1 Un simplexe de dimension 3 $(\xi, \xi_1, \xi_2, \xi_3)$ 100
D.2 Notations utiles dans un simplexe de dimension 2. La trajectoire entre en x et sort en $y \in O$. 101
D.3 La trajectoire touche le bord de l'espace d'état : $y \in \partial O$. . . . 102

Remerciements:

Paul Bourgine, Martine Naillon, Guillaume Deffuant, Hugues Bersini, Jean Saulais, Guy Barles, Andrew Moore, Jean Petitot, équipe LISC du CEMAGREF (Bernard, Isabelle, Olivier, Nicolas, Juliette), Pierre Saurel, Mourad Rezgui, équipe Etudes Avancées de DASSAULT-AVIATION (Isabelle Avenas-Payan, Alexis, ...), mes parents, Guylène, ...

Abréviations et notations utilisées

A/R	Apprentissage par renforcement
DF	Différences finies
EF	Eléments finis
EDS	Equation différentielle stochastique
EPD	Equation de programmation dynamique
HJB	Hamilton Jacobi Bellman
PD	Programmation Dynamique
PDM	Processus de Décision Markovien
SV	Solution de viscosité
g^+	Partie positive de la fonction g : $g^+ = \max\{g, 0\}$
g^-	Partie négative de la fonction g : $g^- = \max\{-g, 0\}$
$\|\|y\|\|_1$	La norme 1 du vecteur y : $\|\|y\|\|_1 = \sum_{i=1}^{d} \|y_i\|$

Hypothèses principales

Hypothèse 1 Pour tout $x \in \partial O$ et $t \in [t_0, t_1[$, en notant $\overrightarrow{n}(x)$ la normale sortante à O en x (par exemple, $\overrightarrow{n}(x_2)$ sur la figure 2.1), on a :

Si $\exists u \in U$ avec $f(t,x,u).\overrightarrow{n}(x) \leq 0$,
alors $\exists v \in U$ avec $f(t,x,v).\overrightarrow{n}(x) < 0$

Si $\exists u \in U$ avec $f(t,x,u).\overrightarrow{n}(x) \geq 0$,
alors $\exists v \in U$ avec $f(t,x,v).\overrightarrow{n}(x) > 0$

Hypothèse 2 : Parabolicité uniforme Il existe $c > 0$ tel que : $\forall x \in \overline{O}, \forall u \in U, \forall y \in \mathbb{R}^d$,

$$\sum_{i,j=1}^{d} a_{ij}(x,u) y_i y_j \geq c||y||^2$$

Hypothèse 3 Pour tout $i = 1..d$, on suppose que :

$$a_{ii}(x,u) - \sum_{j \neq i} |a_{ij}(x,u)| \geq 0.$$

Chapitre 1

Introduction à l'apprentissage par renforcement

Comment un système informatique peut-il apprendre par l'expérience ? C'est à dire comment peut-il acquérir des capacités permettant la réalisation d'un objectif, sans être explicitement programmé (soit parce que cela prendrait trop de temps, soit parce que l'on ne dispose pas de solution connue), mais seulement en réalisant des successions de décisions et en apprenant à partir des échecs ou succès constatés ?

La recherche de réponses à cette question nous mène à concevoir non plus des méthodes de résolution spécifiques à tel ou tel problème, mais des algorithmes d'apprentissage permettant de modifier les paramètres internes d'un système en fonction des conséquences de ses prises de décision. Ainsi, ces algorithmes, dits "*d'Apprentissage par Renforcement*" (A/R) fournissent une dynamique d'évolution des compétences d'un système à partir de son expérience et des rétributions observées.

Nous nous intéressons dans cette thèse au cas où :

- l'ensemble des "états" définissant le système est un espace continu (plus précisément sous-ensemble borné de $\mathbb{R}^d$),
- le temps auquel les décisions sont prises est une variable continue.

Dans le cas d'un espace d'état fini, il existe à l'heure actuelle de nombreux résultats théoriques (nottament dans le cadre des Processus de Décision Markovien) permettant d'établir la convergence des algorithmes d'A/R. Ce n'est pas le cas en continu même si de nombreuses applications utilisant des méthodes d'A/R avec des approximateurs de fonctions donnent de très bons résultats. Cependant, rien n'assure la convergence des méthodes employées.

Il nous semble que la difficulté de définir des algorithmes convergents en continu provient des deux problèmes suivants :

- Si l'on utilise des approximateurs de fonctions pour représenter la fonction valeur, les méthodes de descente de gradient peuvent converger vers des solutions très différentes de la solutions optimale donnée par la Programmation Dynamique.
- Si l'on discrétise l'espace d'état et que l'on utilise des méthodes d'A/R sur l'espace discrétisé, on pert le caractère markovien de la succession d'états visités, ce qui entraîne une divergence des valeurs calculées.

Le travail effectué dans cette thèse part du désir de fournir une base théorique à l'A/R en continu afin d'établir des méthodes pour générer des algorithmes convergents. Nous proposons de discrétiser l'espace d'état par des méthodes aux éléments finis ou aux différences finies et de **définir des algorithmes basés sur des propriétés géométriques de la dynamique d'état (à partir de schémas d'approximation convergents) et non sur des propriétés statistiques des successions d'états**.

Dans un premier temps, au paragraphe 1.1, nous faisons une présentation sommaire de l'A/R, ses origines et les courants scientifiques auxquels il se rapporte. Puis nous traitons au paragraphe 1.2 des méthodes d'A/R lorsque l'espace d'état est discret, en présentant les Processus de Décision Markoviens (PDM), diverses méthodes de résolution par la Programmation Dynamique (PD) et un exemple d'algorithme d'A/R, le Q-learning. Enfin, au paragraphe 1.3, nous traitons l'A/R lorsque l'espace d'état est continu par un résumé des méthodes existantes, une analyse des problèmes rencontrés et une présentation de l'approche développée dans cette thèse.

Le paragraphe 1.4 présente les apports personnels de ce travail et le paragraphe 1.5 donne le plan de la thèse.

1.1 Qu'est ce que l'Apprentissage par Renforcement ?

1.1.1 Introduction

Nous considérons un **système artificiel**, ou **agent**, en relation avec un **environnement** selon une certaine dynamique d'interaction agent/environnement ou **dynamique d'état**. À chaque instant, le système a une influence sur l'environnement grâce à l'exécution d'**actions** (ou **décisions** ou **contrôles**), dont les séquences définissent son comportement.

L'information dont dispose le système pour améliorer ses performances est le **renforcement** (ou **récompense**), rétribution occasionnelle et a posteriori de son comportement, qui fournit une évaluation de la pertinence des prises de décisions passées. L'idée est que cette rétribution permet de rectifier son comportement en renforçant les séquences d'actions qui ont mené à des renforcements avantageux.

Ainsi, l'*objectif du système* est de déterminer le comportement optimal, c'est à dire, en tout instant, la décision qui permet d'optimiser l'espérance de renforcement à venir. Il s'agit donc d'un problème de contrôle optimal adaptatif.

> Les *méthodes d'A/R* fournissent des algorithmes qui permettent de modifier les paramètres internes du système en fonction des séquences d'actions effectuées et des renforcements observés, avec pour but de s'approcher au mieux du comportement optimal.

1.1.2 Origines des modèles

Les modèles d'A/R sont issus des travaux en psychologie expérimentale sur le conditionnement pavlovien (par exemple le modèle de Rescorla-Wagner [RW72]), qui ont permis de construire des modèles d'apprentissage basés sur des notions de dérivées temporelles (voir essentiellement le travail de Sutton et Barto [SB81] et [SB90]).

Des exemples d'algorithmes d'A/R sont les "Différences Temporelles" (Temporal Difference TD) de Sutton [Sut88], le Q-learning de Watkins [Wat89], le "Bucket Brigade" utilisé dans les systèmes de règles, voir [Gol89]. Les points communs entre les algorithmes du Q-learning et du Bucket-Brigade sont étudiés dans [DB94], [MP94] ou [Mun92].

Ils ont été développés dans différents courants scientifiques :

- en *Intelligence Artificielle* , courant "*Vie Artificielle*" où ils fournissent des algorithmes d'adaptation pour des systèmes artificiels dits *agents autonomes* ou *animats* (modèles informatiques et robotiques qui s'inspirent de comportements animaux ; voir [Wil87] et les conférences "Simulation of Adaptive Behavior", par exemple [CHMW94])

- en *Robotique* pour la programmation automatique de comportements adéquats (voir par exemple [MC92], [SA94]),

- en *Contrôle Optimal* comme méthode adaptative pour déterminer le contrôle minimisant un coût (par exemple, le problème du pendule inverse étudié, entre autres, par Barto, Sutton et Anderson [Bar90], un

problème de planification temporelle d'ascenseurs [CB96], un problème d'ordonnancement de tâches sur la navette spatiale américaine [ZD95]).

Quelques références

Une bonne introduction à l'A/R peut être trouvée dans [BBS95] où Barto, Bradtke et Singh proposent une étude des diverses méthodes de Programmation Dynamique, introduisent un algorithme de PD en temps réel (Real-Time DP ou RTDP) et font le lien avec les algorithmes d'A/R. Une autre référence complète est l'article de Kaelbling, Littman et Moore [KBW96]. On peut aussi se référer au livre de Bertsekas et Tsitsiklis [BT96] qui traite des méthodes d'A/R utilisant des réseaux neuronaux ou autres approximateurs de fonctions.

1.2 L'A/R dans le cas discret

L'hypothèse de travail pour les méthodes de PD est une indépendance des événements successifs ; le processus est alors dénommé Markovien.

1.2.1 Les Processus de Décision Markovien

Dans le cas discret, la variable temps est discrétisée : $t \in \mathbb{N}$ et l'espace X des états possibles est fini.

À chaque instant, le processus se trouve dans un certain état $x(t)$. Un contrôle $u(t) \in U$ (ou une action ou une décision) est choisi, ce qui entraîne les conséquences suivantes :

- le processus se retrouve dans un nouvel état $x(t+1)$,
- le système reçoit une récompense (ou renforcement) $r(t)$.

L'**hypothèse Markovienne** s'énonce ainsi :

L'état résultant $x(t+1)$ est une variable aléatoire qui ne dépend que de l'état $x(t)$ et du contrôle $u(t)$ à l'instant t.

Ce qui permet de définir des probabilités de transition :

$$p(x, y, u) = prob\,(x(t+1) = x \mid x(t) = y, u(t) = u)$$

De plus, on fait l'hypothèse suivante sur la récompense :

$r(t)$ est une fonction déterministe de l'état $x(t)$, du contrôle $u(t)$ et de l'état résultant $x(t+1)$.

On note $r(x, y, u)$ la récompense reçue lors de la transition de l'état x à l'état y avec le contrôle u et $r(x, u) = \sum_y p(x, y, u).r(x, y, u)$ la somme pondérée des récompenses.

Ainsi, pour tout état initial $x(t = 0)$, le choix d'une séquence de contrôles $u(t)$ permet de générer des séquences d'états $x(t)$ appelées **trajectoires**. Pour chaque état initial x et chaque fonction contrôle $u(t)$ on peut définir une espérance de gain défini,

- dans le cas à horizon temporel fini T, par :

$$J(x; u(t)) = E_{x;u(t)} \left[\sum_{t=0}^{T} r(t) \right]$$

- dans le cas à horizon temporel infini avec récompense pondérée par un facteur de décroissance géométrique $\gamma < 1$, par :

$$J(x; u(t)) = E_{x;u(t)} \left[\sum_{t=0}^{\infty} \gamma^t.r(t) \right].$$

L'objectif du PDM est de déterminer, pour chaque état initial x, le contrôle $u(t)$ qui maximise l'espérance de gain $J(x; u(t))$. Nous utilisons, pour résoudre ce problème de contrôle optimal, la méthode de la Programmation Dynamique.

1.2.2 La méthode de la Programmation Dynamique

La fonction valeur

La méthode de la PD introduite par Bellman [Bel57] nécessite de définir la **fonction valeur** (ou fonction d'utilité, i.e. opposé de la fonction coût), valeur maximale de la fonctionnelle $J(x; u(t))$ en tant que fonction de l'état initial x :

$$V(x) = \sup_{u(t) \in \mathcal{L}(U)} J(x; u(t)). \tag{1.1}$$

Cependant, pour calculer V, on s'apperçoit que l'on doit réaliser une optimisation sur l'espace $\mathcal{L}(U)$ des fonctions contrôle $u(t)$. La méthode de la PD permet de remédier à cela en remplaçant le problème d'optimisation global

sur l'espace fonctionnel $\mathcal{L}(U)$ par un ensemble de problèmes locaux d'optimisation sur l'espace U. En effet, on montre que la fonction valeur V satisfait (dans le cas horizon infini) l'équation suivante :

$$V(x) = \max_{u \in U} \left\{ \gamma \sum_{y \in X} p(x, y, u) V(y) + r(x, u) \right\}. \tag{1.2}$$

appelée *équation de la PD* (EPD) ou *équation de Bellman*.

Réciproquement, l'équation de la PD (1.2) signifie que V est solution de l'équation :

$$W = F[W]$$

avec l'opérateur $F[W](x) = \max_{u \in U} \left\{ \gamma \sum_{y \in X} p(x, y, u) W(y) + r(x, u) \right\}$. D'où V est le point fixe de F, qui vérifie la propriété de contraction suivante :

$$|F[W_1] - F[W_2]| \leq \gamma |W_1 - W_2|.$$

Donc l'équation de la PD (1.2) admet une solution unique qui est la fonction valeur défini par (1.1). Ainsi, dans le cas discret, en résolvant l'EPD (1.2) sur tout l'espace d'état X, on déduit la fonction valeur. Quelques références complètes sur la PD sont les ouvrages de Bertsekas [Ber87], Bertsekas et Tsitsiklis [BT96], Puterman [Put94].

Remarque 1 *Nous verrons au paragraphe 2.2.1 que dans le cas continu, ce résultat n'est plus vrai au sens usuel, et nous devrons faire appel aux notions de solutions de viscosité pour pouvoir obtenir la fonction valeur à partir de l'équation de la PD (appelée dans le cas continu équation de Hamilton-Jacobi-Bellman).*

Le contrôle optimal

Dans le cadre défini précédemment, le contrôle optimal $u^*(t)$ peut se définir en boucle fermé (contrôle en feed-back), c'est à dire en tant que fonction $\pi : X \to U$ (appelée **stratégie**) de l'état courant :

$$u^*(t) = \pi^*(x(t))$$

où π^* est la stratégie optimale. Et grâce à l'équation de la PD (1.2), nous pouvons déduire de la fonction valeur les stratégies optimales suivantes :

$$\pi^*(x) \in \arg\max_{u \in U} \left\{ \gamma \sum_{y \in X} p(x, y, u) V(y) + r(x, u) \right\}, \tag{1.3}$$

ce qui justifie l'intérêt du calcul de V.

Remarque 2 *De manière équivalente, nous pouvons définir les* ***Q-valeurs*** *(ou* ***qualités****) associées à un état et un contrôle :*

$$Q(x,u) = \gamma \sum_{y \in X} p(x,y,u)V(y) + r(x,u)$$

et vérifiant :

$$V(x) = \max_{u \in U} Q(x,u)$$

et qui permettent de définir les stratégies optimales suivantes :

$$\pi^*(x) \in \arg\max_{u \in U} Q(x,u).$$

La connaissance des Q-valeurs est équivalente à celle de la fonction valeur.

Méthodes de résolution de l'équation de la PD

Il existe classiquement deux types de méthodes pour résoudre des équations de PD :

- **Itérations sur les stratégies** (policy iteration) (voir [Ber87]) : méthode basée sur des successions de deux phases :
  - une phase d'évaluation de la fonction valeur à partir de la stratégie courante π_n :

$$V_{n+1}(x) = \gamma \sum_{y \in X} p(x,y,\pi_n(x))V(y) + r(x,\pi_n(x))$$

  - une phase d'itération sur la statégie à partir de la fonction d'évaluation courante V_n :

$$\pi_{n+1}(x) = \arg\max_{u \in U} \left\{ \gamma \sum_{y \in X} p(x,y,u)V_n(y) + r(x,u) \right\}$$

- **Itérations sur les valeurs** (value iteration) : la fonction valeur est obtenue comme limite d'itérations successives (à partir de V_0 on calcule $V_1 = F(V_0)$ puis $V_2 = F(V_1) = F^2(V_0)$ puis ... $V_n = F^n(V_0)$ et l'on sait que d'après la propriété de contraction, la suite V_n converge vers V). Diverses méthodes peuvent être utilisées selon que les valeurs à l'étape $n+1$ sont itérées exclusivement en fonction des valeurs à l'étape n (synchrone) ou bien qu'elles utilisent aussi des valeurs précédemment itérées (de l'étape $n+1$) (Gauss-Seidel ou asynchrone) :

- *méthode synchrone* : La fonction valeur est itéré simultanément pour tous les états (calcul en parallèle) : V_{n+1} dépend exclusivement des valeurs à l'étape n : pour tout $x \in X$,

$$V_{n+1}(x) = \max_{u \in U} \left\{ \gamma \sum_{y \in X} p(x, y, u) V_n(y) + r(x, u) \right\}$$

- *méthode de Gauss-Seidel* : les valeurs sont itérées selon une combinaison séquentielle de tous les états (par exemple $x_1, x_2, ...$), en utilisant pour chaque état les valeurs des autres états calculées précédemment :

$$V_{n+1}(x_i) = \max_{u \in U} \left\{ \gamma \sum_{x_j \in X} p(x, y, u) \widetilde{V}(x_j) + r(x, u) \right\}$$

avec

$$\widetilde{V}(x_j) = \begin{cases} V_{n+1}(x_j), & \text{si } j < i, \\ V_n(x_j), & \text{sinon.} \end{cases}$$

- *méthode asynchrone* : similaire à Gauss-Seidel sauf que les itérations n'ont pas besoin de suivre une organisation séquentielle de tous les états.

Ces méthodes de PD sont dites "**hors-ligne**" (off-line) car les itérations effectuées ne sont pas en relation avec les pas de temps auxquels le processus de décision est résolu. Les méthodes de PD dites "**en-ligne**" (on-line) réalisent une PD asynchrone de manière simultanée avec le processus d'exécution des contrôles ; ainsi, la phase de PD et le processus de contrôle interagissent de la manière suivant :

- Les décisions (le choix des contrôles) sont réalisées en fonction des valeurs les plus récentes calculées par la PD,
- Les séquences d'états, influencées par le contrôle choisi, déterminent la succession des itérations effectuées par la PD.

Les algorithmes résultant d'une telle interaction sont dits de **PD en temps réel** (Real-time DP ou RTDP).

Nous avons le résultat suivant que *la convergence de toutes ces méthodes vers la fonction valeur est assurée pourvu que tous les états et contrôles soient itérés un nombre de fois suffisamment grand* (voir [BT96]).

1.2.3 L'approche Apprentissage par Renforcement

Dans le paragraphe précédent, nous avons vu des méthodes de résolution exactes de l'équation de la PD à partir d'itérations successives (selon diverses méthodes) de la fonction valeur. Cependant, ces méthodes présuppose la connaissance de la dynamique d'état (les probabilités de transitions $p(x, y, u)$) et de la fonction renforcement $r(x, y, u)$. L'approche A/R consiste à supprimer cette hypothèse. Ainsi, les méthodes d'A/R ont pour objectif de s'appliquer lorsque :

- la dynamique d'état et/ou les récompenses sont (au moins partiellement) inconnues du programmeur,
- la dynamique d'état et les récompenses sont connues mais les méthodes de PD sont inefficaces car l'espace d'état est trop grand pour pouvoir effectuer des itérations de manière exhaustive sur l'ensemble des états.

Nous nous intéresserons dans cette thèse au premier cas, c'est à dire lorsque la dynamique d'état et les récompenses sont a priori inconnues du système. Ainsi, au cours de ses interactions avec l'environnement, le système doit combiner :

- une *phase d'exploration* qui permet de découvrir les dynamiques de transition d'état et les récompenses correspondantes,
- une *phase d'exploitation* qui permet d'agir au mieux en fonction des connaissances actuelles.

Le problème consistant à bien gérer les parts respectives de ces deux phases est appelé le *dilemme exploration/exploitation* (voir [Meu96]). Ce problème n'est pas l'objet de la thèse, mais il est complémentaire de notre préoccupation, qui est :

> **Comment intégrer, à partir des données des successions d'états et récompenses, une représentation interne (la fonction valeur) permettant de définir la meilleure stratégie possible (au sens de la convergence vers la solution optimale).**

On peut distinguer deux classes pour les méthodes d'A/R :

- **Les méthodes indirectes** (model-based) : dans un premier temps, le système construit un modèle de la dynamique d'interaction avec son environnement et des renforcements (approximation des $p(x, y, u)$ et $r(x, y, u)$), puis utilise ces valeurs approchées dans des règles itératives de type PD en temps réel ou autres méthodes de PD.

- **Les méthodes directes** (model-free) : le système apprend une stratégie de contrôle optimal directement sans construire de modèle de la dynamique.

Les exemples d'algorithmes d'A/R les plus connus sont le *Q-learning* de Watkins [Wat89] et les différences temporelles TD(λ) de Sutton [Sut88]. Ces algorithmes ont fait l'objet d'études théoriques poussées (voir [Day92], [WD92], [Sau92], [DS94], [JJS94], [Tsi94] et [BT96]) et il existe maintenant diverses preuves de convergence leur assurant une assise solide, ce qui permet de générer des algorithmes plus sophistiqués pour des applications réelles.

Des exemples d'utilisation d'algorithmes d'A/R ayant donné des résultats intéressants pour des applications réelles sont le programme champion de backgammon de Tesauro [Tes92], [Tes95] (cette application donne un exemple d'utilisation d'algorithmes d'A/R pour un espace d'état très grand : de l'ordre de 10^{20} états), un problème d'ordonnancement de tâches pour la navette spatiale américaine de Zhang et Dietterich [ZD95], un problème de planification temporel d'une série d'ascenseurs d'un grand magasin de Crites et Barto [CB96] (ici aussi, l'espace composé de 10^{22} états le rend intraitable par la PD), un problème d'ajustement de fréquences pour des communications téléphoniques traité dans [BT96], etc.

Voici une très brève introduction à l'algorithme Q-learning de Watkins [Wat89].

1.2.4 Présentation sommaire du Q-learning

Nous disposons d'un ensemble de qualités $Q_t(x, u)$ (vérifiant à l'instant t, $V_t(x) = \max_u Q_t(x, u)$). Supposons qu'à l'instant t, le système se trouve dans l'état x, choisisse le contrôle u, et se retrouve à l'instant $t + 1$ dans l'état y en recevant la récompense r. Alors l'algorithme du Q-learning consiste à modifier la qualité $Q_t(x, u)$ selon :

$$Q_{t+1}(x, u) = (1 - \alpha_t)Q_t(x, u) + \alpha_t \left[\gamma V_t(y) + r\right]$$

avec des valeurs α_t, appelées *pas d'apprentissage*, satisfaisant :

$$\sum_{t=0}^{\infty} \alpha_t = \infty \text{ et } \sum_{t=0}^{\infty} \alpha_t^2 < \infty. \tag{1.4}$$

Et la convergence de l'algorithme (V_t tend vers V presque sûrement) est assurée lorsque le nombre d'itérations de tous les états et contrôles tend vers l'infini. L'idée étant que lorsque l'on effectue des transitions d'un état x avec un contrôle u vers les états successeurs $\{y\}$ en recevant les récompenses $\{r\}$ (donc selon les probabilités $p(x, y, u)$), l'algorithme calcule une qualité $Q_t(x, u)$ qui converge (grâce aux propriétés (1.4), voir par exemple [JJS94]) vers la valeur :

$$\gamma \sum_y p(x, y, u) V_t(y) + r(x, u).$$

1.3 L'A/R dans le cas continu

En continu, une telle base théorique (de preuve de convergence d'algorithmes) n'existe pas. Nous avons vu que l'A/R utilise la méthode de la PD qui consiste à construire une approximation de la fonction valeur pour déduire le contrôle optimal. Lorsque l'espace d'état est continu, les questions suivantes apparaissent :

Q1: Comment représenter cette fonction avec un ensemble fini de paramètres ?

Q2: Comment modifier ces paramètres au cours des interactions système / environnement pour que la fonction représentée approche correctement la fonction valeur ?

1.3.1 Les approches existantes

En général en réponse à la première question (*Q1*) on utilise une des deux méthodes de représentation suivantes :

R1: On discrétise a priori l'espace d'état pour ainsi se ramener à un problème discret.

R2: On utilise des approximateurs de fonctions, par exemple :

- des réseaux de neurones (voir [BSA83], [Bar90], [Wil92], [Lin93], [SW93], [Zom94], [Bai95], l'ouvrage de Bertsekas et Tsitsiklis [BT96], les méthodes de Back-Propagation-Through-Time [BG97], [Wer90], les neurones à loi gaussienne de Gullapalli [Gul90], [Gul92], une étude théorique sur des dynamiques d'état stochastiques, voir [Bre95], etc.),

- des fonctions linéaires, multilinéaires (voir [Dav96]) ou polynomiales par morceaux,
- des ensembles flous (voir [Now95]),
- des fonctions à codage économes (les clusters de Mahadevan et Connell [MC92], les systèmes à agrégation d'états [SJJ94], les grilles CMAC se chevauchant de Sutton [Sut96]),
- des discrétisations selon des résolutions variables de Moore [Moo91], [Moo94], etc.

La réponse à la deuxième question (*Q2*) consiste généralement à utiliser une des deux méthodes suivantes :

M1: On considère le problème comme étant discret et on utilise des algorithme basés sur des méthodes d'A/R en discret (par exemple le Q-learning développé au paragraphe précédent).

M2: On utilise des méthodes de descente de gradient.

La représentation *R1* est utilisée avec la méthode *M1*. Les représentations *R2* sont utilisées soit avec la méthode *M1*, soit avec *M2*. Étudions les problèmes qui apparaissent de ces différentes combinaisons.

1.3.2 Les problèmes posés

Combinaison $R2 \times M2$:

Il apparaît que pour différentes raisons (voir les travaux de Boyan et Moore [BM95], de Baird [Bai95], de Gordon [Gor95] et de Bertsekas et Tsitsiklis [BT96]), l'utilisation d'approximateurs de fonctions (par exemple les réseaux neuronaux) pour résoudre des problèmes de PD pose des problèmes fondamentaux de convergence (même si certaines applications donnent de bon résultats). Une raison majeure peut être exprimée ainsi :

> **Problème 1** : l'équation de la PD définit la fonction valeur de façon implicite alors qu'un approximateur de fonction apprend une fonction définie explicitement.

En effet, les approximateurs de fonctions apprennent une fonction V définie *explicitement*, c'est à dire à partir de la donnée des couples $(x, V(x))$ (dans les réseaux neuronaux, x est l'entrée, $V(x)$ est la sortie désirée). Or l'équation de la PD (en discret, l'équation (1.2), en continu, l'équation (2.10)

plus bas) est *implicite*, c'est à dire que la valeur $V(x)$ en un point x est donnée en tant que fonction des valeurs $V(y)$ aux points voisins y (en continu, $V(x)$ satisfait une équation différentielle).

Diverses approches ont été développées pour palier à cet inconvénient dont l'algorithme *Residual-Gradient* de Baird [Bai95] (combiné à l'*Advantage Updating* de Baird [Bai93], algorithme équivalent au Q-learning, mais utilisable lorsque l'on passe à un espace d'état continu) et essentiellement les algorithmes de descente de gradient développés dans le livre de Bertsekas et Tsitsiklis [BT96]. Cependant, étant donné un système d'approximation de fonctions, nous n'avons pas la certitude que la fonction valeur (a priori inconnue) appartient à la classe des fonctions pouvant être approchées par le système d'approximation en question. De plus, les méthodes de descente de gradient, qui tentent de résoudre le problème en minimisant une fonction erreur, souffrent du fait que :

> **Problème 2** : les méthodes de descente de gradient permettent de converger vers des optima locaux ou globaux (dans les meilleurs cas), c'est à dire vers des fonctions W minimisant une certaine fonction erreur (par exemple $E = \int_{x \in O} [H(x, W, DW)]^2 \, dx$ avec l'hamiltonien H défini par (2.13) au paragraphe 2.1.2) ; or, dans le cas des problèmes de PD, on montre qu'il existe une infinité d'optima globaux (les solutions satisfaisant l'équation de Hamilton Jacobi Bellman presque partout, puisqu'elles fournissent une erreur $E = 0$), qui peuvent être très différents de la fonction valeur cherchée.

En effet, on montre que la fonction valeur satisfait l'équation de la PD, appelée équation de Hamilton-Jacobi-Bellman, qui est une équation différentielle non-linéaire possédant une infinité de solutions généralisées (c'est à dire définies presque partout) autres que la fonction valeur (ce problème est abordé au chapitre 2, paragraphe 2.2.1, en particulier à la remarque 8). Ces solutions généralisées W fournissent une erreur E nulle (puisqu'elles vérifient $H(x, W, DW) = 0$ presque partout), donc ce sont des optima globaux vers lesquels une méthode de descente de gradient peut converger.

Or ces fonctions peuvent être très différentes de la fonction valeur recherchée (voir l'exemple donnée par la figure 2.3), menant ainsi à un contrôle très différent du contrôle optimal. **Ce problème fondamental, concernant les méthodes de PD avec fonctions paramétrées utilisant des algorithmes de descente de gradient, n'est pas à notre connaissance résolu.**

Combinaisons $R1 \times M1$ ou $R2 \times M1$:

On montre que la combinaison $R1 \times M1$ ne permet pas de générer des algorithmes convergents (donc à plus forte raison la combinaison $R2 \times M1$ non plus) à cause du problème suivant :

> **Problème 3** : Lorsque l'espace d'état est discrétisé (par exemple selon des cases), **le processus constitué des successions des cases traversées par les trajectoires n'est pas markovien** ; donc les résultats de convergence des algorithmes d'A/R en discret définis au paragraphe 1.2.1, qui nécessitent l'hypothèse Markovienne, ne s'appliquent pas.

Ce problème, décrit dans [Mun97b], apparaît lorsqu'on discrétise l'espace d'état, par exemple selon une grille régulière composée de cases X (voir la figure 1.1). En effet, considérons une trajectoire $x(t)$ et la séquence des cases X_i qu'elle traverse : $x(t) \in X_i$ pour $t \in [t_i, t_{i+1}]$. Supposons que le contrôle u est conservé constant à l'intérieur d'une case X. Nous observons que la transition de la case X vers la case adjacente X_i ne dépend pas seulement de la case X et du contrôle u mais aussi de l'endroit d'où la trajectoire est entrée en X. Ainsi, la séquence des cases traversées $(X, u) \rightarrow X_i$ n'est pas un processus de décision markovien, *même lorsque le processus continu est déterministe.*

Ainsi, si l'on utilisait l'algorithme du Q-learning avec des qualités $Q_n(X_i, u)$ attribuées aux cases X_i en utilisant une règle itérative du style :

$$Q_{n+1}(X, u) = (1 - \alpha_n) Q_n(X, u) + \alpha_n \left[\gamma^{\tau_n} V_n(X_i) + \tau_n . r(X)\right]$$

(avec un certain temps τ_n et un pas d'apprentissage α_n), on calculerait une qualité $Q_n(X, u)$, fonction des valeurs des cases suivantes $V_n(X_i)$ qui dépendrait des fréquences d'occurrence des cases de sortie X_i. Par exemple, supposons, à partir de la figure 1.1, que la plupart des trajectoires entrent en X par son côté bas (donc selon une certaine stratégie d'exploration), sortant ainsi par son côté droit. Alors la qualité $Q(X, u)$ dépendrait presque exclusivement de $V(X_1)$. Pour une autre stratégie d'exploration (entrée dans X par le côté gauche), cette même qualité pourrait ne dépendre que de $V(X_2)$, ce qui donnerait une toute autre valeur de $Q(X, u)$. Ainsi, les valeurs calculées par ce genre d'algorithme dépendraient des stratégies d'exploration choisies et donc ne convergeraient pas vers la fonction valeur (qui est unique).

1.3.3 La méthode utilisée dans la thèse

Au lieu de calculer les qualité $Q_n(X, u)$ en fonction des fréquences d'occurrence des cases de sortie X_i, on va utiliser les propriétés

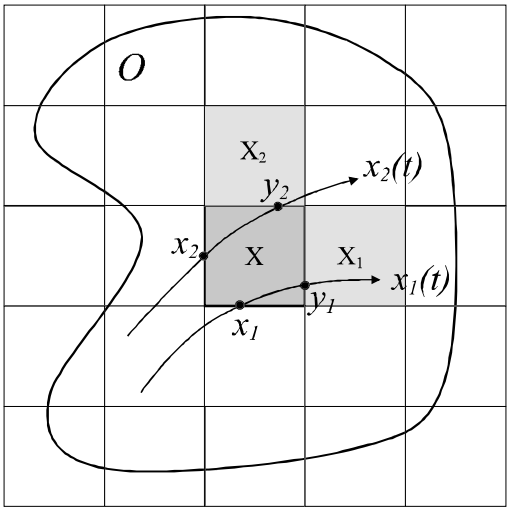

Figure 1.1: *Discrétisation de l'espace d'état* O. La séquence des cases successives X_i traversées par la trajectoire $x(t)$ ne forme pas un PDM. La transition de X vers X_1 ou X_2 selon un contrôle u (se produisant ici pour des trajectoires $x_1(t)$ et $x_2(t)$) dépend des points d'entrée (x_1 et x_2).

géométriques de la dynamique d'état f. Pour cela, nous utilisons les *schémas d'approximation* de Kushner [Kus90], [KD92] qui permettent de discrétiser le problème de contrôle en continu (grâce à des méthodes aux différences finies (DF) et aux éléments finis (EF)) en un Processus de Décision Markovien (PDM).

Les probabilités de transition $p(X, X_i, u)$ de la case X à la case X_i avec un contrôle u de ce PDM peuvent être interprétées, pour les méthodes aux DF, comme les composantes normalisées $\frac{|f_i(\xi,u)|}{\sum_j |f_j(\xi,u)|}$ du vecteur dynamique d'état $f(\xi, u)$ (avec ξ le centre de la case X), et pour les méthodes aux EF, comme des coordonnées barycentriques.

La résolution de ce PDM par des méthodes d'A/R nécessite de faire des approximations sur la dynamique d'état (nous rappelons que dans l'approche A/R, celle-ci est inconnue), par exemple lorsqu'une trajectoire $x_1(t)$ traverse X (voir figure 1.1), nous pouvons approcher $f(\xi, u)$ par $\frac{y_1 - x_1}{\tau_1}$ (où τ_1 est le temps de passage dans la case).

Les algorithmes ainsi générés ont la forme de règles itératives de PD portant sur les valeurs associées à l'espace discretisé (les cases, ou leur centre, pour les méthodes aux DF, les sommets de la triangulation pour les méthodes aux EF). Ainsi en réponse aux questions $Q1$ et $Q2$, nous proposons :

Réponse à $Q1$**:** La fonction est représentée par une discrétisation de l'es-

pace d'état afin d'utiliser des méthodes aux EF (construction de fonctions linéaires par morceaux) ou aux DF (construction de fonctions constantes par morceaux),

Réponse à $Q2$: Les valeurs des états de l'espace discrétisé sont modifiées par des règles itératives de PD basées sur des schémas d'approximation convergents,

ce qui permet de générer des algorithmes d'A/R convergents.

1.4 Apports personnels

Les contributions personnelles à ce travail sont les suivantes :

- Utilisation du formalisme du contrôle optimal, dans le cas déterministe et stochastique (chapitre 2), pour décrire la problématique posée par l'A/R, en particulier :
  - la formulation de l'équation de Hamilton-Jacobi-Bellman,
  - l'utilisation de schémas d'approximation convergents,
  - l'utilisation des solutions de viscosité.
- La méthode générale de construction d'algorithmes d'A/R, basée sur le théorème 7 (chapitre 3) dont la preuve est donnée au paragraphe 3.1.2.
- Les exemples d'algorithmes d'A/R utilisant les méthodes
  - *indirectes*, dans le cas
    - *déterministe* (théorème 8 décrit au paragraphe 3.2.1 dont la preuve est donnée en annexe B),
    - *stochastique* (théorème 9 au paragraphe 3.2.2 dont la preuve est en annexe C)
  - *directes* basées sur des méthodes
    - *aux éléments finis* (théorème 10 au paragraphe 3.3.1, preuve en annexe D),
    - *aux différences finies* (théorème 11 au paragraphe 3.3.2, preuve en annexe E).
- L'utilisation d'une double dynamique d'apprentissage et de modification structurelle (de l'espace discret) au chapitre 4.

1.5 Plan de la thèse

Le *chapitre 2* introduit le formalisme du contrôle optimal utilisé dans la thèse.

Dans un premier temps, nous étudions le cas déterministe (lorsque la dynamique d'état est défini par une équation différentielle). À la section 2.1.1, nous donnons, pour les deux cas horizon temporel fini et infini, la définition de la fonction valeur V, l'équation de Hamilton-Jacobi-Bellman (HJB) et une propriété de continuité pour V. La section 2.2 illustre les problèmes rencontrés lors de l'intégration de l'équation de HJB et introduit la notion de solution de viscosité ainsi que quelques résultats importants. La section 2.3 présente les méthodes d'approximation de la fonction valeur par des schémas numériques basés sur des méthodes aux différences finies (DF) et aux éléments finis (EF) et fournit un théorème général de convergence de tels schémas.

Puis nous étudions à la section 2.4 le cas stochastique (lorsque la dynamique d'état est défini par une équation différentielle stochastique).

Le *chapitre 3* décrit une méthode générale pour construire des algorithmes d'A/R convergents (section 3.1) et fournit des exemples d'algorithmes utilisant des techniques indirectes (section 3.2) et directes (section 3.3).

Le *chapitre 4* propose de combiner la dynamique d'apprentissage défini par les algorithmes précédents à une dynamique structurelle permettent d'affiner la discrétisation au cours de l'apprentissage et d'utiliser des méthodes à résolution variables.

L'*annexe A*, donnée à titre indicative, reprend la démonstration de Barles et Souganidis [BS91] afin de démontrer la convergence des schémas d'approximation.

Les *annexes B, C, D* et *E*, contributions majeures de ce travail, fournissent les preuves de convergence des algorithmes définis au chapitre 3.

Chapitre 2

Le contrôle optimal

Ce chapitre présente le formalisme du contrôle optimal dans l'optique apprentissage par renforcement. Les sections 2.1, 2.2 et 2.3 concernent le cas déterministe, c'est à dire lorsque la dynamique d'état est régie par une équation différentielle. La section 2.4 présente le cas stochastique (la dynamique d'état régie par une équation différentielle stochastique).

Pour le cas déterministe, nous définissons au paragraphe 2.1.1 la fonctionnelle de renforcement, ou gain, ainsi que la fonction valeur V, valeur maximale du gain en tant que fonction des données initiales. Nous énonçons au paragraphe 2.1.2 l'équation de Hamilton-Jacobi-Bellman (HJB), ou équation de PD, qui est une équation différentielle non-linéaire du premier ordre portant sur V, et donnons au paragraphe 2.1.4 un résultat de continuité pour V, conséquence d'un résultat d'unicité forte obtenu à partir de l'hypothèse 1 (paragraphe 2.1.4).

Afin de résoudre l'équation de HJB, nous introduisons à la section 2.2 la notion de solutions de viscosité (SV) et donnons, par le théorème 4, un résultat d'identification de la solution de viscosité avec la fonction valeur.

La section 2.3 est consacrée à des méthodes d'approximation numériques de la fonction valeur à partir de schémas basés sur des méthodes aux différences finies (paragraphe 2.3.2) et aux éléments finis (paragraphe 2.3.3) et dont la convergence est énoncée au paragraphe 2.3.4 et prouvée en annexe A.

Pour le cas stochastique, abordé à la section 2.4, on définit de même la fonction valeur V, valeur maximale de l'espérance de gain, on énonce l'équation de HJB (qui est ici du second ordre) au paragraphe 2.4.2 et donnons un exemple de schéma d'approximation numérique au paragraphe 2.4.3.

2.1 Le cas déterministe

2.1.1 Quelques définitions et propriétés

Nous considérons un système déterministe contrôlé défini à l'instant t par son *état* $x(t) \in \bar{O}$, où $\bar{O}$, la fermeture de O, ouvert borné de $\mathbb{R}^d$, est l'*espace d'état*. L'évolution du système au cours du temps, c'est à dire sa **dynamique d'état**, est déterminée par une équation différentielle contrôlée :

$$\frac{d}{dt}x(t) = f(t, x(t), u(t)) \tag{2.1}$$

où le *contrôle* $u(t)$ est une fonction mesurable, au sens de Lebesgue, bornée, et à valeurs dans un compact U appelé *espace des contrôles admissibles*.

Nous supposons que la dynamique d'état est Lipschitzienne par rapport à la variable d'état : $\exists L_f > 0$ tel que :

$$\|f(t, x, u) - f(t, y, u)\| \leq L_f . \|x - y\| \tag{2.2}$$

pour tout t, x, y et $u \in U$. De plus, nous supposons que la frontière ∂O de O est de classe $\mathcal{C}^2$.

Ainsi, à partir d'un état initial x, le choix d'une fonction contrôle $u(t)$ permet, en intégrant (2.1), de générer une unique **trajectoire** $x(t)$.

Nous allons considérer deux classes possibles de problèmes : les problèmes à *horizon temporel fini* et ceux à *horizon temporel infini avec renforcement à décroissance géométrique*.

Horizon temporel fini

Ici nous considérons le cas d'un horizon temporel $t_1 < \infty$ donné. Soit t_0 l'instant initial. Pour un état initial x_0 au temps t_0, le choix d'une fonction contrôle $u(t)$ génère la trajectoire $x(t)$ (voir figure 2.1).

Considérons le cylindre $Q = [t_0, t_1[\times O$, sa fermeture $\overline{Q} = [t_0, t_1] \times \overline{O}$ et sa frontière :

$$\partial Q = ([t_0, t_1[\times \partial O) \cup (\{t_1\} \times \overline{O})$$

On appelle $([t_0, t_1[\times \partial O)$ sa *frontière latérale* et $(\{t_1\} \times \overline{O})$ sa *frontière terminale*.

Nous définissons le *temps de sortie* de $(t, x(t))$ de $\overline{Q}$:

$$\tau = \inf_{u(.)} \left\{ t, (t, x(t)) \notin \overline{Q} \right\}.$$

Ce qui est équivalent à :

$$\tau = \begin{cases} \inf_{u(.)} \left\{ t, x(t) \notin \overline{O} \right\} \\ t_1 \text{ si } x(t) \in \overline{O} \text{ pour tout } t \in [t_0, t_1] \end{cases}$$

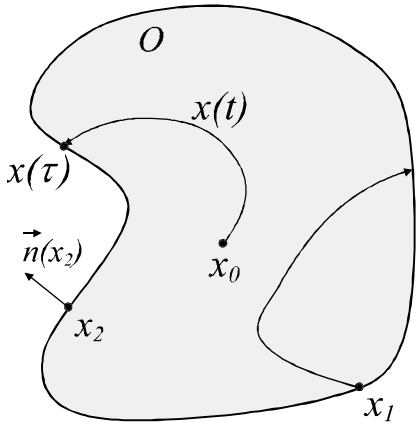

Figure 2.1: L'espace d'état $\overline{O}$. A partir de l'état initial x_0 au temps t_0, le choix d'un contrôle $u(t)$ génère une unique trajectoire $x(t)$ pour $t_0 \leq t \leq \tau$ où τ est le temps de sortie.

Définition 1 (Fonctionnelle de renforcement) *Pour tout état initial x_0, t_0 et pour toute fonction contrôle $u(t)$ pour $t_0 \leq t \leq \tau \leq t_1$, on définit la fonctionnelle de renforcement :*

$$J(t_0, x_0; u(t)) = \int_{t_0}^{\tau} r(t, x(t), u(t))dt + R(\tau, x(\tau)) \tag{2.3}$$

où :

$$R(t,x) = \begin{cases} R_l(t,x) & si\ (t,x) \in [t_0, t_1) \times \partial O \quad (renforcement\ latéral) \\ R_t(x) & si\ t = t_1 \quad (renforcement\ terminal) \end{cases}$$

On appelle $r(t,x,u)$ le ***renforcement courant****, $R_l(t,x)$ le* ***renforcement latéral*** *(qui apparaît lorsque le système sort de $\bar{O}$ avant l'instant t_1) et $R_t(x)$ le* ***renforcement terminal*** *(qui apparaît lorsque le système reste dans $\bar{O}$ jusqu'à l'instant t_1).*

Nous supposons que les fonctions renforcement sont Lipschitziennes par rapport à la variable d'état : $\exists L_r, L_{R_l}, L_{R_t}$ positives telles que :

$$\begin{aligned} |r(t,x,u) - r(t,y,u)| &\leq L_r . \|x - y\| \\ |R_l(t,x) - R_l(t,y)| &\leq L_{R_l} . \|x - y\| \\ |R_t(x) - R_t(y)| &\leq L_{R_t} . \|x - y\| \end{aligned} \tag{2.4}$$

pour tout t, x, y et u dans leur domaine de définition.

Le problème de contrôle optimal consiste à trouver pour toutes conditions initiales x_0, t_0 le contrôle $u(t)$ qui optimise la fonctionnelle de renforcement $J(t_0, x_0; u(t))$.

Remarque 3 *Contrairement au cas discret, dans le cas continu, on doit considérer trois fonctions renforcement distinctes: r est obtenu et accumulé le long de la trajectoire, R_l apparaît si la trajectoire sort de l'espace d'état et R_t correspond au renforcement à l'instant terminal. Ce formalisme permet de traiter de nombreux problèmes de contrôle optimal.*

Pour déterminer le contrôle optimal, nous utilisons la méthode de la PD qui nécessite l'introduction d'une fonction d'anticipation, appelée fonction valeur :

Définition 2 (Fonction Valeur) *La fonction valeur est la valeur maximale de la fonctionnelle de renforcement (2.3) en tant que fonction des données initiales:*

$$V(t, x) = \sup_{u(.)} J(t, x; u(.)) \tag{2.5}$$

Horizon temporel infini

Ici nous considérons un horizon temporel infini ($t_1 = \infty$). Afin d'éviter des problèmes de divergence, on considère généralement le cas d'un renforcement actualisé pour lequel, les fonctions renforcement satisfont une propriété de décroissance temporelle exponentielle :

$$r(t, x, u) = \gamma^t.r(x, u)$$
$$R_l(t, x) = \gamma^t.R(x)$$

avec un *facteur de décroissance* $\gamma \in [0, 1[$ (en économie, $\ln \frac{1}{\gamma}$ est appelé le taux d'actualisation). De plus nous supposons que la dynamique d'état f est indépendante du temps (f est autonome).

Considérons un état initial x à l'instant $t = 0$; le choix d'un contrôle $u(t)$ génère la trajectoire $x(t)$. Soit τ le temps de sortie de $x(t)$ de $\bar{O}$ (avec la convention que si $x(t)$ reste indéfiniment dans $\bar{O}$, alors $\tau = \infty$). De manière similaire, on définit la fonctionnelle de renforcement et la fonction valeur :

Définition 3 (Fonctionnelle de renforcement actualisé) *Pour tout état x et tout contrôle $u(t)$ on définit la fonctionnelle de renforcement actualisé :*

$$J(x; u(t)) = \int_0^\tau \gamma^t r(x(t), u(t))dt + \gamma^\tau R(x(\tau)) \tag{2.6}$$

Où r est le renforcement courant et R le renforcement final.

De même qu'au paragraphe précédent, r et R sont supposées Lipschitziennes avec des constantes de Lipschitz L_r et L_R.

Définition 4 (Fonction valeur) *La fonction valeur est la valeur maximale de la fonctionnelle de renforcement actualisé, en tant que fonction de l'état initial x au temps $t = 0$:*

$$V(x) = \sup_{u(.)} J(x; u(.)) \tag{2.7}$$

Remarque 4 *L'étude théorique et les approximations numériques de la fonction valeur sont de grande importance dans la méthode de la PD puisqu'à partir de V, on peut déduire le contrôle optimal (voir paragraphe 2.1.3).*

Le paragraphe suivant établit que la fonction valeur satisfait une propriété locale : une équation différentielle non-linéaire du premier ordre, appelée *équation de Hamilton-Jacobi-Bellman.*

2.1.2 L'équation de Hamilton-Jacobi-Bellman

A partir du principe de PD (introduit par Bellman dans [Bel57]), on peut montrer que la fonction valeur satisfait une condition locale, appelée équation de Hamilton-Jacobi-Bellman (voir [FS93] ou [KD92], pour une étude approfondie). Dans le cas déterministe étudié ici, il s'agit d'une équation différentielle non-linéaire du premier ordre (dans le cas stochastique nous verrons à la section 2.4 qu'on obtient une équation analogue du second ordre).

Théorème 1 (Hamilton-Jacobi-Bellman) *On considère les deux cas suivants :*

- ***Le cas à horizon temporel fini.*** *Si la fonction valeur V est différentiable en $(t, x) \in [t_0, t_1[\times O$, soit $V_t(t, x)$ sa dérivée temporelle et $D_x V(t, x)$ son gradient ; alors l'équation de Hamilton-Jacobi-Bellman :*

$$V_t(t, x) + \sup_{v \in U} [D_x V(t, x).f(t, x, v) + r(t, x, v)] = 0 \tag{2.8}$$

a lieu en (t, x). De plus, la fonction valeur satisfait les conditions au bord suivantes :

$$V(t, x) \geq R_l(t, x) \text{ si } (t, x) \in [t_0, t_1[\times \partial O \tag{2.9a}$$
$$V(t, x) = R_t(x) \text{ si } t = t_1 \tag{2.9b}$$

- ***Le cas à horizon temporel infini.*** *Si la fonction valeur V est différentiable en $x \in O$, soit $DV(x)$ son gradient ; alors l'équation de Hamilton-Jacobi-Bellman :*

$$V(x)\ln\gamma + \sup_{v\in U}[DV(x).f(x,v) + r(x,v)] = 0 \qquad (2.10)$$

a lieu en x. De plus, la fonction valeur satisfait les conditions au bord suivantes :

$$V(x) \geq R(x) \text{ pour } x \in \partial O \qquad (2.11)$$

Remarque 5 *Les conditions au bord (2.9a) et (2.11) sont des inégalités. En effet, pour un point sur la frontière (par exemple pour $x_1 \in \partial O$, figure 2.1) il peut exister un contrôle $u(t)$ pour lequel la trajectoire rentre à l'intérieur de $\overline{O}$ et dont la fonctionnelle de renforcement est supérieure au renforcement ($R_l(t,x_1)$ ou $R(x_1)$) obtenu en sortant immédiatement de $\overline{O}$. Dans de tels cas, les inégalités (2.9a) et (2.11) sont strictes.*

De manière équivalente, l'équation de HJB (2.8) (respectivement (2.10)) signifie que V est solution de l'équation :

$$H(t,x,W,DW) = 0 \qquad (2.12)$$

avec l'Hamiltonien H défini, dans le cas à horizon temporel fini, par :

$$H(t,x,W,DW) = -W_t(t,x) - \sup_{v\in U}[D_xW(t,x).f(t,x,v) + r(t,x,v)]$$

(respectivement, dans le cas à horizon temporel infini, par :

$$H(x,W,DW) = -W(x)\ln\gamma - \sup_{v\in U}[DW(x).f(x,v) + r(x,v)]\Big) \qquad (2.13)$$

pour toute fonction différentiable W.

Remarque 6 *En tout point, la résolution de l'équation de HJB consiste en une optimisation d'une fonction (calcul d'une borne supérieure sur l'ensemble des contrôle admissibles U). Ainsi, le problème de l'optimisation de la fonctionnelle J (problème très complexe puisqu'il s'agit du calcul d'une borne supérieure sur l'espace fonctionnel constitué de l'ensemble des fonctions contrôle $u(t)$ possibles) est remplacé par un problème d'optimisation de fonction en tout point de l'espace d'état.*

Le paragraphe suivant illustre la relation entre la fonction valeur et le contrôle optimal.

2.1.3 Fonction valeur et contrôle optimal

La méthode de la PD consiste à calculer la fonction valeur afin de générer le contrôle optimal sous forme d'une **stratégie de contrôle en boucle fermée** (feed-back), c'est à dire une application $\pi(t,x) : \overline{Q} \to U$ (pour le cas horizon fini) telle que le contrôle optimal $u^*(t)$ à l'instant t soit exprimé en fonction de l'état courant $x(t)$:

$$u^*(t) = \pi(t, x(t)).$$

A partir de l'équation HJB (2.8) (respectivement (2.10)), on déduit que la connaissance de la fonction valeur mène à la stratégie optimale de contrôle suivante :

$$\pi^*(t,x) \in \arg\sup_{v\in U}[D_x V(t,x).f(t,x,v) + r(t,x,v)]$$

(respectivement à

$$\pi^*(x) \in \arg\sup_{v\in U}[DV(x).f(x,v) + r(x,v)]\Big). \tag{2.14}$$

2.1.4 Continuité de la fonction valeur

Ici nous donnons un résultat de continuité pour la fonction valeur sous l'hypothèse suivante concernant la dynamique d'état f au voisinage de la frontière ∂O :

Hypothèse 1 *Pour tout $x \in \partial O$ et $t \in [t_0, t_1[$, en notant $\overrightarrow{n}(x)$ la normale sortante à O en x (par exemple, $\overrightarrow{n}(x_2)$ sur la figure 2.1),*

Si $\exists u \in U$ avec $f(t,x,u).\overrightarrow{n}(x) \leq 0$, alors $\exists v \in U$ avec $f(t,x,v).\overrightarrow{n}(x) < 0$
Si $\exists u \in U$ avec $f(t,x,u).\overrightarrow{n}(x) \geq 0$, alors $\exists v \in U$ avec $f(t,x,v).\overrightarrow{n}(x) > 0$

Théorème 2 (Continuité de la fonction valeur) *Supposons que l'hypothèse 1 est vérifiée, alors la fonction valeur est continue dans $\overline{Q}$.*

Une preuve de ce théorème, basée sur un résultat d'unicité forte (voir paragraphe 2.2.3, lemme 1), peut être trouvée dans [BP90].

Remarque 7 *Pour le cas à horizon infini, la fonction valeur est continue sur $\overline{O}$ pour les mêmes hypothèses que précédemment avec $f(x,u)$ au lieu de $f(t,x,u)$.*

2.2 Introduction aux Solutions de Viscosité

Le théorème 1 énonce que si la fonction valeur est différentiable, alors elle vérifie l'équation de HJB. Cependant, en général, la fonction valeur n'est pas différentiable partout même lorsque les données du problème sont très régulières. Nous ne pouvons donc pas espérer trouver de solutions classiques (c'est à dire vérifiant HJB partout). Maintenant, si nous recherchons des solutions généralisées (c'est à dire vérifiant HJB presque partout), on s'apperçoit que l'on obtient une infinité de solutions autres que la fonction valeur.

Ainsi, on doit définir une formulation faible des solutions de cette équation. Crandall et Lions ont introduit dans leur article [CL83] une telle formulation en définissant la notion de *Solutions de Viscosité*. Quelques références générales et complètes de ce domaine sont [CIL92], [Bar94] et pour une application au contrôle optimal, [FS93]). Cette notion a été développée pour une grande classe d'équations différentielles non linéaires du premier et second ordre (dont les équations de HJB pour le contrôle déterministe, stochastique, les processus de diffusion contrôlés).

Les propriétés les plus importantes des solutions de viscosité sont le fait qu'il existe une unique solution de viscosité de l'équation de HJB (2.8) avec les conditions aux limites (2.9a) et (2.9b) qui est la fonction valeur du problème de contrôle correspondant ainsi que des résultats de stabilité des solutions d'équations approchées (résultat important qui nous permet d'établir la convergence de schémas d'approximation au paragraphe 2.3).

Tout d'abord, voici une illustration, par un exemple très simple, des problèmes rencontrés lors de l'intégration de l'équation de HJB et de la nécessaire introduction de la notion de solutions de viscosité.

2.2.1 Illustration des problèmes rencontrés

Considérons un problème de contrôle très simple en une dimension. Soit l'état $x(t) \in [0, 1]$. Le contrôle est $u(t) \in \{-1, +1\}$ et la dynamique d'état :

$$\frac{dx}{dt} = f(x, u) = u.$$

Nous considérons le problème à horizon infini avec renforcement à décroissance géométrique avec un facteur γ. Le renforcement courant est $r = 0$. Le renforcement final est : $\{R(0), R(1)\}$. Ainsi, l'équation de HJB s'écrit :

$$V(x) \ln \gamma + \max \{V'(x), -V'(x)\} = 0 \tag{2.15}$$

avec les conditions au bord :

$$V(0) \geq R(0) \text{ et } V(1) \geq R(1). \tag{2.16}$$

Ici, nous pouvons calculer explicitement la fonction valeur qui vaut :

$$V(x) = \max\left\{R(0).\gamma^x, R(1).\gamma^{1-x}\right\}$$

La fonction valeur n'est pas différentiable partout

Prenons $R(0) = 1$ et $R(1) = 2$ avec $\gamma = 0.3$. La fonction valeur correspondante est représentée sur la figure 2.2. On observe que V n'est pas différentiable partout et donc ne vérifie pas l'équation de HJB (2.15) partout.

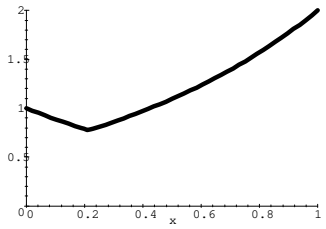

Figure 2.2: La fonction n'est pas differentiable partout

Plusieurs solutions généralisées

Si nous cherchons des solutions généralisées qui satisfont l'équation de HJB (2.15) presque partout, nous trouvons une infinité de fonctions autres que la fonction valeur. Un exemple de fonction vérifiant (2.15) presque partout avec les conditions au bord $R(0) = 1$ et $R(1) = 2$ est représenté sur la figure 2.3. Ainsi, il existe une infinité de solutions généralisées autres que la fonction valeur.

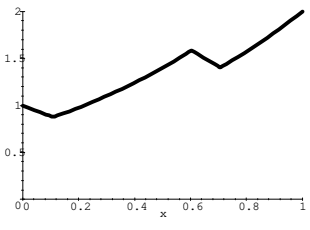

Figure 2.3: Une solutions généralisée autre que la fonction valeur.

Remarque 8 *Le fait qu'il existe une infinité de fonctions W vérifiant l'équation de HJB (2.12) presque partout et différentes de la fonction valeur implique l'inefficacité des méthodes de descente de gradient pour approcher la fonction valeur à partir d'un système d'approximation de fonctions. En effet supposons que l'algorithme de descente de gradient cherche à minimiser une fonction erreur de type :*

$$E(W) = \int_{x \in O} [H(x, W, DW)]^2 \, dx$$

avec l'Hamiltonien H défini par (2.13) au paragraphe 2.1.2, alors la méthode convergera, dans les meilleurs cas, vers une solution généralisée W de (2.12) (puisque ces fonctions sont des optima globaux du problème de minimisation de l'erreur, du fait qu'elles vérifient $E(W) = 0$), donc différente de la fonction valeur V. Le contrôle défini par une telle fonction W (selon la stratégie en boucle fermée (2.14)) peut être très différent du contrôle optimal (défini par V). Par exemple sur la figure 2.3, la fonction représentée engendre un contrôle (donné par la direction du gradient) très différent du contrôle optimal, obtenu par la fonction V représentée sur la figure 2.2.

Le problème de l'inégalité des conditions au bord

Voici une illustration du fait que la condition au bord (2.16) peut être une stricte inégalité. Soit $R(0) = 1$ et $R(1) = 5$. La fonction valeur correspondante est représentée sur la figure 2.4. On observe que $V(0)$ est strictement plus grand que le renforcement final $R(0)$. Cette inégalité stricte a lieu aux points du bord $x \in \partial O$ où il existe un contrôle $u(.)$ pour lequel la trajectoire rentre dans $\overline{O}$ et qui fournit un meilleur gain que le renforcement final $R(x)$ obtenu en sortant de $\overline{O}$ en x.

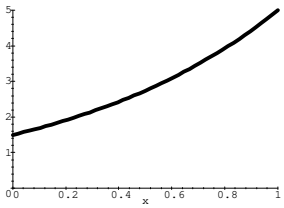

Figure 2.4: La condition au bord est une inégalité. Ici $V(0) > R(0) = 1$.

Nous verrons que l'on peut donner (voir la définition 6 au paragraphe 2.2.3) une formulation faible (de viscosité) des conditions aux bords (2.9a) et (2.11).

2.2.2 Définition des solutions de viscosité

Ici, nous donnons une définition des solutions de viscosité pour des fonctions continues (une définition pour des fonctions discontinues est donnée dans l'annexe A). Considérons le cas à horizon temporel fini.

Définition 5 (Solution de Viscosité) *Soit W une fonction continue, définie sur Q, à valeurs réelles.*

- *W est une* sous-solution de viscosité *de (2.12) dans Q si pour toute fonction $\varphi \in C^1(Q)$, pour tout $(t,x) \in Q$ maximum local de $W - \varphi$ tel que $W(t,x) = \varphi(t,x)$, on a:*

$$H(t,x,\varphi(t,x), D\varphi(t,x)) \leq 0$$

- *W est une* sur-solution de viscosité *de (2.12) dans Q si pour toute fonction $\varphi \in C^1(Q)$, pour tout $(t,x) \in Q$ minimum local de $W - \varphi$ tel que $W(t,x) = \varphi(t,x)$, on a:*

$$H(t,x,\varphi(t,x), D\varphi(t,x)) \geq 0$$

- *W est une* solution de viscosité *de (2.12) dans Q si elle est une sous-solution et une sur-solution de viscosité de (2.12) dans Q.*

Remarque 9 *Pour le cas à horizon infini, la définition est identique, avec O au lieu de Q et $H(x,W,DW)$ au lieu de $H(t,x,W,DW)$.*

2.2.3 Quelques propriétés des solutions de viscosité

Le théorème suivant (dont la preuve peut être trouvé dans [CIL92], [FS93] ou [Bar94]) établit que la fonction valeur est une solution de viscosité.

Théorème 3 *Supposons que l'hypothèse 1 est satisfaites. Alors la fonction valeur V est une solution de viscosité de (2.12) dans Q.*

Afin de pouvoir traiter le problème de l'inégalité des conditions aux bords (2.9a) ou (2.11), nous définissons une formulation de viscosité sous la forme de conditions de type différentiel au lieu de conditions de type purement Dirichlet. Considérons le cas à horizon fini ; pour le cas à horizon infini, la définition et le théorème sont très similaires.

Définition 6 (Condition de viscosité au bord) *Soit W une fonction continue définie sur $\overline{Q}$, à valeurs réelles,*

- *W est une* sous-solution de viscosité *de (2.12) dans Q* avec la condition au bord *(2.9a) si elle est sous-solution de viscosité de (2.12) dans Q et que pour toute fonction $\varphi \in C^1(\overline{Q})$, pour tout $(t,x) \in [t_0, t_1[\times\partial O$ maximum local de $W - \varphi$ tel que $W(t,x) = \varphi(t,x)$ on a:*

$$\min\{H(t,x,W,DW), W(t,x) - \varphi(t,x)\} \leq 0$$

- *W est une* sur-solution de viscosité *de (2.12) dans Q* avec la condition au bord *(2.9a) si elle est sur-solution de viscosité de (2.12) dans Q et que pour toute fonction $\varphi \in C^1(\overline{Q})$, pour tout $(t,x) \in [t_0, t_1) \times \partial O$ minimum local de $W - \varphi$ tel que $W(t,x) = \varphi(t,x)$ on a:*

$$\min\{H(t,x,W,DW), W(t,x) - \varphi(t,x)\} \geq 0 \qquad (2.17)$$

- *W est une* solution de viscosité *de (2.12) dans Q* avec la condition au bord *(2.9a) si elle est sous-solution et sur-solution de viscosité de (2.12) dans Q avec la condition au bord (2.9a).*

Remarque 10 *Lorsque l'Hamiltonien H provient d'un problème de contrôle optimal (ce qui est le cas ici), la condition (2.17) est simplement équivalente à l'inégalité (2.9a).*

Remarque 11 *Il n'est pas nécessaire de donner une formulation de viscosité de la condition terminale (2.9b).*

Avec cette définition, le théorème 3 s'étend aux solutions de viscosité avec les conditions au bord. De plus, nous avons un résultat d'unicité pour la solution de viscosité, ce qui mène au théorème suivant (pour lequel on peut trouver une preuve dans [CIL92], [FS93] ou [Bar94] :

Théorème 4 *Supposons que l'hypothèse 1 est satisfaites. Alors la fonction valeur est l'unique solution de viscosité de (2.12) dans Q avec les conditions au bord latérale (2.9a) et terminale (2.9b).*

Ce théorème très important justifie la pertinence de l'introduction de la notion de solution de viscosité pour les équations de HJB. Ce résultat, ainsi que celui du théorème 3, est obtenu à partir de la continuité de la solution de viscosité (théorème 2) qui est une conséquence d'un principe d'unicité forte énoncé ainsi :

Lemme 1 (Principe d'unicité forte) *Supposons que l'hypothèse 1 est satisfaites, alors (2.12), (2.9a) et (2.9b) satisfont un principe d'unicité forte, i.e. pour toute sous-solution $\underline{W}$ et toute sur-solution $\overline{W}$ de (2.12), (2.9a) et (2.9b), pour tout $(t,x) \in [t_0,t_1] \times O$ on a :*

$$\underline{W}^*(t,x) \leq \overline{W}^*(t,x).$$

Une preuve de ce résultat de comparaison entre sous- et sur-solutions peut être trouvée dans [Bar94], [BP88], [BP90] ou pour des hypothèses légèrement différentes sur la dynamique d'état à la frontière, dans [FS93].

Ce principe sera utilisé (en annexe A) pour prouver la convergence des schémas d'approximation (théorème 5) de la fonction valeur étudiés à la section suivante.

2.3 Approximation par des schémas numériques

2.3.1 Introduction

Dans cette partie, nous tentons de calculer numériquement la fonction valeur à partir d'une discrétisation de l'espace d'état. L'idée principale est que la discrétisation de l'équation de HJB mène à une équation de programmation dynamique (PD) pour un certain Processus de Décision Markovien (PDM). Ce PDM est fini et peut être résolu par des méthodes de PD (voir [Ber87]), ce qui mène pour une résolution δ à la fonction valeur approchée V^δ. On s'intéresse aux propriétés de convergence de telles fonctions V^δ lorsque le pas de discrétisation δ tend vers 0 : est-ce que V^δ converge vers V, la fonction valeur du problème continu ?

Nous définissons deux classes de schémas d'approximation (d'après Kushner [Kus90], [KD92]) basées sur des méthodes aux différences finies (DF) (paragraphe 2.3.2) et aux éléments finis (EF) (paragraphe 2.3.3). Afin d'illustrer plusieurs cas possibles, le paragraphe 2.3.2 traite des problèmes à horizon temporel fini et infini et le paragraphe 2.3.3 de ceux à horizon infini. Le paragraphe suivant 2.3.4 fournit un cadre général assurant la convergence de schémas d'approximation (qui provient de la formulation très générale de [BS91] et qui utilise des propriétés de stabilité des solutions de viscosité) et qui couvre à la fois les méthodes aux DF et aux EF.

Nous considérons par la suite des espaces de contrôle approchés U^δ, espaces finis, approximations de U au sens : $\overline{\cup_\delta U^\delta} = U$ et tels que pour $\delta \leq \delta'$, on ait $U^{\delta'} \subset U^\delta$.

2.3.2 Approximation par des méthodes aux différences finies

Soit $e_1, e_2, ..., e_d$ une base de $\mathbb{R}^d$. La dynamique d'état est : $f = (f_1, ..., f_d)$. Définissons la partie positive (respectivement négative) de f_i étant : $f_i^+ = \max(f_i, 0)$ (resp. $f_i^- = \max(-f_i, 0)$).

Pour un pas de discrétisation spatial δ, on considère les *grilles* suivantes :

$$\begin{aligned} \delta\mathbb{Z}^d &= \left\{\delta . \sum\nolimits_{i=1}^{d} j_i e_i\right\} \text{ avec } j_1, ..., j_d \text{ entiers quelconques} \\ \Sigma^\delta &= \delta\mathbb{Z}^d \cap O \\ \partial\Sigma^\delta &= \{\xi \in \delta\mathbb{Z}^d \setminus O, \text{il existe un point adjacent } \xi \pm \delta e_i \in \Sigma^\delta\} \end{aligned}$$

Σ^δ est *l'espace d'état discrétisé* et $\partial\Sigma^\delta$ est appelée la *frontière* de Σ^δ (voir figure 2.5).

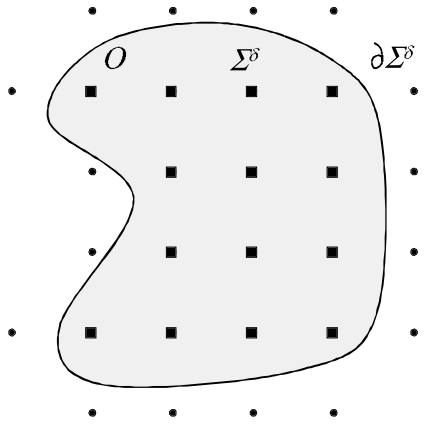

Figure 2.5: L'espace d'état discrétisé Σ^δ (représenté par les points carrés) et sa frontière $\partial\Sigma^\delta$ (les ronds).

La méthode aux DF consiste à remplacer dans l'équation de HJB le gradient $D_x V(t, \xi)$ par les quotients aux différences finies en avant et en arrière dans les directions e_i :

$$\begin{cases} \Delta_i^+ V(t,\xi) = \dfrac{1}{\delta}\left[V(t, \xi + \delta e_i) - V(t,\xi)\right] \\ \Delta_i^- V(t,\xi) = \dfrac{1}{\delta}\left[V(t, \xi - \delta e_i) - V(t,\xi)\right] \end{cases} \tag{2.18}$$

et la dérivée temporelle $V_t(t, \xi)$ par le quotient aux différences :

$$\Delta_t^- V(t,\xi) = \frac{1}{h}\left[V(t,\xi) - V(t-h,\xi)\right]. \tag{2.19}$$

Horizon temporel fini

En utilisant ces quotients aux différences finies, l'équation de HJB (2.8), se discrétise selon :

$$\Delta_t^- V^\delta(t,\xi) + \sup_{v\in U^\delta} \quad \Big\{ \sum_{i=1}^n [f_i^+(t,\xi,v).\Delta_i^+ V^\delta(t,\xi) + f_i^-(t,\xi,v).\Delta_i^- V^\delta(t,\xi)] + r(t,\xi,v)\Big\} = 0\Big\} \text{ pour } \xi \in \Sigma^\delta, t < t_1. \quad (2.20)$$

Nous utilisons la notation $||y||_1 = \sum_{i=1}^d |y_i|$ pour désigner la norme 1 d'un vecteur y.

Choisissons un *pas de discrétisation temporel* h tel que pour tout t, x, u on ait :

$$h.||f(t,x,u)||_1 \leq \delta \quad (2.21)$$

Ce qui est réalisable puisque comme O est borné, d'après (2.2), f est borné dans O. De plus on peut trouver une constante k_h telle que :

$$h \geq k_h.\delta \quad (2.22)$$

(par exemple avec $k_h = \frac{M_f}{2}$ où $M_f = \sup_{t,x,u} ||f(t,x,u)||_1$, on a $k_h.\delta \leq h \leq M_f.\delta$). En remplaçant les quotients aux différences finies par leurs valeurs (2.18) et (2.19) , l'équation (2.20) devient :

$$V^\delta(t,\xi) = \sup_{v\in U^\delta} \left\{ \sum_{\xi'\in\Sigma^\delta} p(t,\xi,\xi',v).V^\delta(t+h,\xi') + h.r(t,\xi,v) \right\} \quad (2.23)$$

avec les valeurs :

$$\begin{aligned} p(t,\xi,\xi',v) &= \frac{h}{\delta} f_i^{\pm}(t,\xi,v) \text{ pour } \xi' = \xi \pm \delta e_i \\ &= 1 - \frac{h}{\delta}||f(t,\xi,u)||_1 \text{ pour } \xi' = \xi \\ &= 0 \text{ sinon} \end{aligned}$$

qui peuvent être interprétées comme des probabilités (la propriété (2.21) assurent qu'elles sont positives). On s'apperçoit que l'équation (2.23) est une équation de PD pour un PDM fini dont *l'espace d'état* est Σ^δ et les *probabilités de transition* de (l'état ξ , le contrôle v) vers les états successeurs suivants ξ' sont les $p(t,\xi,\xi',v)$.

Les conditions terminales et latérales sont :

$$V^\delta(t,\xi) = R_t(\xi) \text{ pour } t = t_1 \quad (2.24)$$

$$V^\delta(t,\xi) = R_l(t,\xi) \text{ pour } \xi \in \partial\Sigma^\delta \quad (2.25)$$

pour lesquelles correspondent les états terminaux absorbants du PDM : pour $t = t_1$ ou $\xi \in \partial\Sigma^\delta$,

$$\begin{aligned} p(t,\xi,\xi,v) &= 1 \\ p(t,\xi,\xi' \neq \xi,v) &= 0 \end{aligned}$$

Ainsi, la discrétisation de l'équation de HJB obtenue en remplaçant les dérivées par des quotients aux différences finies permet de remplacer le problème de contrôle continu par un PDM fini dont l'équation de PD est (2.23) et les conditions aux bords (2.24) et (2.25).

En définissant $F^\delta_{DF_{fini}}$ **le schéma aux DF dans le cas horizon temporel fini :**

$$F^\delta_{DF_{fini}}[\varphi](t,\xi) = \sup_{v\in U^\delta} \left\{ \sum_{\xi'\in\Sigma^\delta} p(t,\xi,\xi',v).\varphi(t+h,\xi') + h.r(t,\xi,v) \right\}$$

l'équation de la PD (2.23) devient

$$V^\delta(t,\xi) = F^\delta_{DF_{fini}}\left[V^\delta\right](t,\xi). \tag{2.26}$$

Pour le problème à horizon temporel fini, la solution V^δ de (2.23) avec les conditions aux bords (2.24) et (2.25) est la fonction valeur du PDM associé et peut être résolu récursivement par rétropropagation dans le temps (voir [Ber87], [Put94] ou [BT96])).

Horizon temporel infini

Avec les quotients aux différences finies, l'équation de HJB (2.10), se discrétise selon :

$$V^\delta(\xi)\ln\gamma + \sup_{v\in U^\delta}\left\{ \sum_{i=1}^{d} \left[f_i^+(\xi,v).\Delta_i^+ V^\delta(\xi) + f_i^-(\xi,v).\Delta_i^- V^\delta(\xi)\right] + r(\xi,v)\right\} = 0$$

En remplaçant les quotients aux différences finies par leur valeur (2.18) on obtient l'équation suivante : pour $\xi \in \Sigma^\delta$,

$$V^\delta(\xi) = \frac{1}{1-\frac{\delta}{||f(\xi,u)||_1}\ln\gamma} \sup_{v\in U^\delta}\left\{ \sum_{\xi'} p(\xi,\xi',v).V^\delta(\xi') + \frac{\delta}{||f(\xi,v)||_1} r(\xi,v)\right\}$$

avec :

$$\begin{aligned} p(\xi,\xi',v) &= \frac{f_i^\pm(\xi,u)}{||f(\xi,u)||_1} \text{ pour } \xi' = \xi \pm \delta e_i \\ &= 0 \text{ sinon.} \end{aligned} \tag{2.27}$$

Sachant que $\frac{1}{1-\Delta t \ln \gamma}$ est équivalent à $\gamma^{\Delta t}$ lorsque Δt tend vers 0, nous pouvons définir le schéma d'approximation équivalent (lorsque δ tend vers 0) suivant :

$$V^\delta(\xi) = \sup_{v \in U^\delta} \left\{ \gamma^{\frac{\delta}{||f(\xi,u)||_1}} \sum_{\xi'} p(\xi, \xi', v).V^\delta(\xi') + \frac{\delta}{||f(\xi,v)||_1} r(\xi, v) \right\} \quad (2.28)$$

Cette équation est de type PD pour un PDM fini dont l'*espace d'état* est Σ^δ, l'espace de contrôle U^δ et les *probabilités de transition* $p(\xi, \xi_i, u)$ (voir figure 2.6 pour une interprétation géométrique des probabilités).

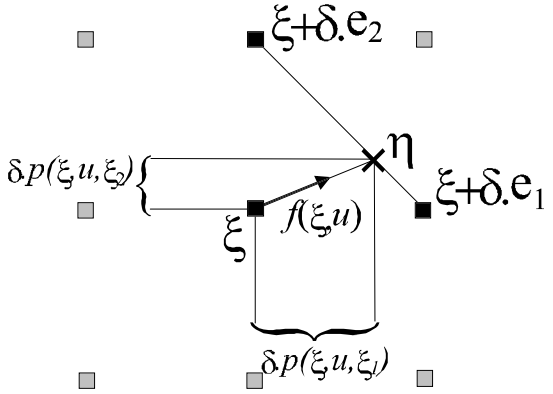

Figure 2.6: *Interprétation géométrique du schéma aux DF pour le cas horizon temporel infini.* Les probabilitiés de transition $p(\xi, \xi', v)$ du PDM correspondant sont les coordonnées du vecteur $\frac{1}{\delta}\xi\eta$ avec η la projection de ξ sur le segment $(\xi + \delta.e_1, \xi + \delta.e_2)$ parallèlement à $f(\xi, v)$.

On définit $F^\delta_{DF_\infty}$ le **schéma aux DF dans le cas horizon temporel infini** :

$$F^\delta_{DF_\infty}[\varphi](\xi) = \sup_{v \in U^\delta} \left\{ \gamma^{\frac{\delta}{||f(\xi,u)||_1}} \sum_{\xi'} p(\xi, \xi', v).\varphi(t+h, \xi') + \frac{\delta}{||f(\xi,v)||_1}.r(\xi, v) \right\} \quad (2.29)$$

L'équation de la PD (2.28) devient : pour $\xi \in \Sigma^\delta$:

$$V^\delta(\xi) = F^\delta_{DF_\infty}\left[V^\delta\right](\xi) \quad (2.30)$$

et pour les états absorbants $\xi \in \partial\Sigma^\delta$,

$$V^\delta(\xi) = R(\xi) \quad (2.31)$$

Cette équation établit que V^δ est un point fixe de $F^\delta_{DF_\infty}$. De plus, comme $||f(\xi,u)||_1$ est bornée supérieurement (par une valeur notée M_f), $F^\delta_{DF_\infty}$ satisfait la propriété de contraction suivante :

$$\left\|F^\delta_{DF_\infty}[\varphi_1] - F^\delta_{DF_\infty}[\varphi_2]\right\| \leq \gamma^{\frac{\delta}{M_f}} . \|\varphi_1 - \varphi_2\|$$

et puisque $\gamma^{\frac{\delta}{M_f}} < 1$, le point fixe est unique. Ainsi, la théorie de la PD assure qu'il existe une unique solution V^δ vérifiant (2.30) et (2.31).

Calcul de V^δ et convergence

Comme indiqué au paragraphe 1.2.2, il existe deux classes de méthodes pour calculer la fonction valeur V^δ du PDM correspondant (voir [Ber87], [Put94], [BBS95] ou [BT96] pour davantage d'information concernant la PD) : les méthodes d'itération sur les stratégies et les méthodes d'itérations sur les valeurs.

Nous espérons que les fonctions $V^\delta(t,\xi)$ et $V^\delta(\xi)$, fonctions valeur des schémas aux DF, $F^\delta_{DF_{fini}}$ et $F^\delta_{FD_\infty}$ sont des "bonnes approximations" de la fonction valeur V du problème continu initial, et en effet nous verrons au paragraphe 2.3.4 que ces schémas sont convergents lorsque δ tend vers zéro.

Remarque 12 *Les schémas d'approximation aux DF $F^\delta_{DF_{fini}}, F^\delta_{DF_\infty}$ définis précédemment sont extrêmement simples puisqu'ils considèrent une discrétisation spatiale homogène dans toutes les directions. On peut construire des schémas dont le pas de discrétisation δ_i dépende de la direction e_i ou encore considérer des discrétisations à résolution variable utilisant des algorithmes multi-grilles (voir [KD92], [Aki90]), voir le chapitre 4.*

2.3.3 Approximation par des méthodes aux éléments finis

Ici nous considérons les problèmes à horizon temporel infini. Nous utilisons une méthode aux EF (avec des simplexes linéaires) et considérons des triangulations Σ^δ (le pas de discrétisation δ représente la largeur maximale des cotés des simplexes) telles que l'ensemble des simplexes recouvre O (voir figure 2.7).

La fonction valeur V est approchée par des fonctions continues, linéaires par morceaux V^δ définies par leurs valeurs aux sommets $\{\xi\}$ de la triangulation Σ^δ. La valeur de V^δ en un point x à l'intérieur d'un simplexe $(\xi_0, ..., \xi_d)$

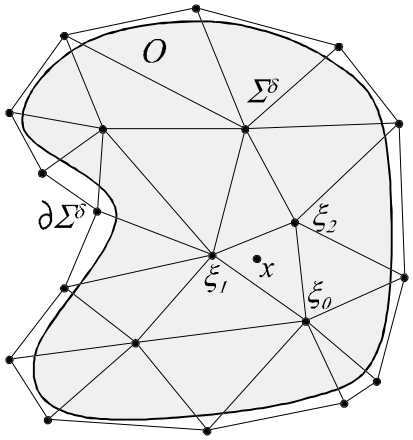

Figure 2.7: Triangulation Σ^δ de l'espace d'état. $V^\delta(x)$ est une combinaison linéaire des $V^\delta(\xi_i)$ pour $x \in$ Simplexe (ξ_0, ξ_1, ξ_2).

est une combinaison linéaires de V^δ aux sommets $\xi_0, ..., \xi_d$ (voir figure 2.7) :

$$V^\delta(x) = \sum_{i=0}^{d} \lambda_{\xi_i}(x) V^\delta(\xi_i) \text{ pour tout } x \in \text{Simplexe } (\xi_0, ..., \xi_d)$$

avec $\lambda_{\xi_i}(x)$ les coordonnées barycentriques de x dans le simplexe $(\xi_0, ..., \xi_d) \ni x$. (On rappelle la définition des coordonnées barycentriques : les $\lambda_{\xi_i}(x)$ satisfont : $\sum_{i=0}^{d} \lambda_{\xi_i}(x).(\xi_i - x) = 0$ et : $\sum_{i=0}^{d} \lambda_{\xi_i}(x) = 1$).

En utilisant un schéma d'approximation aux EF (déduit de [Kus90] ou [KD92]), l'équation de HJB est approchée par l'équation suivante :

$$V^\delta(\xi) = \sup_{u \in U^\delta} \left[\gamma^{\tau(\xi,u)} . V^\delta(\eta(\xi,u)) + \tau(\xi,u) r(\xi,u) \right] \tag{2.32}$$

où $\eta(\xi, u)$ est un point de l'espace couvert par la triangulation tel que $\eta(\xi, u) = \xi + \tau(\xi, u).f(\xi, u)$ (voir figure 2.8) pour une certaine fonction τ : $\Sigma^\delta \times U^\delta \to \mathbb{R}$ satisfaisant pour certaines constantes positives k_1 et k_2 :

$$k_1 \delta \leq \tau(\xi, u) \leq k_2 \delta \tag{2.33}$$

Considérons $(\xi_0, ..., \xi_d)$ le simplexe contenant $\eta(\xi, u)$. Puisque V^δ est linéaire dans le simplexe, l'équation (2.32) s'écrit :

$$V^\delta(\xi) = \sup_{u \in U^\delta} [\gamma^{\tau(\xi,u)} . \sum_{i=1}^{d} \lambda_{\xi_i}(\eta(\xi,u)) V^\delta(\xi_i) + \tau(\xi,u) r(\xi,u)] \tag{2.34}$$

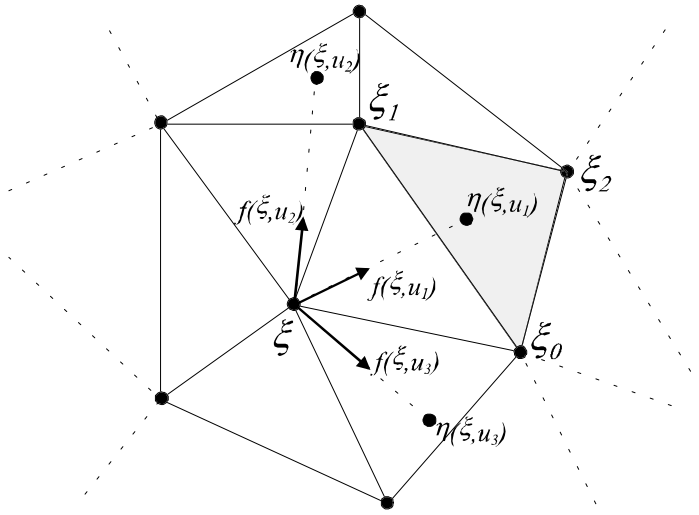

Figure 2.8: Une approximation aux éléments finis. Considérons un sommet ξ de la triangulation et $\eta(\xi, u_1) = \xi + \tau(\xi, u_1).f(\xi, u_1)$. $V^\delta(\eta(\xi, u_1))$ est une combinaison linéaire de $V^\delta(\xi_0), V^\delta(\xi_1)$ et $V^\delta(\xi_2)$ pondérée par les coordonnées barycentriques de $\eta(\xi, u_1)$ dans (ξ_0, ξ_1, ξ_2).

qui est une équation de PD pour un PDM fini dont l'*espace d'état* est l'ensemble des sommets $\{\xi\}$ et les *probabilités de transition* de (l'état ξ, le contrôle u) vers les états successeurs $\xi' \in \{\xi_0, ..., \xi_d\}$ sont les coordonnées barycentriques de $\eta(\xi, u)$ dans le simplexe $(\xi_0, ..., \xi_d)$ (et 0 pour $\xi' \notin \{\xi_0, ..., \xi_d\}$) :

$$\begin{aligned} p(\xi, \xi_i, u) &= \lambda_{\xi_i}(\xi + \tau(\xi, u).f(\xi, u)) \\ &\quad \text{pour } \xi + \tau(\xi, u).f(\xi, u) \in \text{Simplexe } (\xi_0, ..., \xi_d) \\ p(\xi, \xi', u) &= 0 \text{ sinon} \end{aligned} \tag{2.35}$$

Pour les états sur la frontière $\xi \in \partial\Sigma^\delta$ la condition au bord est

$$V^\delta(\xi) = R(\xi). \tag{2.36}$$

En définissant F^δ_{EF} le **schéma aux éléments finis**,

$$F^\delta_{EF}\left[\varphi\right](\xi) = \sup_{v \in U} \left\{ \gamma^{\tau(\xi,u)}. \sum_{i=1}^{n} \lambda_{\xi_i}(\eta(\xi, u))\varphi(\xi_i) + \tau(\xi, u)r(\xi, u) \right\} \tag{2.37}$$

pour toute fonction φ définie sur Σ^δ, la fonction valeur approchée V^δ satisfait l'équation de PD :

$$V^\delta(\xi) = F^\delta_{EF}\left[V^\delta\right](\xi). \tag{2.38}$$

Comme le schéma aux DF, F^{δ}_{EF} vérifie la propriété de contraction suivante :

$$\left\| F^{\delta}_{EF}[\varphi_1] - F^{\delta}_{EF}[\varphi_1] \right\| \leq \gamma^{k_1\delta} . \|\varphi_1 - \varphi_2\|$$

avec $\gamma^{k_1\delta} < 1$. Donc il existe une unique solution V^{δ} de (2.38) avec la condition au bord (2.36). De même que précédemment, nous nous intéressons aux propriétés de convergence de ce schéma aux EF : est-ce que V^{δ} converge vers V lorsque δ tend vers zero ? Cette question est traitée dans le paragraphe suivant.

2.3.4 Convergence des schémas d'approximation

Ici nous présentons un théorème de convergence pour une large classe de schémas d'approximation. En fait, la principale propriété nécessaire à la convergence est la consistance du schéma (propriétés (2.44) et (2.45) ci-dessous). En corollaire, nous déduisons que les schéma aux DF et aux EF étudiés précédemment sont convergents.

Un théorème général de convergence

Dans ce paragraphe, pour des raisons de commodité, nous utilisons le formalisme du cas à horizon temporel fini ; néanmoins, les mêmes résultats sont valables pour le cas à horizon infini. Soit Σ^{δ} et $\partial\Sigma^{\delta}$ deux sous-ensembles finis de $\mathbb{R}^d$. On suppose que pour tout $x \in O, \lim_{\delta\downarrow 0} dist(x, \Sigma^{\delta}) = 0$ et pour tout $x \in \partial O, \lim_{\delta\downarrow 0} dist(x, \partial\Sigma^{\delta}) = 0$. Soit F^{δ} un opérateur de l'espace des fonctions bornées sur Σ^{δ}. On s'intéresse à la convergence des solutions V^{δ} de l'équation de la PD :

$$V^{\delta}(t,x) = F^{\delta}(V^{\delta})(t,x) \text{ pour } x \in \Sigma^{\delta} \tag{2.39}$$

avec les conditions terminale et latérale suivantes :

$$\begin{array}{ll} V^{\delta}(t,x) = R_t(x) & \text{pour } t = t_1 \\ V^{\delta}(t,x) = R_l(t,x) & \text{pour } x \in \partial\Sigma^{\delta} \end{array} \tag{2.40}$$

Nous faisons les hypothèses suivantes :

- *Monotonie* :

$$\text{si } \varphi_1 \leq \varphi_2 \text{ alors } F^{\delta}(\varphi_1) \leq F^{\delta}(\varphi_2) \tag{2.41}$$

- Pour toute constante c,

$$F^{\delta}(\varphi + c) = F^{\delta}(\varphi) + c(1 + O(\delta)) \tag{2.42}$$

- Pour tout δ,

$$\text{Il existe une solution } V^\delta \text{ de (2.39) qui est bornée avec une constante } M_V \text{ indépendante de } \delta. \tag{2.43}$$

- *Consistance* : il existe une constante $k > 0$ telle que :

  - si $H(t,x,\varphi(t,x),D\varphi(t,x)) \geq 0$ alors

$$\liminf_{(t_\delta,x_\delta)\overset{\delta\downarrow 0}{\longrightarrow}(t,x)} \frac{1}{\delta}\left[\varphi - F^\delta[\varphi]\right](t_\delta,x_\delta) \geq k.H(t,x,\varphi(t,x),D\varphi(t,x)) \tag{2.44}$$

  - si $H(t,x,\varphi(t,x),D\varphi(t,x)) \leq 0$ alors

$$\limsup_{(t_\delta,x_\delta)\overset{\delta\downarrow 0}{\longrightarrow}(t,x)} \frac{1}{\delta}\left[\varphi - F^\delta[\varphi]\right](t_\delta,x_\delta) \leq k.H(t,x,\varphi(t,x),D\varphi(t,x)) \tag{2.45}$$

Remarque 13 *Les conditions (2.44) et (2.45) sont vérifiées dans le cas particulier où :*

$$\lim_{(t_\delta,x_\delta)\overset{\delta\downarrow 0}{\longrightarrow}(t,x)} \frac{1}{\delta}\left[\varphi - F^\delta[\varphi]\right](t_\delta,x_\delta) = H(t,x,\varphi(t,x),D\varphi(t,x))$$

Le théorème suivant, dont la preuve est donnée en annexe A, énonce qu'un schéma d'approximation satisfaisant les propriétés précédentes est convergent.

Théorème 5 (Convergence des schémas) *Supposons que l'hypothèse 1 est vérifiée. Supposons de plus que (2.41), (2.42), (2.44) et (2.45) sont satisfaites, alors F^δ est un schéma d'approximation convergent, i.e. les solutions V^δ de (2.39) et (2.40) vérifient :*

$$\lim_{(t_\delta,x_\delta)\overset{\delta\downarrow 0}{\longrightarrow}(t,x)} V^\delta(t_\delta,x_\delta) = V(t,x) \text{ uniformément sur tout compact } \Omega \subset Q \tag{2.46}$$

Idée de la démonstration

On utilise la méthode de Barles et Perthame [BP88]. L'idée est de définir la fonction limite supérieure $V_{\text{sup}} = \limsup V^\delta$ et la fonction limite inférieure $V_{\text{inf}} = \liminf V^\delta$ et de montrer qu'elles sont respectivement sous- et sur-

solutions de viscosité discontinues. Cette preuve, qui reprend le théorème de convergence général de Barles and Souganidis [BS91], est donnée en annexe A. Puis nous utilisons le résultat d'unicité forte du lemme 1 qui établit que si l'hypothèse 1 est vérifiée, alors les sous-solutions de viscosité sont inférieures aux sur-solutions, d'où $V_{\text{sup}} \leq V_{\text{inf}}$. Par définition $V_{\text{sup}} \geq V_{\text{inf}}$, donc $V_{\text{sup}} = V_{\text{inf}} = V$ et la fonction limite V est solution de viscosité de l'équation de HJB donc (par le théorème 4) la fonction valeur du problème.

Les schémas d'approximation aux DF et aux EF convergent

Corollaire 1 *Les schémas d'approximation $F^{\delta}_{DF_{fini}}, F^{\delta}_{DF_{\infty}}$ et F^{δ}_{EF} sont convergents.*

Preuve. Pour les schémas aux DF, il est clair que puisque les valeurs $p(t,x,y,v)$ peuvent être considérées comme des probabilités de transition, les schémas $F^{\delta}_{DF_{fini}}$ et $F^{\delta}_{DF_{\infty}}$ satisfont (2.41) et (2.42). Puisque (2.26) et (2.30) sont des équations de PD pour certains PDM, on en déduit que (2.43) est vrai.

De plus nous pouvons vérifier que les schémas sont consistants : d'après (2.22), les conditions (2.44) et (2.45) sont vraies pour $F^{\delta}_{DF_{finite}}$ avec $k = k_h$. Pour $F^{\delta}_{DF_{\infty}}$, les conditions (2.44) et (2.45) sont vraies avec $k = \frac{1}{M_f}$. Donc $F^{\delta}_{DF_{finite}}$ et $F^{\delta}_{DF_{\infty}}$ satisfont les hypothèses du théorème 5.

De manière similaire, pour le schéma aux EF, les propriétés de base des coordonnées barycentriques $\lambda_{\xi_i}(x)$ entraîne que F^{δ}_{EF} satisfait (2.41). Grâce à (2.33), la condition (2.42) est vraie. La théorie de la PD assure que (2.43) est vraie.

Enfin le schéma est consistant et les conditions (2.44) et (2.45) sont vérifiées pour $k = k_1$. Donc F^{δ}_{EF} satisfait les hypothèses du théorème 5. ■

2.4 Le cas stochastique

2.4.1 La dynamique d'état

Ici nous considérons un processus de diffusion Markovien contrôlé dont la dynamique d'état, c'est à dire la variation de l'état courant $x(t) \in \bar{O}$ dépend du contrôle $u(t) \in U$ (les ensembles O et U satisfont les mêmes hypothèses que dans le cas déterministe) selon l'intégrale stochastique, au sens de Itô, suivante :

$$x(t) = x(0) + \int_0^t f(x(s), u(s))ds + \int_0^t \sigma(x(s), u(s))dw$$

où f, la *dérive locale* (l'aspect déterministe ou variation lente), est commun à la dynamique d'état (2.1) du cas déterministe étudié précédemment, et w, la *diffusion* (l'aspect stochastique ou variation rapide), est un processus standard de Wiener de dimension r. Enfin, σ est une matrice de dimension $d \times r$. Cette intégrale stochastique peut s'exprimer sous la forme différentielle :

$$dx = f(x(t), u(t))dt + \sigma(x(t), u(t))dw \tag{2.47}$$

appelée *équation différentielle stochastique* (EDS), au sens de Itô. Nous ne développerons pas les aspects théoriques des EDS pour lesquels on pourra se référer à [Kry80], [IW89], [FS93], [Kry95] ou [Oks95].

L'aspect stochastique de (2.47) peut provenir de fluctuation aléatoires, d'influences bruitées, d'un manque de précision dans le contrôle, etc.

Par soucis de simplicité, nous considérons ici une dynamique d'état f et σ indépendante du temps. Pour un état initial x, le choix d'une fonction contrôle $u(t)$ permet, en intégrant (2.47) de générer une infinité de trajectoires possibles. Pour une telle trajectoire $x(t)$ (voir figure 2.9), on définit τ son temps de sortie de $\bar{O}$.

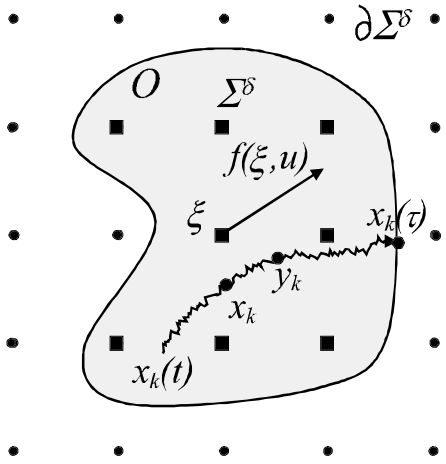

Figure 2.9: L'espace d'état O, sa discrétisation Σ^δ(les points carrés) et la frontière discrétisée $\partial\Sigma^\delta$ (les ronds). Une trajectoire $x_k(t)$ traverse le voisinage de ξ.

De manière similaire au cas déterministe, nous considérons deux classes de problèmes : les problèmes à horizon temporel fini et ceux à horizon temporel infini avec renforcement à décroissance exponentielle. Dans les deux cas, nous définissons la fonctionnelle de renforcement :

Définition 7 (Fonctionnelle de renforcement) *On définit la fonctionnelle de renforcement comme une espérance sur toutes les trajectoires possibles :*

- *Horizon temporel fini : pour tout état initial x_0, t_0 et pour toute fonction contrôle $u(t)$ pour $t_0 \leq t \leq \tau \leq t_1$,*

$$J(t_0, x_0; u(t)) = E_{t_0, x_0; u(.)} \left\{ \int_{t_0}^{\tau} r(t, x(t), u(t))dt + R(\tau, x(\tau)) \right\} \quad (2.48)$$

- *Horizon temporel infini : pour tout état initial x_0 et pour toute fonction contrôle $u(t)$ pour $0 \leq t \leq \tau \leq \infty$,*

$$J(x_0; u(t)) = E_{x_0; u(.)} \left\{ \int_{t_0}^{\tau} r(x(t), u(t))dt + R(x(\tau)) \right\} \quad (2.49)$$

Nous faisons l'hypothèse que f, σ sont de classe $\mathcal{C}^2$, r et R sont Lipschitziennes (avec les constantes L_r et L_R) et la frontière ∂O est $\mathcal{C}^2$.

On définit de même la fonction valeur :

Définition 8 *La fonction valeur est la valeur maximale de la fonctionnelle de renforcement (2.48) (respectivement (2.49)) en tant que fonction des données initiales :*

$$V(t, x) = \sup_{u(.)} J(t, x; u(.)) \quad (2.50)$$

(respectivement :

$$V(t, x) = \sup_{u(.)} J(x; u(.))). \quad (2.51)$$

2.4.2 L'équation de Hamilton-Jacobi-Bellman

A partir du principe de PD, on montre que la fonction valeur satisfait une équation différentielle non linéaire du second ordre appelée également équation de Hamilton-Jacobi-Bellman (voir [Kry80], [KD92] ou [FS93]).

Soit la matrice $a = \sigma.\sigma^T$ de dimension $d \times d$ (avec ^{T} désignant la matrice transposée). Nous considérons dans toute la suite, le cas *parabolique uniforme,* i.e. vérifiant l'hypothèse suivante :

Hypothèse 2 (Parabilicité uniforme) *Il existe $c > 0$ tel que : $\forall x \in \overline{O}, \forall u \in U, \forall y \in I\!R^d$,*

$$\sum_{i,j=1}^{d} a_{ij}(x, u) y_i y_j \geq c||y||^2 \quad (2.52)$$

Sous cette hypothèse, on montre que V est $\mathcal{C}^2$ (voir [Kry80]). Soit V_x le gradient de V et $V_{x_i x_j}$ ses dérivées partielles du second ordre.

Théorème 6 (Hamilton-Jacobi-Bellman) *Nous avons, pour les deux cas suivants :*

- ***Horizon temporel fini*** *: pour tout $x \in O$, tout $t \in [t_0, t_1[$, l'équation de Hamilton-Jacobi-Bellman suivante :*

$$V_t(t,x) + \sup_{u \in U} \left[r(t,x,u) + V_x(t,x).f(x,u) + \frac{1}{2} \sum_{i,j=1}^{n} a_{ij}(x,u) V_{x_i x_j}(t,x) \right] = 0 \tag{2.53}$$

a lieu. De plus, la fonction valeur satisfait les conditions au bord suivantes :

$$V(t,x) = R_l(t,x) \text{ si } (t,x) \in [t_0, t_1[\times \partial O \tag{2.54a}$$

$$V(t,x) = R_t(x) \text{ si } t = t_1 \tag{2.54b}$$

- ***Horizon temporel infini*** *: pour tout $x \in O$, l'équation de Hamilton-Jacobi-Bellman suivante :*

$$V(x) \ln \gamma + \sup_{u \in U} \left[r(x,u) + V_x(x).f(x,u) + \frac{1}{2} \sum_{i,j=1}^{n} a_{ij}(x,u) V_{x_i x_j}(x) \right] = 0 \tag{2.55}$$

a lieu. De plus, la fonction valeur satisfait les conditions au bord suivantes :

$$V(x) = R(x) \text{ pour } x \in \partial O \tag{2.56}$$

Remarque 14 *L'hypothèse de parabolicité uniforme entraîne que les conditions au bord (2.54a) et (2.56) sont des égalités (contrairement au cas déterministe). Cette hypothèse permet de simplifier le problème au sens que nous ne sommes pas confrontés aux difficultés du cas déterministe (développées au paragraphe 2.2.1) du fait du sens de l'intégration de l'équation de HJB (puisque V est $\mathcal{C}^2$ donc satisfait HJB partout) et des conditions aux limites. Cependant, sans cette hypothèse, nous pourrions, de même, introduire la notion de solution de viscosité pour les équations du deuxième ordre (voir par exemple [CIL92], [FS93]).*

Remarque 15 *De même qu'au chapitre précédent, à partir de l'équation HJB (2.53), on déduit que la stratégie optimale de contrôle en boucle fermé est :*

$$\pi^*(t,x) \in \arg\sup_{u \in U} [r(t,x,u) + V_x(t,x).f(x,u) + \frac{1}{2} \sum_{i,j=1}^{n} a_{ij}(x,u) V_{x_i x_j}(t,x)].$$

2.4.3 Approximation par des schémas numériques convergents

Pour un pas de discrétisation δ, on définit les grilles suivantes :

$$\begin{aligned}\Sigma^\delta &= \delta\mathbb{Z}^d \cap O \\ \partial\Sigma^\delta &= \left\{ \xi + \delta \sum_{i=1}^{d} \varepsilon(i) e_i \text{ avec } \xi \in \Sigma^\delta \text{ et } \varepsilon(i) \in \{-1, 0, 1\} \right\} \setminus \Sigma^\delta\end{aligned}$$

(voir figure 2.9). Nous faisons l'hypothèse suivante :

Hypothèse 3 *Pour tout $i = 1..d$, on suppose que :*

$$a_{ii}(x,u) - \sum_{j \neq i} |a_{ij}(x,u)| \geq 0.$$

Horizon temporel fini

De même qu'au paragraphe 2.3.2, nous remplaçons la dérivée temporelle $V_t(t,\xi)$ par le quotient aux différences :

$$\Delta_t^- V(t,\xi) = \frac{1}{h} \left[V(t,\xi) - V(t-h,\xi) \right],$$

le gradient $V_x(t,\xi)$ par les quotients aux différences finies

$$\Delta_{x_i}^{\pm} V(t,\xi) = \frac{1}{\delta} \left[V(t,\xi \pm \delta e_i) - V(t,\xi) \right]$$

et les dérivées partielles du second ordre $V_{x_i x_j}(t,\xi)$ par les quotients aux différences finies du second ordre :

$$\begin{aligned}\Delta_{x_i}^2 V(t,\xi) &= \frac{1}{\delta^2} \left[V(t,\xi+\delta e_i) + V(t,\xi-\delta e_i) - 2V(t,\xi) \right] \\ \Delta_{x_i x_j}^+ V(t,\xi) &= \frac{1}{2\delta^2} \left[V(t,\xi+\delta e_i+\delta e_j) + V(t,\xi-\delta e_i-\delta e_j) + 2V(t,\xi) \right. \\ &- \left. V(t,\xi+\delta e_i) - V(t,\xi-\delta e_i) - V(t,\xi+\delta e_j) - V(t,\xi-\delta e_j) \right] \\ \Delta_{x_i x_j}^- V(t,\xi) &= \frac{1}{2\delta^2} \left[V(t,\xi+\delta e_i-\delta e_j) + V(t,\xi-\delta e_i+\delta e_j) + 2V(t,\xi) \right. \\ &- \left. V(t,\xi+\delta e_i) - V(t,\xi-\delta e_i) - V(t,\xi+\delta e_j) - V(t,\xi-\delta e_j) \right]\end{aligned}$$

dans l'équation de HJB (2.53) qui, discrétisée, devient :

$$V^\delta(t-h,\xi) = V^\delta(t,\xi) + h.H^\delta \left(t, \xi, \Delta_{x_i}^{\pm} V^\delta(t,\xi), \Delta_{x_i}^2 V^\delta(t,\xi), \Delta_{x_i x_j}^+ V^\delta(t,\xi) \right)$$

avec l'Hamiltonien discrétisé :

$$H^\delta(t,x,p^\pm,A_{ii}, \quad A^\pm_{ij}) = \sup_{u\in U^\delta}\left\{ r(t,x,u) + \left[\sum_{i=1}^d f_i^+(t,x,u).p^+ + f_i^-(t,x,u).p^- \right.\right.$$
$$\left.\left. +\frac{1}{2}\left(a_{ii}(x,u).A_{ii} + \sum_{j\neq i}\left[a^+_{ij}(x,u).A^+_{ij} + a^-_{ij}(x,u).A^-_{ij}\right]\right)\right]\right\}.$$

et en arrangeant les termes, cette équation s'écrit :

$$V^\delta(t,\xi) = \sup_{u\in U^\delta}\left\{ \sum_{\xi'\in\Sigma^\delta} p(t,\xi,\xi',u).V^\delta(t+h,\xi') + h.r(t,\xi,u)\right\} \qquad (2.57)$$

avec les valeurs :

$$\begin{aligned}
p(t,\xi,\xi',u) &= \frac{h}{2\delta^2}\left[a_{ii}(\xi,u) - \sum_{j\neq i}|a_{ij}(\xi,u)| + 2\delta f_i^\pm(t,\xi,u)\right] \text{ pour } \xi' = \xi \pm \delta e_i \\
&= \frac{h}{2\delta^2} a^\pm_{ij}(\xi,u) \text{ pour } \xi' = \xi + \delta e_i \pm \delta e_j \text{ avec } i\neq j \\
&= \frac{h}{2\delta^2} a^\mp_{ij}(\xi,u) \text{ pour } \xi' = \xi - \delta e_i \pm \delta e_j \text{ avec } i\neq j \\
&= 1 - \frac{h}{\delta^2}\left[\delta\,|f_i(\xi,u)| + \sum_{i=1}^d a_{ii}(\xi,u) - \frac{1}{2}\sum_{j\neq i}|a_{ij}(\xi,u)|\right] \text{ pour } \xi' = \xi \\
&= 0 \text{ sinon.}
\end{aligned}$$

qui peuvent être interprétées comme des probabilités (l'hypothèse 2.52 assurent qu'elles sont positives). L'équation (2.57) est une équation de PD pour un PDM fini dont *l'espace d'état* est Σ^δ et les *probabilités de transition* de (l'état ξ , le contrôle u) vers les états successeurs suivants ξ' sont les $p(t,\xi,\xi',u)$.

Horizon temporel infini

De même qu'au paragraphe 2.3.2, nous obtenons après substitutions dans (2.55) des dérivées d'ordre 1 et 2 par les quotients aux différences finies, l'équation suivante :

$$V^\delta(\xi)\ln\gamma + \sup_{u\in U^\delta}\left\{ r(\xi,u) + \sum_{i=1}^d\left[f_i^+(\xi,u).\Delta^+_{x_i}V^\delta(\xi) - f_i^-(\xi,u).\Delta^-_{x_i}V^\delta(\xi)\right.\right.$$
$$\left.\left. +\tfrac{a_{ii}(\xi,u)}{2}\Delta_{x_ix_i}V(\xi) + \sum_{j\neq i}\left(\tfrac{a^+_{ij}(\xi,u)}{2}\Delta^+_{x_ix_j}V(\xi) - \tfrac{a^-_{ij}(\xi,u)}{2}\Delta^-_{x_ix_j}V(\xi)\right)\right]\right\} = 0.$$

En procédant de même qu'au paragraphe 2.3.2, cette équation est équivalente (lorsque δ tend vers 0) à :

$$V^\delta(\xi,u) = \sup_{u\in U^\delta}\left[\gamma^{\tau(\xi,u)}\sum_{\xi'\in\Sigma^\delta} p(\xi,\xi',u)V^\delta(\xi') + \tau(\xi,u)r(\xi,u)\right] \quad (2.58)$$

$$\text{avec } \tau(\xi,u) = \frac{\delta^2}{\sum_{i=1}^d\left[\delta\left|f_i(\xi,u)\right| + a_{ii}(\xi,u) - \frac{1}{2}\sum_{j\neq i}\left|a_{ij}(\xi,u)\right|\right]} \quad (2.59)$$

et les valeurs :

$$\begin{aligned}
p(\xi,\xi',u) &= \frac{\tau(\xi,u)}{2\delta^2}\left[2\delta\left|f_i^{\pm}(\xi,u)\right| + a_{ii}(\xi,u) - \sum_{j\neq i}\left|a_{ij}(\xi,u)\right|\right] \text{ pour } \xi' = \xi \pm \delta e_i \\
p(\xi,\xi',u) &= \frac{\tau(\xi,u)}{2\delta^2}a_{ij}^{\pm}(\xi,u) \text{ pour } \xi' = \xi + \delta e_i \pm \delta e_j \text{ avec } i\neq j, \qquad (2.60)\\
p(\xi,\xi',u) &= \frac{\tau(\xi,u)}{2\delta^2}a_{ij}^{\mp}(\xi,u) \text{ pour } \xi' = \xi - \delta e_i \pm \delta e_j \text{ avec } i\neq j, \\
p(\xi,\xi',u) &= 0 \text{ sinon.}
\end{aligned}$$

L'équation (2.58) est une équation de PD pour un PDM fini dont l'*espace d'état* est Σ^δ et les *probabilités de transition* de (l'état ξ , le contrôle u) vers les états successeurs suivants ξ' sont les $p(\xi,\xi',u)$.

Remarque 16 *L'hypothèse (2.52) assure que les $p(\xi,\xi',u)$ sont positifs. Si cette condition n'est pas satisfaite, il existe plusieurs possibilités pour contourner le problème, voir [Kus90].*

De même que pour le cas déterministe, nous définissons le **schéma aux DF dans le cas stochastique et horizon temporel infini :**

$$F^\delta_{Stochas}\left[\varphi\right](\xi) = \sup_{u\in U^\delta}\left[\gamma^{\tau(\xi,u)}\sum_{\xi'\in\Sigma^\delta} p(\xi,\xi',u)V^\delta(\xi') + \tau(\xi,u)r(\xi,u)\right] \quad (2.61)$$

et l'équation de PD (2.58) s'écrit : pour $\xi\in\Sigma^\delta$

$$V^\delta(\xi) = F^\delta_{Stochas}\left[V^\delta\right](\xi)$$

et les conditions aux limites sont : pour $\xi\in\partial\Sigma^\delta$

$$V^\delta(\xi) = R(\xi).$$

De même, le schéma $F^{\delta}_{Stochas}$ vérifie la propriété de contraction :

$$\left\| F^{\delta}_{Stochas}\left[\varphi_1\right] - F^{\delta}_{Stochas}\left[\varphi_2\right] \right\| \leq \gamma^{\frac{\delta}{M_f}} . \left\| \varphi_1 - \varphi_2 \right\|$$

(avec M_f majorant de $\|f\|_1$) et il existe une solution unique (le point fixe) V^{δ} de (2.58) pour $\xi \in \Sigma^{\delta}$ avec les conditions aux limites $V^{\delta}(\xi) = R(\xi)$ pour $\xi \in \partial\Sigma^{\delta}$.

Nous pouvons montrer que $F^{\delta}_{Stochas}$ est un schéma qui vérifie les hypothèses (2.41), (2.42), (2.44) et (2.45) du théorème 5 et donc les V^{δ} convergent vers V lorsque $\delta \downarrow 0$:

$$\lim_{\xi \xrightarrow{\delta \downarrow 0} x} V^{\delta}(\xi) = V(x) \text{ uniformément sur } \overline{O}$$

Chapitre 3

Algorithmes d'apprentissage par renforcement

Dans ce chapitre, nous abordons les méthodes d'apprentissage par renforcement pour résoudre les équations de PD établies aux chapitres précédents. La méthode utilisée est très générale et permet de définir des algorithmes basés sur les schémas aux DF, aux EF pour des dynamiques déterministes ou stochastiques, qui peuvent être utilisés "en-ligne" ou "hors-ligne" (voir paragraphe 1.2.2), selon des méthodes dites "directes" ou "indirectes" (voir plus bas).

Ces algorithmes sont décrits sous la forme d'une règle itérative de type PD qui itère, à l'étape n, pour un ensemble d'états Ξ_n^δ (partie d'un espace discrétisé : $\Xi_n^\delta \subset \Sigma^\delta \cup \partial\Sigma^\delta$), les valeurs $V_n^\delta(\xi)$ (ou des Q-valeurs $Q_n^\delta(\xi,u)$) des états $\xi \in \Xi_n^\delta$. Par exemple pour la méthode de résolution dite synchrone, $\Xi_n^\delta{=}\Sigma^\delta$ pour tout n ; pour les méthodes en ligne, $\Xi_n^\delta{=}\{\xi_n\}$, etc.

La règle consiste à itérer à l'étape n, les valeurs des états $\xi \in \Xi_n^\delta$ selon :

$$V_{n+1}^\delta(\xi) \leftarrow \widetilde{F_n^\delta}\left[V_n^\delta\right](\xi)$$

en utilisant un **schéma d'algorithme** $\widetilde{F_n^\delta} : \Sigma^\delta \cup \partial\Sigma^\delta \to \Xi_n^\delta$, à l'étape n (qui défini l'algorithme utilisé). Afin de définir un algorithme convergent (au sens de la convergence des valeurs calculées V_n^δ vers la fonction valeur V) nous imposerons que le schéma d'algorithme soit "proche" d'un schéma numérique convergent F^δ (traités au chapitre précédent), c'est à dire vérifiant une propriété de "contraction faible" (définition donnée au paragraphe 3.1.1).

Nous donnons la définition suivante d'un algorithme :

Définition 9 (Algorithme) *Un algorithme est la donnée des couples : (ensemble des états itérés Ξ_n^δ, schéma d'algorithme $\widetilde{F_n^\delta}$) pour toute étape n.*

La partie 3.1 définit la méthode générale pour générer des algorithmes d'A/R. Le théorème 7, dont la preuve est établie au paragraphe 3.1.2 (ou dans [Mun97e]), fournit les hypothèses suffisantes pour démontrer la convergence.

La partie 3.2 utilise ce théorème pour définir des algorithmes indirects (c'est à dire nécessitant la construction d'un "modèle" de la dynamique d'état) pour des dynamiques déterministes (paragraphe 3.2.1) et stochastiques (paragraphe 3.2.2). Les preuves de convergence de ces algorithmes sont données respectivement en annexes B et C.

La partie 3.3 définit des algorithmes directs (qui ne nécessitent pas l'élaboration d'un modèle de la dynamique d'état) basés sur des méthodes aux EF (paragraphe 3.3.1) et aux DF (paragraphe 3.3.2). Les preuves de convergence sont données respectivement en annexes D et E.

3.1 La méthode générale

Afin de résoudre les équations de la PD (2.23), (2.28), (2.34), (2.57) ou (2.58) on peut utiliser des méthodes de PD dites "*hors-ligne*" (off-line) (synchrones, de Gauss-Seidel, asynchrones) ou "*en-ligne*" (on-line), ou en temps réel (voir [Ber87], [BBS91], [Put94], [BBS95], [BT96]).

Par exemple, en introduisant les Q-valeurs $Q_n^\delta(\xi, u)$, l'équation (2.28), (2.34) ou (2.58) peut être résolue par des itérations successives des états ξ et contrôles u (dans n'importe quel ordre pourvu que tous les états et contrôles soient itérés régulièrement) par la règle :

$$\begin{aligned} Q_{n+1}^\delta(\xi, u) &\leftarrow \gamma^{\tau(\xi,u)} . \sum_{\xi'} p(\xi, \xi', u) . V_n^\delta(\xi') + \tau(\xi, u) r(\xi, u) \qquad (3.1) \\ \text{avec } V_n^\delta(\xi) &= \sup_{u \in U^\delta} Q_n^\delta(\xi, u) \end{aligned}$$

ce qui fournit des algorithmes dont les valeurs calculées convergent vers la solutions V^δ.

Cependant, dans l'approche A/R, nous avons vu que la dynamique d'état et les fonctions renforcement r, R sont inconnues du système. Donc la partie droite de la règle itérative (3.1) doit être approchée à partir des connaissances disponibles. Selon la terminologie de l'A/R, il existe deux approches possibles :

- l'**approche indirecte** (*model-based)* consiste à apprendre dans un premier temps un modèle de la dynamique d'état et des fonctions renforcement pour ensuite utiliser des algorithmes de PD (par exemple en utilisant la règle (3.1)) avec les valeurs du modèle. L'apprentissage (la

modification des qualités Q_n^δ) peut se faire soit au cours des trajectoires (apprentissage en-ligne) soit à la fin (au temps de sortie) de chaque ou de plusieurs trajectoires (apprentissage hors-ligne).

- l'**approche directe** (*model-free)* consiste à modifier itérativement les qualités estimées Q_n^δ des états parcourus sans nécessairement apprendre de modèle de la dynamique d'état et des renforcements.

Dans le paragraphe suivant, nous proposons un théorème de convergence dans le cas à horizon temporel infini qui s'applique à une grande classe d'algorithmes d'A/R (directs, indirects, en-ligne, hors-ligne).

3.1.1 Théorème de convergence des algorithmes d'A/R

Nous considérons le cas à horizon temporel infini. Pour un pas de discrétisation δ, considérons des sous-ensembles finis Σ^δ et $\partial\Sigma^\delta$ vérifiant les propriétés du paragraphe 2.3.4.

La propriété requise pour la convergence d'un algorithme $(\Xi_n^\delta, \widetilde{F_n^\delta})$ **est que le schéma d'algorithme** $\widetilde{F_n^\delta}$ **satisfasse une propriété de "contraction faible" par rapport à un schéma d'approximation convergent** F^δ**,** au sens que leur différence satisfait la propriété de contraction faible définie ci-dessous :

Définition 10 *Un schéma d'algorithme* $\widetilde{F_n^\delta}$ *satisfait une propriété de "contraction faible" par rapport à un schéma d'approximation convergent* F^δ *(par exemple satisfaisant les hypothèses du théorème 5) s'il existe deux constante positives* k *et* k_R *ainsi qu'une fonction* $e(\delta)$ *qui tend vers* 0 *lorsque* $\delta \downarrow 0$*, tels que pour toute étape* n*,*

- *pour tout* $\xi \in \Xi_n^\delta \cap \Sigma^\delta$*, on ait :*

$$\left|\widetilde{F_n^\delta}\left[V_n^\delta\right](\xi) - F^\delta[V^\delta](\xi)\right| \leq (1 - k.\delta) \sup_{\xi \in \Sigma^\delta \cup \partial\Sigma^\delta} \left|V_n^\delta(\xi) - V^\delta(\xi)\right| + e(\delta).\delta$$

- *pour tout* $\xi \in \Xi_n^\delta \cap \partial\Sigma^\delta$*, on ait :*

$$\left|\widetilde{F_n^\delta}\left[V_n^\delta\right](\xi) - F^\delta[V^\delta](\xi)\right| \leq k_R.\delta$$

En notant $N(\xi, n)$ le nombre de fois que l'état ξ apparaît dans $\Xi_1^\delta \cup \Xi_2^\delta \cup ... \cup \Xi_n^\delta$ (nombre d'itérations de l'état ξ en n étapes), nous pouvons donner la définition d'un algorithme exhaustif :

Définition 11 (Algorithme exhaustif) *Un algorithme est dit exhaustif si tous les états de Σ^δ sont itérés régulièrement et tous les états $\partial\Sigma^\delta$ sont itérés au moins une fois chacun, i.e. :*

$$\begin{array}{lll} \forall \xi \in \Sigma^\delta, & \forall m, \exists n, & N(\xi, n) = m \\ \forall \xi \in \partial\Sigma^\delta, & \exists n, & N(\xi, n) = 1 \end{array}$$

Le théorème suivant donne une condition générale pour que la fonction valeur calculée par un algorithme d'A/R converge vers la fonction valeur du problème continu.

Théorème 7 (Convergence des algorithmes d'A/R) *Soit F^δ un schéma d'approximation convergent (par exemple vérifiant les hypothèses du théorème 5) et V^δ les solutions de (2.39) et (2.40). Considérons un algorithme exhaustif (Ξ_n^δ, F^δ) dont le schéma d'algorithme $\widetilde{F_n^\delta}$ satisfait une propriété de "contraction faible" par rapport au schéma F^δ, alors, pour pour tout compact $\Omega \subset O$, pour tout $\varepsilon > 0$, il existe Δ tel que pour tout $\delta \leq \Delta$, en utilisant la règle d'itération suivante (à l'étape n) :*

$$\begin{array}{ll} V_{n+1}^\delta(\xi) \leftarrow \widetilde{F_n^\delta}\left[V_n^\delta\right](\xi) & \textit{pour } \xi \in \Xi_n^\delta \\ V_{n+1}^\delta(\xi) \leftarrow V_n^\delta(\xi) & \textit{pour } \xi \notin \Xi_n^\delta \end{array}$$

il existe N, pour tout $n \geq N$,

$$\sup_{\xi \in \Omega \cap (\Sigma^\delta \cup \partial\Sigma^\delta)} \left| V_n^\delta(\xi) - V(\xi) \right| \leq \varepsilon.$$

Ce théorème établit un résultat de convergence uniforme sur tout compact inclu dans O lorsque le pas de discrétisation δ tend vers 0 et le nombre d'itérations n tend vers l'infini.

Les propriétés requises dans la définition de "contraction faible" d'un schéma d'algorithme par rapport à un schéma d'approximation sont équivalentes, d'après la règle d'itération précédente, à :

- Pour tout $\xi \in \Sigma^\delta$,

$$\left| V_{n+1}^\delta(\xi) - V^\delta(\xi) \right| \leq (1 - k.\delta) \sup_{\xi \in \Sigma^\delta \cup \partial\Sigma^\delta} \left| V_n^\delta(\xi) - V^\delta(\xi) \right| + e(\delta).\delta \quad (3.2)$$

- Pour tout $\xi \in \partial\Sigma^\delta$,

$$\left| V_{n+1}^\delta(\xi) - V^\delta(\xi) \right| \leq k_R.\delta \quad (3.3)$$

Remarque 17 *En pratique, pour obtenir la limite lorsque $\delta \downarrow 0$ et $n \to \infty$, il paraît nécessaire de combiner une dynamique d'apprentissage (jouant sur le nombre d'itérations n) et une dynamique structurelle (jouant sur le pas de discrétisation δ), par exemple en partant d'une discrétisation grossière (δ élevé) qui serait affinée progressivement selon un critère estimant l'erreur d'approximation. Ceci est l'objet du chapitre 4.*

3.1.2 Preuve du théorème

Remarque 18 *Si la condition (3.2) était une propriété de contraction forte du style :*

$$\left|\widetilde{F_n^\delta}\left[V_n^\delta\right](\xi) - F^\delta[V^\delta](\xi)\right| \leq \lambda \sup_{\xi \in \Sigma^\delta \cup \partial\Sigma^\delta} \left|V_n^\delta(\xi) - V^\delta(\xi)\right| \tag{3.4}$$

avec une constante $\lambda < 1$, alors la convergence serait évidente puisque de (2.39) et du fait que tous les états sont régulièrement itérés, on déduirait que pour un δ donné, V_n^δ convergerait vers V^δ pour $n \to \infty$; et du fait de la convergence de V^δ vers V pour $\delta \downarrow 0$ (théorème 5) on déduirait que $V_n^\delta \to V$ lorsque $\delta \downarrow 0$ et $n \to \infty$.

Lorsque ce n'est pas le cas, on ne peut plus espérer que $V_n^\delta \to V^\delta$ pour $n \to \infty$. Cependant, si (3.2) est satisfaite, on prouve que pour tout $\varepsilon > 0$, il existe des valeurs suffisamment petites de δ telles que à une étape N, $|V_n^\delta - V^\delta| \leq \varepsilon$ pour $n \geq N$. Ce résultat combiné à la convergence du schéma mène à la convergence de l'algorithme lorsque $\delta \downarrow 0$ et $n \to \infty$.

Idée de la preuve

Nous savons, d'après la convergence du schéma d'approximation (théorème 5), que pour tout compact $\Omega \subset O$, pour tout $\varepsilon_1 > 0$, il existe un pas de discrétisation δ tel que :

$$\sup_{x \in \Omega} |V^\delta(x) - V(x)| \leq \varepsilon_1.$$

Définissons $E_n^\delta = \sup_{\xi \in \Sigma^\delta \cup \partial\Sigma^\delta} \left|V_n^\delta(\xi) - V^\delta(\xi)\right|$.

Si la propriété de contraction forte (3.4) était vraie, on aurait pour tout δ, la convergence de E_n^δ vers 0 pour $n \to \infty$. Puisque nous avons seulement la propriété de contraction faible (3.2) :

$$\left|V_{n+1}^\delta(\xi) - V^\delta(\xi)\right| \leq (1 - k.\delta)E_n^\delta + e(\delta).\delta$$

l'idée est de montrer (objet du lemme 2) que pour tout $\varepsilon_2 > 0$, il existe δ et une étape N, tels que pour $n \geq N$,

$$E_n^\delta = \sup_{\xi \in \Sigma^\delta \cup \partial\Sigma^\delta} |V_n^\delta(\xi) - V^\delta(\xi)| \leq \varepsilon_2. \tag{3.5}$$

Alors nous pouvons déduire que pour tout $\varepsilon > 0$, en définissant $\varepsilon_1 > 0$ et $\varepsilon_2 > 0$ tels que $\varepsilon_1 + \varepsilon_2 = \varepsilon$, on a :

$$\begin{aligned} \sup_{\xi \in \Omega \cap (\Sigma^\delta \cup \partial\Sigma^\delta)} |V_n^\delta(\xi) - V(\xi)| &\leq \sup_{x \in \Omega} |V^\delta(x) - V(x)| + \sup_{\xi \in \Sigma^\delta \cup \partial\Sigma^\delta} |V_n^\delta(\xi) - V^\delta(\xi)| \\ &\leq \varepsilon_1 + \varepsilon_2 = \varepsilon \end{aligned}$$

Une condition suffisante pour que $\sup \left| V_n^\delta(\xi) - V^\delta(\xi) \right| \leq \varepsilon_2$

Lemme 2 *Supposons qu'il existe une constante positive α telle que pour tout état $\xi \in \Xi_n^\delta$, la condition suivante est satisfaite :*

$$\begin{aligned} \textit{Si } E_n^\delta &> \varepsilon_2 \textit{ alors } \left| V_{n+1}^\delta(\xi) - V^\delta(\xi) \right| \leq E_n^\delta - \alpha & (3.6) \\ \textit{Si } E_n^\delta &\leq \varepsilon_2 \textit{ alors } \left| V_{n+1}^\delta(\xi) - V^\delta(\xi) \right| \leq \varepsilon_2 & (3.7) \end{aligned}$$

alors il existe N tel que pour $n \geq N, E_n^\delta \leq \varepsilon_2$.

Preuve. Puisque l'algorithme permet d'itérer tous les états $\xi \in \Sigma^\delta$ regulièrement, il existe un entier m tel que à l'étape $n+m$ tous les états $\xi \in \Sigma^\delta$ ont été itérés au moins une fois depuis l'étape n. Donc, on déduit de (3.6) et (3.7) que :

$$\begin{aligned} \text{Si } E_n^\delta &> \varepsilon_2 \text{ alors } \sup_{\xi \in \Sigma^\delta} |V_{n+m}^\delta(\xi) - V^\delta(\xi)| \leq E_n^\delta - \alpha \\ \text{Si } E_n^\delta &\leq \varepsilon_2 \text{ alors } \sup_{\xi \in \Sigma^\delta} |V_{n+m}^\delta(\xi) - V^\delta(\xi)| \leq \varepsilon_2. \end{aligned}$$

D'où l'on déduit qu'il existe N_1 tel que $\forall n \geq N_1$,

$$\sup_{\xi \in \Sigma^\delta} |V_n^\delta(\xi) - V^\delta(\xi)| \leq \varepsilon_2. \tag{3.8}$$

De plus, tous les états $\xi \in \partial\Sigma^\delta$ sont itérés au moins une fois avec les valeurs estimées $\widetilde{F_n^\delta}\left[V_n^\delta\right](\xi)$. Donc il existe N_2 tel que $\forall n \geq N_2$, pour tout $\xi \in \partial\Sigma^\delta$:

$$|V_n^\delta(\xi) - V^\delta(\xi)| = \left| \widetilde{F_{n-1}^\delta}\left[V_{n-1}^\delta\right](\xi) - F^\delta[V^\delta](\xi) \right| \leq k_R.\delta \leq \varepsilon_2 \tag{3.9}$$

pour tout $\delta \leq \Delta_1 = \frac{\varepsilon_2}{k_R}$.

Ainsi, de (3.8) et (3.9) nous déduisons que pour $n \geq N = \max\{N_1, N_2\}$,

$$E_n^\delta = \sup_{\xi \in \Sigma^\delta \cup \partial\Sigma^\delta} |V_n^\delta(\xi) - V^\delta(\xi)| \leq \varepsilon_2.$$

■

Lemme 3 *Pour tout $\varepsilon_2 > 0$, il existe Δ_2 tel que pour $\delta \leq \Delta_1$, les conditions (3.6) et (3.7) sont satisfaites.*

Preuve. Considérons une valeur $\varepsilon_2 > 0$. D'après la convergence de $e(\delta)$ vers 0 lorsque $\delta \downarrow 0$, il existe Δ_1 tel que pour $\delta \leq \Delta_1$ on a :

$$e(\delta) - k.\frac{\varepsilon_2}{2} \leq 0 \tag{3.10}$$

Montrons alors que (3.6) et (3.7) sont vraies. Soit $E_n^\delta > \varepsilon_2$, on déduit de (3.2) que :

$$\left|V_{n+1}^\delta(\xi) - V^\delta(\xi)\right| \leq (1 - k.\delta)E_n^\delta + e(\delta).\delta \leq E_n^\delta - k.\delta.\varepsilon_2 + e(\delta).\delta.$$

D'après (3.10),

$$\left|V_{n+1}^\delta(\xi) - V^\delta(\xi\right| \leq E_n^\delta - k.\delta.\frac{\varepsilon_2}{2} + e(\delta).\delta - k.\delta.\frac{\varepsilon_2}{2} \leq E_n^\delta - k.\delta.\frac{\varepsilon_2}{2}$$

et la condition (3.6) est vraie pour $\alpha = k.\delta.\frac{\varepsilon_2}{2}$.

Maintenant, si $E_n^\delta \leq \varepsilon_2$, on déduit de (3.2) que :

$$\left|V_{n+1}^\delta(\xi) - V^\delta(\xi)\right| \leq (1 - k.\delta)\frac{\varepsilon_2}{2} + \frac{\varepsilon_2}{2} + e(\delta)\delta - k.\delta\frac{\varepsilon_2}{2} \leq \frac{\varepsilon_2}{2} + \frac{\varepsilon_2}{2} = \varepsilon_2$$

et la condition (3.7) est vraie. ■

Démonstration du théorème 7

Preuve. Pour tout compact $\Omega \subset O$, pour tout $\varepsilon > 0$, on considère $\varepsilon_1 > 0$ et $\varepsilon_2 > 0$ tels que $\varepsilon_1 + \varepsilon_2 = \varepsilon$. D'après le lemme 3, pour $\delta \leq \Delta_1$, les conditions (3.6) et (3.7) sont vraies et par le lemme 2, il existe N, pour tout $n \geq N$,

$$E_n^\delta = \sup_{\xi \in \Sigma^\delta \cup \partial\Sigma^\delta} |V_n^\delta(\xi) - V^\delta(\xi)| \leq \varepsilon_2.$$

De plus, d'après la convergence du schéma d'approximation, le théorème 5 établit qu'il existe Δ_2 tel que pour tout $\delta \leq \Delta_2$,

$$\sup_{x \in \Omega} |V^\delta(x) - V(x)| \leq \varepsilon_1.$$

Ainsi, pour $\delta \leq \Delta = \min\{\Delta_1, \Delta_2\}$, pour tous espaces d'états discrets Σ^δ et $\partial\Sigma^\delta$ satisfaisant les propriétés du paragraphe 2.3.4, il existe N, pour tout $n \geq N$,

$$\begin{aligned}\sup_{\xi\in\Omega\cap(\Sigma^\delta\cup\partial\Sigma^\delta)} |V_n^\delta(\xi) - V(\xi)| &\leq \sup_{x\in\Omega} |V^\delta(x) - V(x)| + \sup_{\xi\in\Sigma^\delta\cup\partial\Sigma^\delta} |V_n^\delta(\xi) - V^\delta(\xi)| \\ &\leq \varepsilon_1 + \varepsilon_2 = \varepsilon.\end{aligned}$$

■

3.1.3 Comment concevoir des algorithmes convergents ?

En général, la propriété de contraction forte (3.4) est impossible à obtenir à moins d'avoir une connaissance parfaite de la dynamique d'état f et des fonctions renforcement r et R. Dans l'approche A/R, ces informations sont estimées et approchées au cours de l'apprentissage. Ainsi, les valeurs estimées $\widetilde{F_n^\delta}\left[V_n^\delta\right](\xi)$ sont imparfaites mais peuvent être suffisamment précises pour satisfaire la propriété de contraction faible (3.2). **La recherche de telles approximations est la base pour la conception d'algorithmes d'A/R convergents.**

Dans le paragraphe 3.2, nous indiquons une méthode générale pour concevoir des algorithmes indirectes, et dans le paragraphe 3.3, nous présentons deux algorithmes directs (respectivement basés sur des schémas aux EF et aux DF). Les preuves de la convergence de ces algorithmes sont données dans les annexes B, C, E et D.

3.2 Algorithmes indirects

L'idée directrice est de construire un modèle, c'est à dire une approximation de la dynamique d'état f (et a dans le cas stochastique) et des fonctions renforcement r et R aux points ξ à partir d'échantillons de trajectoires passant dans leur voisinage. Ainsi, si une trajectoire $x_k(t)$ passe près d'un point ξ à l'instant t_k en choisissant un contrôle u pendant une durée τ_k, dans le cas déterministe, $f(\xi, u)$ est approché par $\frac{x_k(t_k+\tau_k)-x_k(t_k)}{\tau_k}$ (dans le cas stochastique, $f(\xi, u)$ est approché par une moyenne de telles approximations et $a(\xi, u)$ par leur écarts-type). Ensuite, à partir du modèle, nous définissons des probabilités de transitions approchées que nous utilisons à la place des vraies valeurs $p(\xi, \xi', u)$ dans la règle itérative (3.1).

Nous supposons dans ce qui suit que la dynamique d'état f est minoré (il existe m_f tel que $\|f\|_1 \geq m_f$).

3.2.1 Le cas déterministe

- **Estimation pour $\xi \in \Sigma^\delta$:** Nous construisons pour tout $\xi \in \Sigma^\delta$, tout $u \in U^\delta$, un modèle $\widetilde{f}$ et $\widetilde{r}$, approximations de f et r à partir de trajectoires $x(t)$ passant au voisinage des ξ : on considère des états $x = x(t_0)$ et $y = x(t_0 + \tau)$ tels que :

  - $x \in N(\xi)$ voisinage de ξ de diamètre inférieur à $k_N.\delta$ pour une constante positive k_N.

  - le contrôle u est constant pour $t \in [t_0, t_0 + \tau]$,

  - le temps τ satisfait pour deux constantes positive k_1 et k_2,

$$k_1\delta \leq \tau \leq k_2\delta. \tag{3.11}$$

  On peut alors établir le modèle pour l'état ξ, le contrôle u :

$$\begin{aligned} \widetilde{f}(\xi, u) &= \frac{y - x}{\tau} \\ \widetilde{r}(\xi, u) &= r(x, u) \end{aligned}$$

  Ce qui permet de calculer les probabilités approchées $\widetilde{p}(\xi, \xi', u)$ en utilisant dans (2.27) ou (2.35) le modèle $\widetilde{f}$ au lieu de f et obtenir la règle itérative suivante basée sur (3.1) :

$$\begin{aligned} Q^\delta_{n+1}(\xi, u) &\leftarrow \gamma^{\frac{\delta}{\|\widetilde{f}(\xi,u)\|_1}} . \sum_{\xi'} \widetilde{p}(\xi, \xi', u).V^\delta_n(\xi') + \frac{\delta}{\left\|\widetilde{f}(\xi, u)\right\|_1}.\widetilde{r}(\xi, u) \\ \text{avec } V^\delta_n(\xi) &= \sup_{u \in U^\delta} Q^\delta_n(\xi, u). \end{aligned} \tag{3.12}$$

  qui peut être utilisée en méthode de PD "hors-ligne" (synchrone, Gauss-Seidel, asynchrone) ou "en-ligne" (par exemple en itérant $V^\delta_n(\xi)$ dès qu'une trajectoire sort du voisinage de ξ) (voir [Ber87], [BBS91], [Put94], [BBS95], [BT96]).

- **Estimation pour $\xi \in \partial\Sigma^\delta$** : Lorsqu'une trajectoire $x(t)$ sort de l'espace d'état en $y \in \partial O$, nous considérons les points $\xi \in \partial\Sigma^\delta$ dont les voisinages $N(\xi)$ correspondants contiennent y et nous itérons leur valeur selon :

$$V^\delta_{n+1}(\xi) \leftarrow R(y) \tag{3.13}$$

Le théorème suivant, dont la démonstration se trouve en annexe B, énonce que l'algorithme consistant à itérer régulièrement tous les états $\xi \in \Sigma^\delta$ avec la règle (3.12) et tous les états $\xi \in \partial\Sigma^\delta$ (au moins une fois chacun) avec (3.13) satisfait les conditions du théorème 7, ce qui en fait un algorithme d'A/R indirect convergent dans le cas déterministe.

Théorème 8 (Convergence de l'algorithme indirect. Cas déterministe) *Les règles itératives (3.12) et (3.13) satisfont les propriétés (3.2) (resp. (3.3)) donc le théorème 7 s'applique et l'algorithme d'A/R indirect dans le cas déterministe ainsi défini est convergent.*

3.2.2 Le cas stochastique

- **Estimation pour $\xi \in \Sigma^\delta$:** On approche la dynamique d'état f et a en construisant un modèle $\widetilde{f}$ et $\widetilde{a}$ à partir d'échantillons de trajectoires $x_k(t)$: on considère des suites d'états successifs $x_k = x_k(t_k)$ et $y_k = x_k(t_k + \tau_k)$ tels que :

  - $\forall t \in [t_k, t_k + \tau_k]$, $\quad x(t) \in N(\xi)$ voisinage de ξ de diamètre inférieur à $k_N.\delta$ pour une constante positive k_N,

  - le contrôle u est constant pour $t \in [t_k, t_k + \tau_k]$,

  - τ_k satisfait pour deux constantes positive k_1 et k_2,

$$k_1\delta \leq \tau_k \leq k_2\delta. \tag{3.14}$$

  On peut alors affiner le modèle pour l'état ξ, le contrôle u en calculant (de manière incrémentale) :

$$\begin{aligned}
\widetilde{f}_n(\xi,u) &= \frac{1}{n}\sum_{k=1}^{n}\frac{y_k - x_k}{\tau_k} \\
\widetilde{a}_n(\xi,u) &= \frac{1}{n}\sum_{k=1}^{n}\frac{\left(y_k - x_k - \tau_k.\widetilde{f}_n(\xi,u)\right)\left(y_k - x_k - \tau_k.\widetilde{f}_n(\xi,u)\right)^T}{\tau_k} \\
\widetilde{r}_n(\xi,u) &= \frac{1}{n}\sum_{k=1}^{n} r(x_k,u)
\end{aligned} \tag{3.15}$$

  ainsi que le temps $\widetilde{\tau}_n(x,u)$ et les probabilités de transition $\widetilde{p}_n(\xi,\xi',u)$ estimés en remplaçant dans (2.59) et (2.60), f et a par $\widetilde{f}_n$ et $\widetilde{a}_n$:

$$\widetilde{\tau}(\xi,u) = \frac{\delta^2}{\sum_{i=1}^{d}\left[\delta|\widetilde{f}_i(\xi,u)| + \widetilde{a}_{ii}(\xi,u) - \frac{1}{2}\sum_{j\neq i}|\widetilde{a}_{ij}(\xi,u)|\right]}$$

$$\begin{aligned}
\widetilde{p}(\xi,\xi \pm \delta e_i,u) &= \frac{\widetilde{\tau}(\xi,u)}{2\delta^2}\left[2\delta|\widetilde{f}_i^{\pm}(\xi,u)| + \widetilde{a}_{ii}(\xi,u) - \sum_{j\neq i}|\widetilde{a}_{ij}(\xi,u)|\right], \\
\widetilde{p}(\xi,\xi + \delta e_i \pm \delta e_j,u) &= \frac{\widetilde{\tau}(\xi,u)}{2\delta^2}\widetilde{a}_{ij}^{\pm}(\xi,u) \text{ pour } i \neq j,
\end{aligned}$$

$$\begin{aligned}\widetilde{p}(\xi, \xi - \delta e_i \pm \delta e_j, u) &= \frac{\widetilde{\tau}(\xi,u)}{2\delta^2}\widetilde{a}_{ij}^{\mp}(\xi,u) \text{ pour } i \neq j,\\ \widetilde{p}(\xi,\xi',u) &= 0 \text{ sinon.}\end{aligned}$$

(en utilisant les notations allégées $\widetilde{f}$, $\widetilde{a}, \widetilde{\tau}$ et $\widetilde{p}$ au lieu de $\widetilde{f}_n, \widetilde{a}_n, \widetilde{\tau}_n$ et $\widetilde{p}_n$). Nous obtenons alors la règle itérative suivante :

$$V_{n+1}^{\delta}(\xi) \leftarrow \sup_{u \in U^{\delta}} \left[\gamma^{\widetilde{\tau}(x,u)} \sum_{\xi'} \widetilde{p}(\xi,\xi',u) V_n^{\delta}(\xi') + \widetilde{\tau}(x,u)\widetilde{r}(\xi,u) \right] \quad (3.16)$$

qui, de même que dans le cas déterministe, peut être utilisée comme méthode de PD "hors-ligne" ou "en-ligne".

- **Estimation pour** $\xi \in \partial\Sigma^{\delta}$: De manière analogue au cas déterministe, lorsqu'une trajectoire $x(t)$ sort en $y \in \partial O$, nous itérons la valeur des points $\xi \in \partial\Sigma^{\delta}$ dont les voisinages $N(\xi)$ contiennent y selon :

$$V_{n+1}^{\delta}(\xi) \leftarrow R(y) \quad (3.17)$$

De même que dans le cas déterministe, le théorème suivant, dont la démonstration se trouve en annexe C (ou dans [Mun97d]), énonce que l'algorithme utilisant les règle itératives (3.16) et (3.17) satisfait les conditions du théorème 7, ce qui en fait un algorithme d'A/R indirect convergent dans le cas stochastique.

Théorème 9 (Convergence de l'algorithme indirect. Cas stochastique) *Les règle itératives (3.16) et (3.17) satisfont les conditions (3.2) et (3.3), donc le théorème 7 s'applique et l'algorithme d'A/R indirect dans le cas stochastique ainsi défini converge.*

3.3 Algorithmes directs

L'approche directe pour l'A/R consiste à définir des estimations $\widetilde{F_n^{\delta}}\left[V_n^{\delta}\right](\xi)$ (satisfaisant les conditions (3.2) ou (3.3)) et à les itérer en-ligne sans construction a priori d'un modèle de la dynamique d'état. Ici, nous nous limitons à l'étude du cas déterministe.

3.3.1 L'algorithme d'A/R aux EF

Soit une triangulation Σ^{δ} satisfaisant les propriétés de la section 2.3.3. Nous supposons qu'il existe une constante k_{Σ} telle que

$$dist(\Sigma^{\delta}, O) \leq k_{\Sigma}\delta^2 \quad (3.18)$$

Nous considérons le schéma aux EF (2.37) avec le temps $\tau(\xi, u)$ défini de telle façon que :

$\eta(\xi, u) = \xi + \tau(\xi, u).f(\xi, u)$ est la projection de ξ dans la direction $f(\xi, u)$ sur le côté opposé du simplexe (voir figure 3.1).

Nous supposons que les simplexes sont non-dégénérés : il existe une constante $k_\rho > 0$ telle que

le rayon des sphères inscrites dans chaque simplexe est $\geq k_\rho \delta$. (3.19)

De plus, afin d'obtenir (2.33), nous supposons que la dynamique d'état f est bornée inférieurement : il existe m_f tel que pour tout x et u,

$$||f(x, u)|| \geq m_f. \tag{3.20}$$

Supposons qu'une trajectoire $x(t)$ du système dynamique traverse un simplexe T avec un contrôle u constant à l'intérieur de ce dernier. La trajectoire entre en $x = x(t_1)$ le point d'entrée par le côté d'entrée (qui est un simplexe de dimension $d-1$) de T, noté $T_{entrée}$. Soit elle ressort de T par un côté de sortie (qui est aussi un simplexe de dimension $d-1$), noté T_{sortie}, en $y = x(t_2)$ (dans ce cas $y \in O$, voir figure 3.1), soit elle sort de l'espace d'état à l'intérieur de T en $y \in \partial O$ (voir figure 3.2).

Soit $\tau = t_2 - t_1$ le temps mis par la trajectoire pour traverser le simplexe.

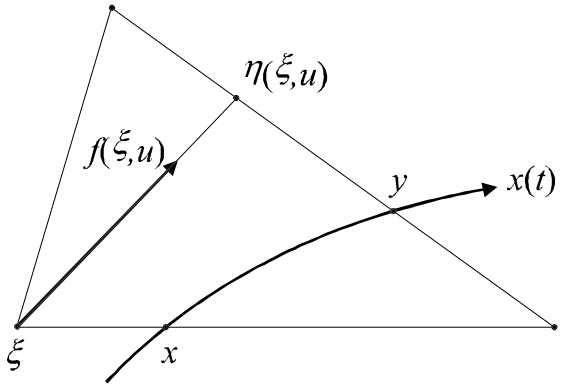

Figure3.1: Une trajectoire traversant un simplexe. $\eta(\xi, u)$ est la projection de ξ sur le côté opposé du simplexe. $\frac{y-x}{\lambda_\xi(x)}$est une bonne approximation de $\eta(\xi, u) - \xi$.

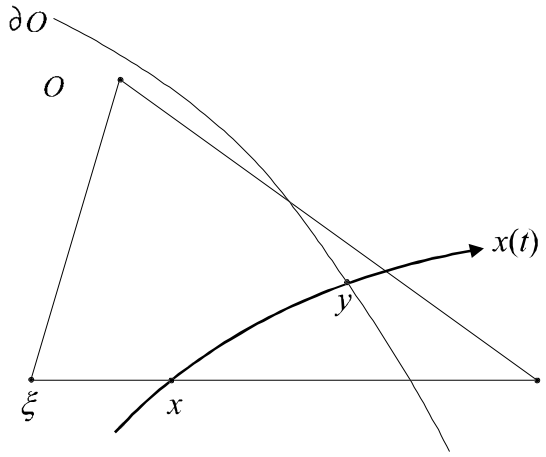

Figure 3.2: Une trajectoire touche la frontière en $y \in \partial O$.

Approximations considérées : puisque la dynamique d'état est inconnue du système, il en est de même des valeurs $\tau(\xi, u)$ et $\eta(\xi, u)$. Nous faisons alors les approximations suivantes :

$$\tau(\xi, u) \text{ est approché par } \frac{\tau}{\lambda_\xi(x)} \tag{3.21}$$

(où $\lambda_\xi(x)$ est la $\xi^{ième}$-coordonnée barycentrique de x dans le simplexe)

$$\eta(\xi, u) \text{ est approché par } \xi + \frac{y - x}{\lambda_\xi(x)} \tag{3.22}$$

L'idée sous-jacente à ces approximations est que pour des valeurs faibles de δ, la dynamique d'état f est presque constante dans le simplexe et donc, d'après le théorème de Thales, $\frac{y-x}{\lambda_\xi(x)}$ est proche de $\eta(\xi, u) - \xi$. Ces approximations n'utilisent que les informations des points d'entrée et de sortie (x et y), du temps de parcours τ de la trajectoire dans le simplexe et de la coordonnée barycentrique $\lambda_\xi(x)$ (laquelle peut être calculée dès que le système connaît les sommets du côté d'entrée du simplexe).

Ainsi, grâce à la linéarité de V^δ sur le simplexe T, $V^\delta(\eta(\xi, u))$ peut être approché par $V^\delta(\xi) + \frac{V^\delta(y) - V^\delta(x)}{\lambda_\xi(x)}$.

De plus, $r(\xi, u)$ est approché par le renforcement courant $r(x, u)$ au point d'entrée.

Algorithme : Considérons une constante $\lambda > 0$ choisie arbitrairement proche de 0.

- Supposons que la trajectoire sorte du simplexe en $y \in O$ (figure 3.1). Si nous avons pour tout ξ_i de $T_{entrée} \cap T_{sortie}$,

$$\lambda_{\xi_i}(y) \geq \lambda_{\xi_i}(x) + \lambda \tag{3.23}$$

(afin que le vecteur $y - x$ appartienne strictement au cône de sommet ξ et de base T_{sortie}), ainsi que

$$\lambda_{\xi}(x) \geq \lambda \tag{3.24}$$

(afin que $\lambda_\xi(x)$ ne soit pas trop proche de 0), alors l'algorithme consiste à itérer la qualité $Q_n^\delta(\xi, u)$ selon :

$$\begin{aligned} Q_{n+1}^\delta(\xi,u) &\leftarrow \gamma^{\frac{\tau}{\lambda_\xi(x)}}.\left[V_n^\delta(\xi) + \frac{V_n^\delta(y) - V_n^\delta(x)}{\lambda_\xi(x)}\right] + \frac{\tau}{\lambda_\xi(x)}.r(x,u) \\ \text{avec } V_n^\delta(\xi) &= \sup_{u \in U^\delta} Q_n^\delta(\xi,u). \end{aligned} \tag{3.25}$$

- Supposons maintenant que la trajectoire sorte de l'espace d'état dans le simplexe en $y \in \partial O$, (figure 3.2), alors on itère le sommet $\xi' \in \partial\Sigma^\delta$ le plus proche de y selon :

$$V_{n+1}^\delta(\xi') \leftarrow R(y). \tag{3.26}$$

Le théorème suivant, dont la démonstration est donnée en annexe D (ou dans [Mun96a]), énonce que l'algorithme consistant à itérer régulièrement tous les états $\xi \in \Sigma^\delta$, pour tous les contrôles $u \in U^\delta$ selon la règle itérative (3.25) et les états $\xi \in \partial\Sigma^\delta$ au moins une fois selon (3.26) satisfait les conditions du théorème 7, ce qui en fait un algorithme d'A/R direct convergent.

Théorème 10 (Convergence de l'algorithme direct aux EF) *Les règles itératives (3.25) et (3.26) satisfont les conditions (3.2) et (3.3) donc le théorème 7 s'applique et l'algorithme d'A/R direct aux EF ainsi défini converge.*

3.3.2 L'algorithme d'A/R aux DF

On considère les grilles G^δ et ∂G^δ composées de cases régulières de côté δ telles que les centres des cases appartiennent à l'ensemble $\delta\mathbb{Z}^d$. G^δ est composé des cases X inclues dans O et ∂G^δ, appelée frontière de G^δ, est composé des cases ayant une intersection non vide avec ∂O. Les centres des cases des grilles G^δ et ∂G^δ forment respectivement des ensembles notés Σ^δ et $\partial\Sigma^\delta$ (voir figure 3.3) qui vérifient les propriétés du paragraphe 2.3.4. L'algorithme, décrit dans

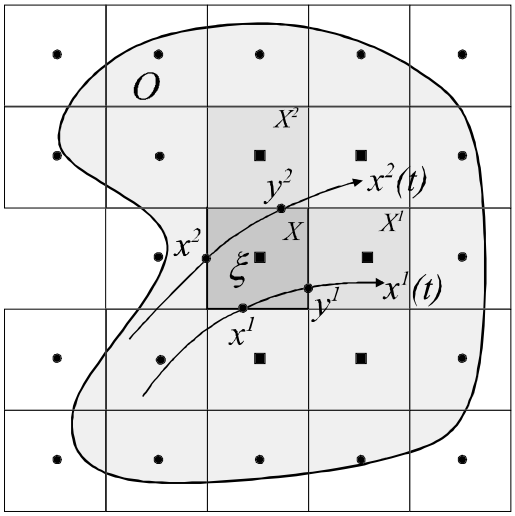

Figure 3.3: Les grilles G^δ et ∂G^δ recouvrant O sont composées de cases dont les centres ξ forment les ensembles Σ^δ (points carrés) et $\partial\Sigma^\delta$ (les ronds). Deux trajectoires $x^1(t)$ et $x^2(t)$ traversent la case X et ressortent vers les cases X^1 et X^2.

[Mun97b] (et dans [Mun96b]) est basé sur le schéma aux DF du cas à horizon infini $F^\delta_{FD_\infty}$ (2.29).

Soit $V^\delta_n(\xi)$ la valeur de la case X de centre ξ calculée par l'algorithme à l'étape n. Soit $Q^\delta_n(\xi, u)$ les qualités correspondantes à une case et un contrôle u (qui sont telles que $V^\delta_n(\xi) = \sup_u Q^\delta_n(\xi, u)$). Nous introduisons les qualités $Q^\delta_n(\xi, u, k)$, représentant des *contributions partielles* à $Q^\delta_n(\xi, u)$ et correspondant à une transition de la case X à la case X^k de centre $\xi + \delta e_k$ pour un contrôle u.

Soit $\lambda > 0$, constante choisie arbitrairement proche de 0.

Algorithme : Considérons une trajectoire $x^1(t)$ entrant dans X en un point x^1 ; alors un contrôle $u \in U^\delta$ est choisi et conservé constant pendant que la trajectoire reste dans X.

- Si $X \in G^\delta$ alors la trajectoire sort de X en un point $y^1 \in X \cap X^{k_1}$ pour une case voisine X^{k_1} de centre $\xi^{k_1} = \xi + \delta e_{k_1}$ avec $k_1 \in \{1..d\}$ (par exemple $y^1 \in X \cap X^1$ ($k_1 = 1$) sur la figure 3.3). L'idée principale est d'approcher $f(\xi, u)$ par $\frac{y^1 - x^1}{\tau^1}$.

  Soit $r(z^1, u)$ le renforcement courant obtenu en un point z^1 de la trajectoire à l'intérieur de X. Soit τ^1 le temps de parcours de la trajectoire dans X.

Supposons que le vecteur $y^i - x^i$ satisfasse l'hypothèse :

$$\left\|y^i - x^i\right\|_1 \geq \lambda.\delta \tag{3.27}$$

(afin que $y^i - x^i$ ne soit pas de norme trop proche de 0), alors on itère $Q_n^\delta(\xi, u, k_i)$, avec $i = 1$, selon :

$$Q_{n+1}^\delta(\xi, u, k_i) \leftarrow \gamma^{\frac{\delta.\tau^i}{\|y^i - x^i\|_1}} V_n^\delta(\xi^{k_i}) + \frac{\delta.\tau^i}{\|y^i - x^i\|_1} r(z^i, u) \tag{3.28}$$

Au cours des simulations, lorsqu'une autre trajectoire $x^2(t)$ traverse X de x^2 à $y^2 \in X \cap X^{k_2}$ (avec $X^{k_2} \neq X^{k_1}$) avec le même contrôle u, nous itérons $Q_n^\delta(\xi, u, k_2)$ selon la règle (3.28) avec $i = 2$.

Les valeurs $Q_n^\delta(\xi, u, k_i)$ sont itérées jusqu'à ce que des trajectoires $x^i(t)$ aient effectuées avec un contrôle u les transitions de X vers toutes les cases X^{k_i} (pour $i = 1..d$) voisines (donc $k_1 \neq k_2 \neq ... \neq k_d$) au moins une fois, menant ainsi à itérer selon (3.28) d fois (dans le cas général où la dynamique d'état $f(\xi, u)$ n'est pas parallèle à une dimension de l'espace).

Alors nous considérons un vecteur $y - x = y^l - x^l$ pour une trajectoire $x^l(t)$ (par exemple, la dernière) et nous itérons la qualité $Q_n^\delta(\xi, u)$ selon :

$$Q_{n+1}^\delta(\xi, u) \leftarrow \sum_{i=1}^{d} \frac{|y_i - x_i|}{\|y - x\|_1} Q_n^\delta(\xi, u, i) \tag{3.29}$$

où x_i signifie la $i^{ème}$ coordonnée de x. Par exemple, sur la figure 3.3, si $x^1(t)$ est exécutée en premier, $Q_n^\delta(\xi, u, 1)$ est itérée puis lorsque $x^2(t)$ est exécutée, $Q_n^\delta(\xi, u, 2)$ puis $Q_n^\delta(\xi, u)$ sont itérées.

- Si $X \in \partial G^\delta$ et si la trajectoire touche la frontière $y^i \in \partial O$ dans X, alors nous itérons selon :

$$V_{n+1}^\delta(\xi) = R(y^i). \tag{3.30}$$

Le théorème suivant, dont la preuve est donnée en annexe E ou dans [Mun97b], énonce que l'algorithme consistant à itérer régulièrement tous les états $\xi \in \Sigma^\delta$, pour tous les contrôles $u \in U^\delta$ selon la règle (3.29) et tous les états $\xi \in \partial\Sigma^\delta$ selon (3.30) satisfait les conditions du théorème 7, ce qui en fait un algorithme d'A/R direct convergent.

Théorème 11 (Convergence de l'algorithme direct aux DF) *Les règles itératives (3.29) et (3.30) satisfont les conditions (3.2) et (3.3), donc le théorème 7 s'applique et l'algorithme d'A/R direct aux DF ainsi défini converge.*

Chapitre 4

Une double dynamique

Les algorithmes d'A/R décrits au chapitre précédent convergent vers la fonction valeur lorsque le nombre d'itérations n tend vers l'infini et lorsque le pas de discrétisation δ tend vers 0. En effet, pour un pas δ fixé, les valeurs V_n^δ ne convergent pas vers celles du schéma V^δ lorsque $n \to \infty$ (voir la discussion du paragraphe 3.1.2 à propos des propriétés de contraction forte (3.4) et faible (3.2)). En pratique, il convient donc de modifier l'espace discrétisé Σ^δ au cours de l'apprentissage. Nous proposons de combiner les deux dynamiques suivantes :

- *Une dynamique d'apprentissage :* pour une discrétisation donnée Σ^{δ_i}, nous itérons les valeurs associées $V_n^{\delta_i}$ selon un algorithme d'A/R décrit au chapitre 3,
- *Une dynamique structurelle :* à partir de Σ^{δ_i}, on génère un nouvel espace discrétisé $\Sigma^{\delta_{i+1}}$ "plus fin", au sens : $\lim_{i\to\infty} \delta_i = 0$.

Cette dynamique structurelle, doit résoudre le problème de l'explosion combinatoire du nombre de valeurs à estimer lorsque le pas de discrétisation δ_i tend vers 0, surtout lorsque la dimension de l'espace d'état est importante.

Nous proposons dans les exemples suivants une dynamique structurelle dite "top-down", c'est à dire du "général vers le spécifique" : la discrétisation initiale est composée d'un nombre restreint d'états ; la dynamique structurelle consiste alors à discrétiser davantage (en ajoutant des états) aux endroits où l'on estime que l'erreur d'approximation de la fonction valeur est importante.

Exemple basé sur une dynamique d'apprentissage aux EF.

Dans [Mun97c], nous avons étudié des triangulations de Delaunay (voir par exemple [Mid93], [Rup94]) adaptatives établies à partir d'un ensemble de sommets Σ^δ. La double dynamique est la suivante :

- Apprentissage : Pour un ensemble de points Σ^δ, nous utilisons l'algorithme d'A/R direct aux éléments finis décrit au paragraphe 3.3.1 sur la triangulation de Delaunay construite à partir de Σ^δ.

- Structurelle : Nous ajoutons des points à l'ensemble des sommets Σ^δ et considérons la triangulation de Delaunay résultante.

Nous renvoyons le lecteur à [Mun97c] pour la description détaillée de cet algorithme ainsi qu'une illustration de la méthode par une simulation numérique pour un problème d'atteinte de cible dans un espace de dimension 2.

Exemple basé sur une dynamique d'apprentissage aux DF.

M. Rezgui [Rez97] utilise un algorithme d'A/R aux DF avec une discrétisation de l'espace d'état utilisant une résolution variable (basé sur les idées de résolutions adaptatives de Moore [Moo91], [Moo94]). La double dynamique est celle-ci :

- Apprentissage : algorithme d'A/R direct aux différences finis décrit au paragraphe 3.3.2.

- Structurelle : selon un critère local, des cases sont subdivisées afin d'obtenir une discrétisation à résolution variable.

Cet algorithme est illustré dans [Rez97] par une simulation numérique pour un problème d'atteinte de cible en 2 dimensions (le "mountain car problem").

Nous pouvons aussi envisager de construire une dynamique structurelle comportant des processus "bottom up", c'est à dire du "spécifique vers le général", ou mécanisme de généralisation, en supprimant certains états de la discrétisation lorsque ceux-ci ne sont pas nécessaires. Ainsi des processus de spécification et de généralisation pourraient être en œuvre simultanément afin de déterminer localement le niveau de résolution de la discrétisation adéquat pour la qualité d'approximation souhaitée.

Chapitre 5

Conclusion

Ce travail est une tentative pour fournir un support théorique à l'A/R dans le cas continu. Après avoir énoncé les problèmes rencontrés lorsque l'on utilise des algorithmes de descente de gradient avec des système d'approximation de fonctions ou des algorithmes d'A/R en discret (de type Q-learning) calculant des moyennes selon les fréquences d'occurrences, nous avons proposé une approche, basée sur l'utilisation des propriétés géométriques de la dynamique d'état, permettant d'obtenir la convergence des algorithmes développés.

Nous avons situé l'A/R dans le formalisme du contrôle optimal, énoncé l'équation de Hamilton-Jacobi-Bellman, introduit les notions de solutions de viscosité et présenté des méthodes de résolution numérique par l'utilisation de schémas d'approximation.

Nous avons proposé une méthode générale pour concevoir des algorithmes convergents, basée sur la notion de schéma d'algorithme vérifiant une propriété de "contraction faible" par rapport à un schéma d'approximation. Nous avons donné des exemples d'algorithmes convergents utilisant des méthodes directes, indirectes, basés sur des schémas aux DF, aux EF, pour une dynamique d'état déterministe ou stochastique.

Il apparaît de nombreuses voies de recherche afin d'approfondir ce travail :

- l'étude des ordres de convergence (quel pas de discrétisation δ faut-il choisir pour obtenir une erreur d'approximation ε souhaitée ?),
- l'étude des algorithmes basés sur des schémas d'approximation multi-résolution,
- l'étude de méthodes directes pour une dynamique d'état stochastique,
- l'étude de dynamiques structurelles permettant de traiter des problèmes dans des espaces de dimension élevée,

- la réalisation de simulations numériques pour tester et comparer les différents algorithmes,
- l'implémentation des algorithmes pour des problèmes réels de contrôle,
- l'établissement d'une analogie cognitive entre la dynamique structurelle et un processus de catégorisation (voir [Mun97a]),
- etc.

Annexe A

Convergence des schémas

A.1 Idée principale de la démonstration.

On utilise la méthode de Barles and Perthame (voir [BP88]). Tout d'abord, on donne une définition des solutions de viscosité discontinues. Ensuite, on définit la fonction limite supérieure V_{sup} et la fonction limite inférieure V_{inf} et on montre (en reprenant la démonstration de Barles et Souganidis [BS91]) dans le lemme 4 que V_{sup} (respectivement V_{inf}) est une sous-solution (resp. sur-solution) de viscosité. Puis on utilise le résultat d'unicité forte (lemme 1) qui établit que si les conditions sur la dynamique d'état à la frontière (hypothèse 1) sont satisfaites, alors les sous-solutions sont inférieures aux sur-solutions, donc que $V_{\text{sup}} \leq V_{\text{inf}}$. Or, par définition $V_{\text{sup}} \geq V_{\text{inf}}$, donc $V_{\text{sup}} = V_{\text{inf}} = V$ et la fonction limite V est la solution de viscosité de l'équation de HJB donc la fonction valeur du problème.

A.2 Définition des solutions de viscosité discontinues

Nous rappelons que les enveloppes semi-continues supérieurement W^* et semi-continues inférieurement W_* d'une fonction réelle W sont :

$$\begin{aligned} W^*(t,x) &= \limsup_{(s,y)\to(t,x)} W(s,y) \\ W_*(t,x) &= \liminf_{(s,y)\to(t,x)} W(s,y) \end{aligned}$$

Définition 12 *Soit W une fonction réelle locallement bornée, définie dans $\overline{Q}$,*

- *W est une sous-solution de viscosité de $H(t,x,W,DW)=0$ dans $\overline{Q}$ si pout toute fonction $\varphi \in C^1(\overline{Q})$, pour tout $(t,x) \in \overline{Q}$ maximum local de $W^* - \varphi$ tel que $W^*(t,x) = \varphi(t,x)$, on ait :*

$$H_*(t,x,\varphi(x),D\varphi(x)) \leq 0$$

- *W est une sur-solution de viscosité de $H(t,x,W,DW)=0$ dans $\overline{Q}$ si pout toute fonction $\varphi \in C^1(\overline{Q})$, pour tout $(t,x) \in \overline{Q}$ minimum local de $W_* - \varphi$ tel que $W_*(t,x) = \varphi(t,x)$, on ait :*

$$H^*(t,x,\varphi(x),D\varphi(x)) \geq 0$$

- *W est une solution de viscosité si W est une sous-solution et une sur-solution de viscosité de $H(t,x,W,DW)=0$ dans $\overline{Q}$.*

A.3 V_{sup} et V_{inf} sont des sous- et sur-solutions de viscosité

Lemme 4 *Les deux fonctions limites V_{sup} et V_{inf}*

$$\begin{aligned} V_{\text{sup}}(t,x) &= \limsup_{\substack{\delta \downarrow 0 \\ (s,\xi) \to (t,x)}} V^\delta(s,\xi) \\ V_{\text{inf}}(t,x) &= \liminf_{\substack{\delta \downarrow 0 \\ (s,\xi) \to (t,x)}} V^\delta(s,\xi) \end{aligned}$$

sont respectivement sous- et sur-solution de viscosité.

Preuve. Montrons que V_{sup} est une sous-solution. La preuve que V_{inf} est une sur-solution est similaire. Soit φ une fonction teste régulière telle que $V_{\text{sup}} - \varphi$ atteint un maximum (qui peut être supposé stricte) en (t,x) et tel que $V_{\text{sup}}(t,x) = \varphi(t,x)$. Soit δ_n une suite tendant vers zéro. Alors $V^{\delta_n} - \varphi$ atteint un maximum en (t_n,ξ_n) qui tend vers (t,x) lorsque δ_n tend vers 0. Donc, pour tout $\xi \in \Sigma^{\delta_n}$,

$$V^{\delta_n}(t_n,\xi) - \varphi(t_n,\xi) \leq V^{\delta_n}(t_n,\xi_n) - \varphi(t_n,\xi_n)$$

D'après (2.41), on a :

$$F^{\delta_n}\left[V^{\delta_n}(t_n,\xi) - V^{\delta_n}(t_n,\xi_n) - \varphi(t_n,\xi_n)\right] \leq F^{\delta_n}\left[\varphi(t_n,\xi)\right]$$

De (2.42), on déduit :

$$F^{\delta_n}\left[V^{\delta_n}(t_n,.)\right](\xi_n) - (1+O(\delta_n))\left[V^{\delta_n}(t_n,\xi_n) - \varphi(t_n,\xi_n)\right] \leq F^{\delta_n}\left[\varphi(t_n,.)\right](\xi_n)$$

Or par (2.43), $F^{\delta_n}\left[V^{\delta_n}\right] = V^{\delta_n}$, donc :

$$\frac{1}{\delta_n} O(\delta_n) \left[V^{\delta_n}(t_n, \xi_n) - \varphi(t_n, \xi_n)\right] \leq \frac{1}{\delta_n} \left[F^{\delta_n}\left[\varphi(t_n, .)\right](\xi_n) - \varphi(t_n, \xi_n)\right].$$

Puisque $V^{\delta_n}(t_n, \xi_n) - \varphi(t_n, \xi_n)$ tend vers 0, la partie gauche de cette inégalité tend vers 0 pour $\delta_n \downarrow 0$. Donc, d'après (2.45), on a :

$$H(t, x, \varphi, D\varphi) \geq 0.$$

et V_{sup} est sous-solution de viscosité. ■

A.4 Preuve du théorème

Preuve du théorème 5. Le lemme 4 établit que la fonction limite supérieure V_{sup} et la fonction limite inférieure V_{inf} sont respectivement des sous- et sur-solutions de viscosité de l'équation de HJB. Le résultat d'unicité forte du lemme 1 donne $V_{\text{sup}} \leq V_{\text{inf}}$. Mais d'après leur définition $V_{\text{sup}} \geq V_{\text{inf}}$, donc $V_{\text{sup}} = V_{\text{inf}} = V$ et le schéma d'approximation V^{δ} converge vers la fonction limite V qui est solution de viscosité de l'équation de HJB donc la fonction valeur du problème et (2.46) est satisfaite. ■

Annexe B

Convergence algorithme A/R indirect, cas déterministe

A partir d'une trajectoire $x(t)$ passant aux voisinages d'un point ξ, on considère des états $x = x(t_0)$, $y = x(t_0+\tau)$ vérifiant les propriétés du paragraphe 3.2.1, ce qui permet de construire le modèle de la dynamique d'état :

$$\widetilde{f}(\xi, u) = \frac{y - x}{\tau}.$$

Le paragraphe B.1.1 donne une comparaison entre le temps $\frac{\delta}{||f(\xi,u)||_1}$ utilisé par le schéma $F^{\delta}_{DF_\infty}$ dans l'équation de PD (2.28) et le temps approché $\frac{\delta}{\left\|\widetilde{f}(\xi,u)\right\|_1}$ utilisé par l'algorithme (3.12). On établit au paragraphe B.1.2 une estimation de l'erreur entre les probabilités de transition $p(\xi, \xi', u)$ données par (2.27) et celles approchées :

$$\begin{aligned} \widetilde{p}(\xi, \xi', u) &= \frac{\widetilde{f}_i^{\pm}(\xi, u)}{\left\|\widetilde{f}(\xi, u)\right\|_1} \text{ pour } \xi' = \xi \pm \delta e_i \\ &= 0 \text{ sinon} \end{aligned}$$

Enfin, le paragraphe B.2 évalue la différence entre la valeur itérée :

$$V^{\delta}_{n+1}(\xi) = \sup_{u \in U^{\delta}} \left\{ \gamma^{\frac{\delta}{\|\widetilde{f}(\xi,u)\|_1}} . \sum_{\xi'} \widetilde{p}(\xi, \xi', u).V^{\delta}_n(\xi') + \frac{\delta}{\left\|\widetilde{f}(\xi, u)\right\|_1}.\widetilde{r}(\xi, u) \right\} \tag{B.1}$$

et la valeur du schéma (2.28), ce qui permet de montrer (3.2) et de prouver la convergence de l'algorithme (théorème 8).

Notations : m_f un minorant de $||f||_1$, M_f un majorant de $||f||_1$, L_r la constant de Lipschitz de r, M_{V^δ} un majorant des V^δ (propriété (2.43)).

B.1 Quelques majorations

B.1.1 Comparaison des temps $\frac{\delta}{||f(\xi,u)||_1}$ et $\frac{\delta}{\left\|\widetilde{f}(\xi,u)\right\|_1}$

D'après le caractère Lipschitzien de f, on peut écrire une majoration provenant du développement de Taylor de f en x :

$$||y - x - f(x,u).\tau||_1 \leq \frac{1}{2} L_f.\tau^2.$$

Puisque f est lipschitzienne et que le voisinage de ξ est de diamètre inférieur à $k_N.\delta$, on a :

$$||f(x,u) - f(\xi,u)||_1 \leq L_f.||x - \xi||_1 \leq L_f k_N \delta.$$

Or $||y-x-f(\xi,u).\tau||_1 = ||y-x-f(x,u).\tau+\tau[f(x,u)-f(\xi,u)]||_1$, donc d'après (3.11), on a :

$$||y - x - f(\xi,u).\tau||_1 \leq \left(\frac{k_2}{2} + k_N\right) L_f k_2 \delta^2.$$

et puisque la dynamique d'état f est bornée inférieurement par m_f et que $\tau \geq k_1\delta$, on a $||y-x||_1 \geq k_1 m_f \delta$, d'où :

$$\left| \frac{\delta}{\|f(\xi,u)\|_1} - \frac{\delta}{\left\|\widetilde{f}(\xi,u)\right\|_1} \right| \leq k_\tau \delta^2 \tag{B.2}$$

avec : $k_\tau = \frac{(\frac{k_2}{2}+k_N)L_f k_2}{k_1 m_f^2}$. On en déduit, en utilisant une propriété de la fonction exponentielle que :

$$\left| \gamma^{\frac{\delta}{\|f(\xi,u)\|_1}} - \gamma^{\frac{\delta}{\|\widetilde{f}(\xi,u)\|_1}} \right| \leq k_\tau \ln\frac{1}{\gamma}.\delta^2 \tag{B.3}$$

B.1.2 Comparaison des probabilités $p(\xi,\xi',v)$ et $\widetilde{p}(\xi,\xi',v)$

Pour $\xi' \neq \xi \pm \delta e_i$, $\widetilde{p}(\xi,\xi',v) = p(\xi,\xi',v) = 0$ et pour $\xi' = \xi \pm \delta e_i$, nous avons :

$$\widetilde{p}(\xi,\xi',v) - p(\xi,\xi',v) = \frac{\widetilde{f}_i^{\pm}(\xi,u)}{\left\|\widetilde{f}(\xi,u)\right\|_1} - \frac{f_i^{\pm}(\xi,u)}{||f(\xi,u)||_1}.$$

Or :

$$\left| \frac{\widetilde{f}_i^{\pm}(\xi,u)}{\left\|\widetilde{f}(\xi,u)\right\|_1} - \frac{f_i^{\pm}(\xi,u)}{||f(\xi,u)||_1} \right| \leq \frac{\left| \left\|\widetilde{f}(\xi,u)\right\|_1 - ||f(\xi,u)||_1 \right| + \left| f_i^{\pm}(\xi,u) - \widetilde{f}_i^{\pm}(\xi,u) \right|}{||f(\xi,u)||_1}$$

D'après le paragraphe précédent,

$$\left\|\widetilde{f}(\xi,u) - f(\xi,u)\right\|_1 \leq \left(\frac{k_2}{2} + k_N\right) L_f \delta$$

d'où nous déduisons :

$$|\widetilde{p}(\xi,\xi',v) - p(\xi,\xi',v)| \leq \frac{L_f}{m_f}(k_2 + 2k_N)\delta \tag{B.4}$$

B.2 Convergence de l'algorithme

La Q-valeur $Q_n^{\delta}(\xi,u)$ est itérée avec la valeur :

$$Q_{n+1}^{\delta}(\xi,u) = \gamma^{\frac{\delta}{\left\|\widetilde{f}(\xi,u)\right\|_1}} . \sum_{\xi'} \widetilde{p}(\xi,\xi',u).V_n^{\delta}(\xi') + \frac{\delta}{\left\|\widetilde{f}(\xi,u)\right\|_1}.\widetilde{r}(\xi,u)$$

selon la règle (3.12) dont la différence, notée Λ, au schéma d'approximation :

$$Q^{\delta}(\xi,u) = \gamma^{\frac{\delta}{||f(\xi,u)||_1}} \sum_{\xi'} p(\xi,\xi',u).V^{\delta}(\xi') + \frac{\delta}{||f(\xi,v)||_1} r(\xi,u)$$

vaut :

$$\begin{aligned}
\Lambda &= \sum_{\xi'} \left[\gamma^{\frac{\delta}{||f(\xi,u)||_1}} p(\xi,\xi',u).V^{\delta}(\xi') - \gamma^{\frac{\delta}{\left\|\widetilde{f}(\xi,u)\right\|_1}} \widetilde{p}(\xi,\xi',u).V_n^{\delta}(\xi') \right] \\
&\quad + \frac{\delta}{||f(\xi,v)||_1} r(\xi,u) - \frac{\delta}{\left\|\widetilde{f}(\xi,u)\right\|_1}.\widetilde{r}(\xi,u) \\
&= \gamma^{\frac{\delta}{||f(\xi,u)||_1}} \sum_{\xi'} \left[p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u) \right] V^{\delta}(\xi') \\
&\quad + \left[\gamma^{\frac{\delta}{||f(\xi,u)||_1}} - \gamma^{\frac{\delta}{\left\|\widetilde{f}(\xi,u)\right\|_1}} \right] \sum_{\xi'} \widetilde{p}(\xi,\xi',u).V^{\delta}(\xi') \\
&\quad + \gamma^{\frac{\delta}{\left\|\widetilde{f}(\xi,u)\right\|_1}} \sum_{\xi'} \widetilde{p}(\xi,\xi',u). \left[V^{\delta}(\xi') - V_n^{\delta}(\xi') \right] \\
&\quad + \frac{\delta}{\left\|\widetilde{f}(\xi,u)\right\|_1} \left[r(\xi,u) - \widetilde{r}(\xi,u) \right] + \left[\frac{\delta}{||f(\xi,v)||_1} - \frac{\delta}{\left\|\widetilde{f}(\xi,u)\right\|_1} \right] r(\xi,u)
\end{aligned}$$

Or d'après (B.3), (B.2) et du caractère lipschtzien de r, on déduit la majoration suivante :

$$\begin{aligned} |\Lambda| \;\leq\; & \gamma^{\frac{\delta}{||f(\xi,u)||_1}} . \left| \sum_{\xi'} \left[p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u) \right] V^\delta(\xi') \right| \\ & + \gamma^{\frac{\delta}{||f(\xi,u)||_1}} \sum_{\xi'} \widetilde{p}(\xi,\xi',u) . \left| V^\delta(\xi') - V_n^\delta(\xi') \right| \\ & + k_\tau \ln \frac{1}{\gamma} . M_{V^\delta} . \delta^2 + \frac{k_2 L_r}{k_1 m_f} \delta^2 + k_\tau M_r \delta^2 . \end{aligned} \tag{B.5}$$

B.2.1 Majoration de $\sum_{\xi'} \left[p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u) \right] V^\delta(\xi')$

Nous avons : $V^\delta(\xi') = V^\delta(\xi) + \left[V^\delta(\xi') - V^\delta(\xi) \right]$. Or d'après les propriétés des probabilités $p(\xi,\xi',u)$ et $\widetilde{p}(\xi,\xi',u)$, nous déduisons :

$$\sum_{\xi'} \left[p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u) \right] V^\delta(\xi') = \sum_{\xi'} \left[p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u) \right] \left[V^\delta(\xi') - V^\delta(\xi) \right] \tag{B.6}$$

De plus,

$$|V^\delta(\xi') - V^\delta(\xi)| \leq |V^\delta(\xi') - V(\xi')| + |V(\xi') - V(\xi)| + |V(\xi) - V^\delta(\xi)|$$

Or d'après le théorème 5, l'erreur d'approximation $\sup_\Omega \left| V^\delta - V \right|$ du schéma tend vers 0 lorsque $\delta \downarrow 0$ pour tout compact $\Omega \subset O$ et grâce à la continuité de V (théorème 2), $\sup_{\substack{z\in\Omega \\ \|h\|\leq\delta}} |V(z) - V(z+h)|$ tend vers 0 lorsque $\delta \downarrow 0$. On en déduit la majoration :

$$\left| V^\delta(\xi') - V^\delta(\xi) \right| \leq \varepsilon(\delta)$$

avec $\varepsilon(\delta) = 2 \sup_{z\in\Omega} \left| V^\delta(z) - V(z) \right| + \sup_{\substack{z\in\Omega \\ \|h\|\leq\delta}} |V(z) - V(z+h)|$, fonction qui tend vers 0 lorsque $\delta \downarrow 0$.

De (B.6) et de (B.4), nous obtenons :

$$\left| \sum_{\xi'} \left[p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u) \right] V^\delta(\xi') \right| \leq \frac{L_f}{m_f} (k_2 + 2k_N) \delta . \varepsilon(\delta) \tag{B.7}$$

Nous pouvons maintenant établir la preuve de convergence de l'algorithme.

B.2.2 Convergence de l'algorithme :

Condition (3.2) :

Soit $E_n^\delta = \sup_{\xi \in (\Sigma^\delta \cup \partial\Sigma^\delta) \cap O} \left| V_n^\delta(\xi) - V^\delta(\xi) \right|$.

De la propriété de la fonction exponentielle $\gamma^{\Delta t} \leq 1 - \frac{\Delta t}{2} \ln \frac{1}{\gamma}$ pour Δt suffisamment petit, on déduit : $\gamma^{\frac{\delta}{\|\tilde{f}(\xi,u)\|_1}} \leq 1 - \frac{\delta}{2.\left\|\tilde{f}(\xi,u)\right\|_1} \ln \frac{1}{\gamma}$, d'où :

$$\gamma^{\frac{\delta}{\|\tilde{f}(\xi,u)\|_1}} \leq 1 - \frac{\delta}{2M_f} \ln \frac{1}{\gamma}$$

pour δ assez petit, et nous déduisons de (B.5) et (B.7) :

$$|\Lambda| \leq (1 - k.\delta) E_n^\delta + e(\delta).\delta \tag{B.8}$$

avec :

$$\begin{aligned} k &= \frac{1}{2M_f} \ln \frac{1}{\gamma} \\ e(\delta) &= \frac{L_f}{m_f}(k_2 + 2k_N)\varepsilon(\delta) + k_\tau \ln \frac{1}{\gamma}.M_{V^\delta}.\delta + \frac{k_2 L_r}{k_1 m_f}\delta + k_\tau M_r \delta \end{aligned}$$

Puisque $\varepsilon(\delta) \downarrow 0$ lorsque $\delta \downarrow 0$, $e(\delta)$ aussi tend vers 0, (B.8) donne :

$$\left| V_{n+1}^\delta(\xi) - V^\delta(\xi) \right| \leq (1 - k.\delta) \sup_{\xi \in \Sigma^\delta \cup \partial\Sigma^\delta} \left| V_n^\delta(\xi) - V^\delta(\xi) \right| + e(\delta).\delta$$

et la condition (3.2) est satisfaite.

Condition (3.3) : lorsque la trajectoire touche la frontière en $y \in \partial O$, avec y dans le voisinage de ξ, la valeur en ξ est itérée selon :

$$V_{n+1}^\delta(\xi) \leftarrow R(y).$$

donc :

$$\left| V_{n+1}^\delta(\xi) - V^\delta(\xi) \right| = |R(y) - R(\xi)| \leq L_R \|y - \xi\|_1 \leq L_R.k_N.\delta$$

et la condition (3.3) est satisfaite avec $k_R = L_R.k_N$.

Ainsi le théorème 7 s'applique et l'algorithme indirect dans le cas déterministe est convergent. ■

Annexe C

Convergence algorithme A/R indirect, cas stochastique

A partir de trajectoires $x_k(t)$ passant aux voisinages des points ξ, on considère des séquences d'états $x_k = x_k(t)$, $y_k = x_k(t+\tau_k)$ vérifiant les propriétés du paragraphe 3.2.2. Un modèle de la dynamique d'état $f(\xi,u)$ et $a(\xi,u)$ est construit à partir de ces données $\{x_k, y_k\}$:

$$\begin{aligned}
\widetilde{f}_n(\xi,u) &= \frac{1}{n}\sum_{k=1}^{n}\frac{y_k - x_k}{\tau_k} \\
\widetilde{a}_n(\xi,u) &= \frac{1}{n}\sum_{k=1}^{n}\frac{\left(y_k - x_k - \tau_k.\widetilde{f}_n(\xi,u)\right)\left(y_k - x_k - \tau_k.\widetilde{f}_n(\xi,u)\right)^T}{\tau_k}.
\end{aligned}$$

Nous donnons une estimation de l'erreur commise $\sup_{n\to\infty}\left|\widetilde{f}_n - f\right|$ et $\sup_{n\to\infty}|\widetilde{a}_n - a|$ au paragraphe C.1.2 (lemmes 6 et 7). Pour établir ce résultat, nous avons besoin d'étendre le développement de Taylor au cas d'équations différentielles stochastiques, ce qui est l'objet du paragraphe C.1.1.

Les probabilités de transition du PDM (2.58) :

$$\begin{aligned}
p(\xi,\xi',u) &= \frac{\tau(\xi,u)}{2\delta^2}\left[2\delta\left|f_i^{\pm}(\xi,u)\right| + a_{ii}(\xi,u) - \sum_{j\neq i}|a_{ij}(\xi,u)|\right] \text{ pour } \xi' = \xi \pm \delta e_i \\
p(\xi,\xi',u) &= \frac{\tau(\xi,u)}{2\delta^2}a_{ij}^{\pm}(\xi,u) \text{ pour } \xi' = \xi + \delta e_i \pm \delta e_j \text{ avec } i \neq j, \qquad \text{(C.1)} \\
p(\xi,\xi',u) &= \frac{\tau(\xi,u)}{2\delta^2}a_{ij}^{\mp}(\xi,u) \text{ pour } \xi' = \xi - \delta e_i \pm \delta e_j \text{ avec } i \neq j, \\
p(\xi,\xi',u) &= 0 \text{ sinon.}
\end{aligned}$$

ainsi que le temps :

$$\tau(\xi,u)=\frac{\delta^2}{\sum_{i=1}^d\left[\delta\left|f_i(\xi,u)\right|+a_{ii}(\xi,u)-\frac{1}{2}\sum_{j\neq i}\left|a_{ij}(\xi,u)\right|\right]} \tag{C.2}$$

sont approchés par les probabilités $\widetilde{p}_n(\xi,\xi',u)$ et le temps $\widetilde{\tau}_n(\xi,u)$ calculés en remplaçant dans (C.1) et (C.2) f et a par le modèle $\widetilde{f_n}$ et $\widetilde{a_n}$.

Une estimation de l'erreur des probabilités approchées $\sup_{n\to\infty}|\widetilde{p_n}-p|$ est donnée au paragraphe C.1.3.

Enfin, le paragraphe C.2 évalue la différence entre la valeur itérée

$$V_{n+1}^{\delta}(\xi)=\sup_{u\in U^{\delta}}\left[\gamma^{\widetilde{\tau}(x,u)}\sum_{\xi'}\widetilde{p}_n(\xi,\xi',u)V_n^{\delta}(\xi')+\widetilde{\tau}_n(x,u)\widetilde{r}_n(\xi,u)\right] \tag{C.3}$$

et la valeur du schéma (2.58), ce qui permet de montrer (3.2) et de prouver la convergence de l'algorithme (théorème 9).

Notations : Par soucis de simplicité, l'indice n sera oté par la suite (sauf lorsqu'il est indiqué explicitement) et le modèle sera noté $\widetilde{f}$, $\widetilde{a}$ et $\widetilde{r}$, les probabilités approchées $\widetilde{p}$ et le temps $\widetilde{\tau}$.

Soient M_f, M_a, M_{f_x} et M_{σ_x} des majorants de f, a, f_x et σ_x, et m_f un minorant de f. Soient $E^{\delta}=\sup_{\xi\in\Sigma^{\delta}}|V^{\delta}(\xi)-V(\xi)|$ et $E_n^{\delta}=\sup_{\xi\in\Sigma^{\delta}}\left|V_n^{\delta}(\xi)-V^{\delta}(\xi)\right|$.

C.1 Quelques majorations

C.1.1 Majoration de Itô-Taylor

On donne ici une généralisation du développement de Taylor au cas stochastique (de telles majorations de Itô-Taylor sont développées à un ordre quelconque dans le livre de Kloeden et Platen [KP95]).

Lemme 5 (Majoration Itô-Taylor) *On considère des trajectoires $x(t)$ satisfaisant l'EDS (2.47) passant en $t=0$ par un point x. Les majorations des espérances (effectuées pour t et x donnés) suivantes ont lieu :*

$$E\left[x(t)\right] = x+f(x,u)t+O(t^2) \tag{C.4}$$

$$E\left[\left(x(t)-x\right)\left(x(t)-x\right)^T\right] = a(x,u)t+f(x,u)f(x,u)^Tt^2+O(t^3) \tag{C.5}$$

Preuve. Notons pour toute matrice symétrique définie positive A,

$$tr(aA)=\sum_{i,j=1}^d a_{i,j}A_{i,j}.$$

Pour toute fonction $\Phi \in \mathcal{C}^{1,2}(\overline{Q})$, notons $D_t\Phi$ sa dérivée temporelle, $D_x\Phi$ son gradient et $D_x^2\Phi$ la matrice de ses dérivées secondes. D'après la formule de différentiation d'Itô (voir [Kry80], [IW89], [FS93], [Kry95], [Oks95] ou [KP95]), pour un contrôle constant $u(t) = u$, on a :

$$\Phi(t,x(t)) = \Phi(0,x) + \int_0^t [D_t\Phi(s,x(s)) + f(x(s),u).D_x\Phi(s,x(s)) \quad \text{(C.6)}$$
$$+\frac{1}{2}tr\left(a(x(s),u).D_x^2\Phi(s,x(s))\right)\Big] ds + \int_0^t \sigma(x(s),u)D_x\Phi(s,x(s)).dw(s)$$

ou de manière équivalente :

$$\Phi(t,x(t)) = \Phi(0,x) + \int_0^t L^0\Phi(s,x(s))ds + \int_0^t L^1\Phi(s,x(s))dw(s)$$

avec les opérateurs

$$\begin{aligned} L^0 &= D_t + f.D_x + \frac{1}{2}tr(aD_x^2) \\ L^1 &= \sigma.D_x \end{aligned}$$

En calculant l'espérance, on obtient la formule de Dynkin :

$$E\left[\Phi(t,x(t))\right] = \Phi(0,x) + E\left[\int_0^t L^0\Phi(s,x(s))ds\right]. \quad \text{(C.7)}$$

Majoration (C.4) : A partir de la forme intégrale du processus de diffusion (équivalent à (C.6) pour $\Phi = x$) :

$$x(t) = x + \int_0^t f(x(t),u)ds + \int_0^t \sigma(x(t),u)dw(s)$$

et en appliquant (C.6) avec $\Phi = f$ et $\Phi = \sigma$, on obtient :

$$\begin{aligned} x(t) &= x + \int_0^t \left[f(x,u) + \int_0^s L^0 f(x(s),u)dz + \int_0^s L^1 f(x(s),u)dw(z)\right] ds \\ &\quad + \int_0^t \left[\sigma(x,u) + \int_0^s L^0\sigma(x(s),u)dz + \int_0^s L^1\sigma(x(s),u)dw(z)\right] dw(s). \end{aligned}$$

En prenant l'espérance, on en déduit :

$$E\left[x(t)\right] = x + f(x,u).t + E\left[R(t)\right] \quad \text{(C.8)}$$

avec le reste

$$R(t) = \int_0^t \int_0^s L^0 f(x(z),u)dzds.$$

Puisque f est $\mathcal{C}^2$, $L^0 f$ est continue sur le compact $\overline{O}$ donc bornée par une valeur M, d'où $R(t) \leq M\frac{t^2}{2} = O(t^2)$ et on déduit de (C.8) que (C.4) a lieu.

Majoration (C.5) :

A partir de la formule d'Itô pour $\Phi_1 = (x(t) - x)(x(t) - x)^T$, on obtient l'espérance :

$$E\left[(x(t) - x)(x(t) - x)^T\right] = E\left[\int_0^t L^0\Phi_1 ds\right]$$

avec $L^0\Phi_1(x(s)) = 2f.(x(s) - x)^T + a(x(s), u)$, et en appliquant à nouveau Itô avec $\Phi_2 = L^0\Phi_1$, on obtient :

$$E\left[(x(t) - x)(x(t) - x)^T\right] = E\left[\int_0^t \left(a(x,u) + \int_0^s L^0\Phi_2 dz\right) ds\right]$$

avec $L^0\Phi_2 = 2f.f^T$. En appliquant Itô avec $\Phi_3 = L^0\Phi_2$, il vient :

$$\begin{aligned} E\left[(x(t) - x)(x(t) - x)^T\right] &= a(x,u)t + E\left[\int_0^t \int_0^s \left(2f(x,u).f^T(x,u)\right.\right. \\ &\quad \left.\left. + \int_0^z L^0\Phi_3 dr\right) dz ds\right] \\ &= a(x,u)t + f(x,u)f^T(x,u)t^2 + E[R(t)] \quad \text{(C.9)} \end{aligned}$$

avec le reste

$$R(t) = \int_0^t \int_0^s \int_0^z L^0\Phi_3 dr dz ds$$

or $\Phi_3 = 2f.f^T$ est $\mathcal{C}^2$ donc $L^0\Phi_3$ est continue sur le compact $\overline{O}$ donc borné par une valeur M d'où $R(t) \leq M\frac{t^3}{3!} = O(t^3)$ et on déduit de (C.9) que (C.5) a lieu. ■

C.1.2 Erreur d'approximation du modèle $\widetilde{f_n}$ et $\widetilde{a_n}$

Les deux lemmes qui suivent donnent des majorations des erreurs d'approximation du modèle $\widetilde{f}$ et $\widetilde{a}$.

Lemme 6 *Le modèle $\widetilde{f_n}$ tend (lorsque $n \to \infty$) à approcher la dynamique d'état f à un $O(\delta)$ près, au sens :*

$$\limsup_{n\to\infty} \left\| \widetilde{f_n}(\xi, u) - f(\xi, u) \right\| = O(\delta) \text{ presque sûrement} \quad \text{(C.10)}$$

Preuve :

Considérons une trajectoire $x_k(t)$ réalisée pour une certaine occurrence $w_k(t)$ du processus de Wiener ; c'est à dire :

$$x_k(t) = x_k + \int_{t_k}^{t} f(x_k(t), u)dt + \int_{t_k}^{t} \sigma(x_k(t), u)dw_k(t).$$

On peut alors construire une trajectoire $z_k(t)$ partant de ξ au temps t_k et déterminé par le même processus de Wiener $w_k(t)$:

$$z_k(t) = \xi + \int_{t_k}^{t} f(z_k(t), u)dt + \int_{t_k}^{t} \sigma(z_k(t), u)dw_k(t).$$

Soit $z_k = z_k(t_k + \tau_k)$. Alors on peut estimer la différence de parcours entre ces deux trajectoires :

$$\begin{aligned}(y_k - x_k) - (z_k - \xi) &= \int_{t_k}^{t_k+\tau_k} [f(x_k(t), u) - f(z_k(t), u)]\, dt \\ &\quad + \int_{t_k}^{t_k+\tau_k} [\sigma(x_k(t), u) - \sigma(z_k(t), u)]\, dw_k(t)\end{aligned}$$

et puisque f et σ ont des gradients majorés respectivement par M_{f_x} et M_{σ_x}, on a :

$$\|(y_k - x_k) - (z_k - \xi)\| \leq (M_{f_x} + M_{\sigma_x}).k_N.\tau_k.\delta. \qquad \text{(C.11)}$$

Le lemme 5, majoration (C.4), appliqué à de telles trajectoire $z_k(t)$ (passant par ξ) donne :

$$E[z_k] = \xi + \tau_k.f(\xi, u) + O(\tau_k^2)$$

qui s'écrit, grâce à (3.14) :

$$E\left[\frac{z_k - \xi}{\tau_k}\right] = f(\xi, u) + O(\delta)$$

et d'après la loi des grands nombres,

$$f(\xi, u) = E\left[\frac{z_k - \xi}{\tau_k}\right] + O(\delta) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \frac{z_k - \xi}{\tau_k} + O(\delta) \text{ presque sûrement.}$$

Donc d'après (C.11), on obtient :

$$\begin{aligned}\limsup_{n\to\infty} \left\| \widetilde{f}_n(\xi, u) - f(\xi, u) \right\| &= \limsup_{n\to\infty} \left\| \frac{1}{n} \sum_{k=1}^{n} \left[\frac{y_k - x_k}{\tau_k} - \frac{z_k - \xi}{\tau_k} \right] \right\| + O(\delta) \\ &= (M_{f_x} + M_{\sigma_x}).k_N.\delta + O(\delta) = O(\delta) \qquad \text{(C.12)}\end{aligned}$$

presque sûrement. ■

Lemme 7 *Le modèle $\widetilde{a_n}$ tend (lorsque $n \to \infty$) à approcher la matrice de covariance a à un $O(\delta^2)$ près, au sens :*

$$\limsup_{n\to\infty} \|\widetilde{a_n}(\xi,u) - a(\xi,u)\| = O(\delta^2) \text{ presque sûrement}$$

Preuve : Le lemme 5, majoration (C.5) donne :

$$E\left[(z_k-\xi)(z_k-\xi)^T\right] = a(\xi,u)\tau_k + f(\xi,u).f(\xi,u)^T.\tau_k^2 + O(\tau_k^3)$$

qui s'écrit, grâce à (3.14) :

$$E\left[\frac{(z_k-\xi-\tau_k f(\xi,u))(z_k-\xi-\tau_k f(\xi,u))^T}{\tau_k}\right] = a(\xi,u) + O(\delta^2).$$

Soient les valeurs :

$$\begin{aligned} r_k &= z_k - \xi - \tau_k f(\xi,u) \\ \text{et } \widetilde{r_k} &= y_k - x_k - \tau_k \widetilde{f_n}(\xi,u) \end{aligned}$$

qui satisfont (grâce à (C.11) et (C.12)) :

$$\|r_k - \widetilde{r_k}\| = (M_{f_x} + M_{\sigma_x}).\tau_k.k_N.\delta + \tau_k.O(\delta) \tag{C.13}$$

et d'après la loi des grands nombres,

$$a(\xi,u) = E\left[\frac{r_k.r_k^T}{\tau_k}\right] + O(\delta^2) \text{ presque sûrement.}$$

Nous en déduisons que :

$$\limsup_{n\to\infty} \|\widetilde{a_n}(\xi,u) - a(\xi,u)\| = \limsup_{n\to\infty} \left\| \frac{1}{n}\sum_{k=1}^{n} \frac{\widetilde{r_k}.\widetilde{r_k}^T}{\tau_k} - \frac{r_k.r_k^T}{\tau_k} \right\| + O(\delta^2) \text{ p.s.}$$

et d'après (C.12) et (C.13), nous obtenons :

$$\begin{aligned} \limsup_{n\to\infty} \|\widetilde{a_n}(\xi,u) - a(\xi,u)\| &= \|\widetilde{r_k} - r_k\| \limsup_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \left(\left\| \frac{\widetilde{r_k}}{\tau_k} \right\| + \left\| \frac{r_k}{\tau_k} \right\| \right) + O(\delta^2) \\ &= O(\delta^2) \text{ presque sûrement.} \end{aligned}$$

Ainsi, il existe des constantes k_f et k_a telles que $\exists \Delta_1, \forall \delta \le \Delta_1, \exists N_1, n \ge N_1$,

$$\begin{aligned} \left\| \widetilde{f_n}(\xi,u) - f(\xi,u) \right\| &\le k_f.\delta \text{ p.s.} \\ \|\widetilde{a_n}(\xi,u) - a(\xi,u)\| &\le k_a.\delta^2 \text{ p.s.} \end{aligned} \tag{C.14}$$

C.1.3 Erreur d'approximation du temps $\widetilde{\tau}_n$ et des probabilités $\widetilde{p_n}$

De (C.2) et (C.14), on déduit :

$$|\tau(\xi,u)-\widetilde{\tau}_n(\xi,u)| \leq \frac{d.(k_f.\delta^2+d.k_a\delta^2)}{(d.m_f.\delta)^2}\delta^2 \leq k_\tau.\delta^2 \tag{C.15}$$

et grâce à une propriété de la fonction exponentielle,

$$\left|\gamma^{\tau(\xi,u)}-\gamma^{\widetilde{\tau}_n(\xi,u)}\right| = k_\tau.\ln\frac{1}{\gamma}.\delta^2. \tag{C.16}$$

On déduit de (C.14) que :

$$\limsup_{n\to\infty}|p(\xi,\xi',u)-\widetilde{p_n}(\xi,\xi',u)| \leq \frac{(2.\delta.M_f+d.M_a)(2.k_f+d.k_a)\delta^2}{\delta m_f-(2.k_f+d.k_a)\delta^2} \leq k_p\delta \text{ p.s.} \tag{C.17}$$

avec $k_p = 4(d.M_a)(2.k_f+d.k_a)$ pour $\delta \leq \Delta_2 = \min\left\{\frac{m_f}{2.k_f+d.k_a}, \frac{d.M_a}{2.\delta.M_f}\right\}$.

C.2 Convergence de l'algorithme

La valeur $V_n^\delta(\xi)$ est itérée avec l'estimation donné par (C.3) dont la différence, notée Λ, au schéma d'approximation (2.58) vaut :

$$\begin{aligned}\Lambda &= \sum_{\xi'}\left[\gamma^{\tau(\xi,u)}p(\xi,\xi',u)V^\delta(\xi')-\gamma^{\widetilde{\tau}(\xi,u)}\widetilde{p}(\xi,\xi',u)V_n^\delta(\xi')\right]\\ &\quad +\tau(\xi,u)r(\xi,u)-\widetilde{\tau}(\xi,u)\widetilde{r}(\xi,u)\end{aligned}$$

$$\begin{aligned}\Lambda &= \gamma^{\tau(\xi,u)}\sum_{\xi'}\left[p(\xi,\xi',u)-\widetilde{p}(\xi,\xi',u)\right]V^\delta(\xi')\\ &\quad +\left(\gamma^{\tau(\xi,u)}-\gamma^{\widetilde{\tau}(\xi,u)}\right)\sum_{\xi'}\widetilde{p}(\xi,\xi',u)V^\delta(\xi')\\ &\quad +\gamma^{\widetilde{\tau}(\xi,u)}\sum_{\xi'}\widetilde{p}(\xi,\xi',u)\left[V^\delta(\xi')-V_n^\delta(\xi')\right]\\ &\quad +\widetilde{\tau}(\xi,u)\left[r(\xi,u)-\widetilde{r}(\xi,u)\right]-\left[\widetilde{\tau}(\xi,u)-\tau(\xi,u)\right]r(\xi,u)\end{aligned}$$

Or d'après (C.15) et (C.16) et du caractère lipschitzien de r (d'où l'on tire $|r(\xi,u)-\widetilde{r}(\xi,u)| \leq k_V.L_r.\delta$) on déduit :

$$|\Lambda| \leq \gamma^{\tau(\xi,u)}\left|\sum_{\xi'}\left[p(\xi,\xi',u)-\widetilde{p}(\xi,\xi',u)\right]V^\delta(\xi')\right|$$

$$+\gamma^{\widetilde{\tau}(\xi,u)} \sum_{\xi'} \widetilde{p}(\xi,\xi',u) \left[V^\delta(\xi') - V_n^\delta(\xi')\right] \qquad \text{(C.18)}$$

$$+k_\tau . \ln \frac{1}{\gamma} . M_{V^\delta} . \delta^2 + k_V . L_r . (k_2 + k_\tau \delta)\delta^2 + k_\tau M_r \delta^2$$

C.2.1 Majoration de $\sum_{\xi'} \left[p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u)\right] V^\delta(\xi')$

Puisque V est de classe $\mathcal{C}^2$, on peut écrire son développement de Taylor :

$$V(\zeta) = V(\xi) + V_x . (\zeta - \xi) + o(\|\zeta - \xi\|).$$

Considérons alors la fonction linéaire $\widetilde{V}$ définie par :

$$\widetilde{V}(x) = V(\xi) + V_x . (x - \xi) .$$

On a alors :

$$\begin{aligned} [p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u)]\, V^\delta(\xi') &= [p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u)] . \left[V^\delta(\xi') - V(\xi')\right] \\ &\quad + [p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u)]\, V(\xi') \end{aligned}$$

donc d'après (C.17) :

$$\begin{aligned} \sum_{\xi'} [p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u)] \quad V^\delta(\xi') & \\ &= k_p . E^\delta . \delta \quad + \sum_{\xi'} [p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u)]\, \widetilde{V}(\xi') + o(\delta) \\ &= k_p . E^\delta . \delta \quad + \left[\widetilde{V}(\eta) - \widetilde{V}(\widetilde{\eta})\right] + o(\delta) \qquad \text{(C.19)} \end{aligned}$$

avec :

$$\begin{aligned} \eta &= \sum_{\xi'} p(\xi,\xi',u)\,(\xi' - \xi) \\ \text{et } \widetilde{\eta} &= \sum_{\xi'} \widetilde{p}(\xi,\xi',u)\,(\xi' - \xi) \end{aligned}$$

D'après la linéarité de $\widetilde{V}$

$$\left|\widetilde{V}(\eta) - \widetilde{V}(\widetilde{\eta})\right| \leq \|\eta - \widetilde{\eta}\| . M_{V_x}$$

et d'après (C.17), puisque $\|\xi' - \xi\| \leq \sqrt{d}\delta$ pour ξ' tels que $p(\xi,\xi',u) \neq 0$, on obtient :

$$\left|\widetilde{V}(\eta) - \widetilde{V}(\widetilde{\eta})\right| \leq \sqrt{d} k_p M_{V_x} \delta^2 .$$

De plus, d'après la convergence du schéma $F^{\delta}_{Stochas}$, le théorème 5 signifie que $E^{\delta}.\delta = o(\delta)$. On déduit de (C.19) que :

$$\sum_{\xi'} \left[p(\xi,\xi',u) - \widetilde{p}(\xi,\xi',u)\right] V^{\delta}(\xi') = o(\delta) \tag{C.20}$$

C.2.2 Convergence de l'algorithme :

Notons $E^{\delta}_{n} = \sup_{\xi\in(\Sigma^{\delta}\cup\partial\Sigma^{\delta})\cap O} \left|V^{\delta}(\xi) - V^{\delta}_{n}(\xi)\right|$.

Condition (3.2) :

On déduit de (C.18) et de (C.20) que :

$$\Lambda = \gamma^{\widetilde{\tau}(\xi,u)} \sum_{\xi'} \widetilde{p}(\xi,\xi',u) \left|V^{\delta}(\xi') - V^{\delta}_{n}(\xi')\right| + o(\delta). \tag{C.21}$$

De la propriété de la fonction exponentielle $\gamma^{\widetilde{\tau}(\xi,u)} \leq 1 - \frac{\widetilde{\tau}(\xi,u)}{2} \ln\frac{1}{\gamma}$ pour $\widetilde{\tau}(\xi,u)$ assez petit, on a :

$$\gamma^{\widetilde{\tau}(\xi,u)} \leq 1 - \frac{(\tau(\xi,u) - k_{\tau}\delta^2)}{2} \ln\frac{1}{\gamma} \leq 1 - \frac{(k_1\delta - k_{\tau}\delta^2)}{2} \ln\frac{1}{\gamma}$$

et nous déduisons de (C.21) que :

$$\Lambda = (1 - k.\delta)E^{\delta}_{n} + o(\delta)$$

avec $k = \frac{k_1}{2} \ln\frac{1}{\gamma}$ et la condition (3.2) a lieu.

Condition (3.3) :

Lorsque la trajectoire touche la frontière en $y \in \partial O$, avec y dans le voisinage de ξ, la valeur en ξ est itérée avec la valeur :

$$V^{\delta}_{n+1}(\xi) \leftarrow R(y).$$

donc :

$$\left|V^{\delta}_{n+1}(\xi) - V^{\delta}(\xi)\right| = |R(y) - R(\xi)| \leq L_R \left\|y - \xi\right\|_1 \leq L_R.k_N.\delta$$

et la condition (3.3) est satisfaite avec $k_R = L_R.k_N$.

Ainsi le théorème 7 s'applique et l'algorithme indirect dans le cas stochastique est convergent. ■

Annexe D

Convergence algorithme A/R direct, aux éléments finis

Le *paragraphe D.1* introduit quelques notations utiles.

Le *paragraphe D.2* permet de montrer que les points y', y'_1 et η (définis plus bas) appartiennent au côté de sortie du simplexe, il fournit une comparaison entre le temps $\tau(\xi, u)$ utilisé par le schéma F^{δ}_{EF} dans l'équation de PD (2.34) et et le temps approché $\frac{\tau}{\lambda_\xi(x)}$ utilisé par l'algorithme. Enfin il donne un important résultat de majoration entre les valeurs de V^{δ} au point $\eta(\xi, u)$ défini par le schéma F^{δ}_{EF} et au point $\xi + \frac{y-x}{\lambda_\xi(x)}$ utilisé par l'algorithme.

Le *paragraphe D.3* évalue la différence entre l'estimation $Q^{\delta}_{n+1}(\xi, u)$ donnée par (3.25) et la valeur du schéma $Q^{\delta}(\xi, u)$ décrit au paragraphe 3.3.1 démontrant ainsi (3.2) et permettant de prouver la convergence de l'algorithme (théorème 10).

D.1 Quelques définitions

Soit T le simplexe (de dimension d) courant. La trajectoire $x(t)$ entre dans T à l'instant t_1 en $x = x(t_1) \in T_{entrée}$ le simplexe (de dimension $d-1$) d'entrée (par exemple, le triangle (ξ, ξ_1, ξ_2) sur la figure D.1) et en ressort à l'instant t_2 en $y = x(t_2) \in T_{sortie}$ le simplexe (de dimension $d-1$) de sortie (par exemple, le triangle (ξ_1, ξ_2, ξ_3)).

Soit $\tau = t_2 - t_1$ le temps de parcours de la trajectoire à l'intérieur du simplexe. Pour alléger les notations, notons τ_ξ le temps $\tau(\xi, u)$. Nous avons :

$$\eta = \xi + \tau_\xi f(\xi, u) \tag{D.1}$$

Soit P l'hyperplan de dimension $(d-1)$ parallèle à T_{sortie} passant par y (dans le cas où $y \in O$, cas décrit par la figure (D.2), P est l'hyperplan

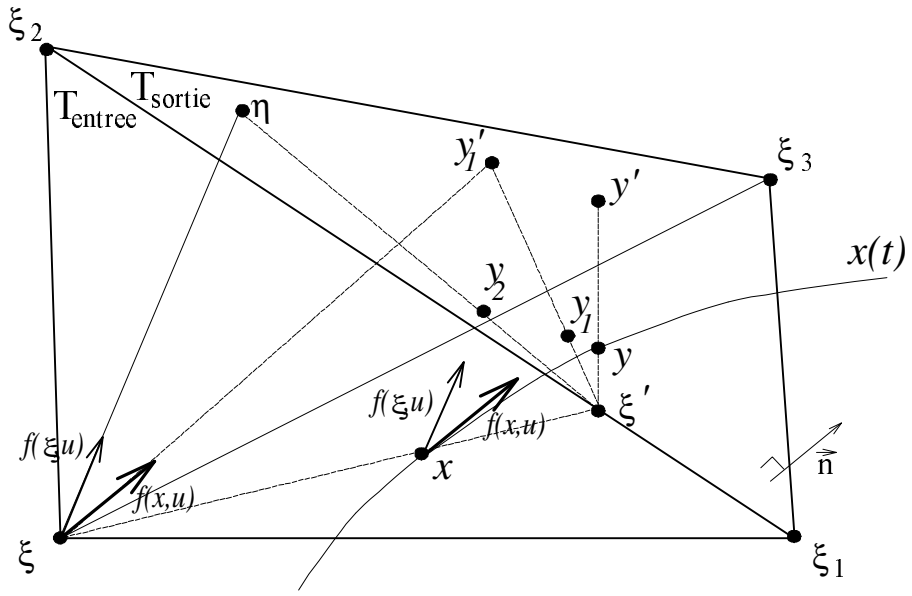

Figure D.1: Un simplexe de dimension 3 $(\xi, \xi_1, \xi_2, \xi_3)$

généré par T_{sortie}). Nous définissons les projections sur P suivantes (voir figure (D.2)) :

y_2	la projection de	x	parallèlement à	$f(\xi, u)$
η	la projection de	ξ	parallèlement à	$f(\xi, u)$
y_1	la projection de	x	parallèlement à	$f(x, u)$
y'_1	la projection de	ξ	parallèlement à	$f(x, u)$
ξ'	la projection de	ξ	parallèlement à	$x - \xi$
y'	la projection de	ξ	parallèlement à	$y - x$

donc, d'après le théorème de Thales, $y'_1 = \xi' + \frac{1}{\lambda_\xi(x)}(y_1 - \xi')$ et $y' = \xi' + \frac{1}{\lambda_\xi(x)}(y - \xi')$. On en déduit que $y' = \xi + \frac{1}{\lambda_\xi(x)}(y - x)$ est l'approximation de η (3.22) utilisée dans l'algorithme Nous définissons les temps suivants :

τ_1	est tel que	$y_1 = x + \tau_1 f(x, u)$
τ_2	est tel que	$y_2 = x + \tau_2 f(\xi, u)$
τ'_1	est tel que	$y'_1 = \xi + \tau'_1 f(x, u)$

Soit $y_3 = x + \tau . f(x, u)$.

Enfin, nous notons M_{V^δ} un majorant des V^δ : M_{V^δ} (propriété (2.43)), M_r un majorant de r : M_r , M_f un majorant de f : M_f, les constantes de Lipschitz de f, r,et R : L_f, L_r, L_R et enfin m_f un minorant de f (propriété (3.20)).

D.2 Quelques propriétés

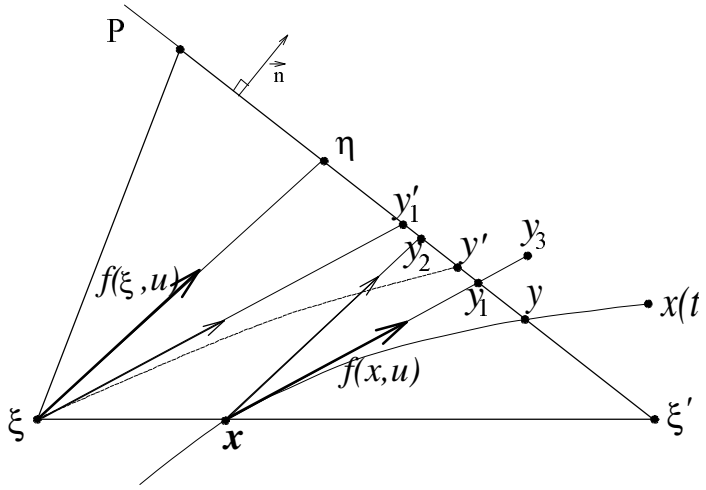

Figure D.2: Notations utiles dans un simplexe de dimension 2. La trajectoire entre en x et sort en $y \in O$.

D.2.1 Les points y', y_1' et η appartiennent à T_{sortie}

Dans ce paragraphe, nous étudions le cas où $y \in O$ (décrit par la figure (D.2)) et nous montrons que si les hypothèses (3.23) et (3.24) sont satisfaites, alors pour des valeurs faibles de δ, on a y', y_1' et $\eta \in T_{sortie}$. A partir de l'hypothèse (3.23) et de la définition de y', on déduit que y' appartient strictement à T_{sortie}.

Des considérations géométriques donnent :

$$||y - y_1|| \leq \frac{||f(x,u)||}{f(x,u)\vec{n}}||y - y_3||. \tag{D.2}$$

De plus, la majoration de Taylor donne :

$$||y - y_3|| \leq \frac{1}{2}L_f\tau^2 \tag{D.3}$$

Donc pour des valeurs faibles de δ, $y_1 \in T_{sortie}$, d'où $f(x,u)$ appartient au cône de sommet ξ engendré par la base T_{sortie}. Aussi, il existe un sommet $\xi_i \neq \xi$ de T tel que :

$$\frac{f(x,u)\vec{n}}{||f(x,u)||} \geq \frac{(\xi_i - \xi)\vec{n}}{||\xi_i - \xi||}.$$

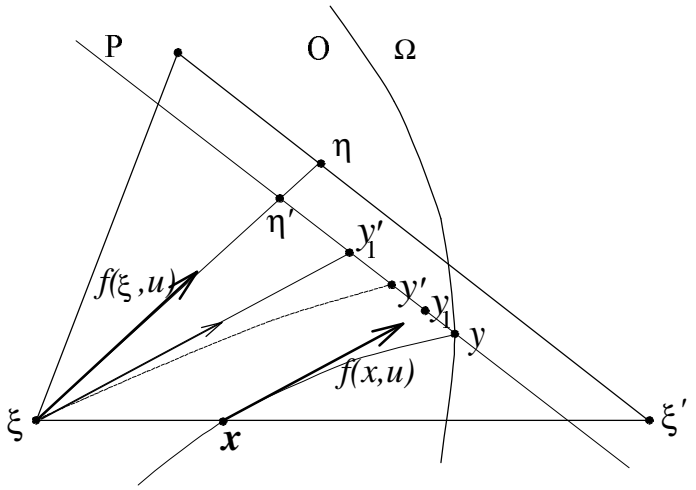

Figure D.3: La trajectoire touche le bord de l'espace d'état : $y \in \partial O$.

Soit $\bar{\xi}$ la projection orthogonale de ξ sur P. Nous déduisons, en utilisant (3.19) que :

$$\frac{f(x,u)\vec{n}}{||f(x,u)||} \geq \frac{||\bar{\xi}-\xi||}{||\xi_i-\xi||} \geq \frac{2k_\rho\delta}{\delta} \geq 2k_\rho.$$

qui donne à partir de (D.2) et (D.3) :

$$||y-y_1|| \leq \frac{1}{2k_\rho}||y-y_3|| \leq \frac{L_f\tau^2}{4k_\rho}. \tag{D.4}$$

Dans le triangle (ξ', y', y'_1), le segment (y, y_1) est colinéaire à (y', y'_1) donc d'après le théorème de Thales, $\frac{||y-y_1||}{||y'-y'_1||} = \frac{||y-\xi'||}{||y'-\xi'||} = \frac{||x-\xi'||}{||\xi-\xi'||} = \lambda_\xi(x)$, d'où :

$$||y'-y'_1|| = \frac{1}{\lambda_\xi(x)}||y-y_1||,$$

donc d'après l'hypothèse (3.24) et (D.4) et sachant que $\tau \leq \frac{\delta}{m_f}$, nous obtenons :

$$||y'-y'_1|| \leq \frac{L_f\delta^2}{4\lambda m_f^2 k_\rho}$$

et pour des valeurs faibles de δ, y'_1 est aussi proche de y' que souhaité ; ainsi pour des valeurs suffisamment faibles de δ, y'_1 appartient strictement à T_{sortie}.

De plus en utilisant la majoration (D.9) obtenue plus bas, on déduit, en utilisant (3.24), que :

$$||\eta - y_1'|| \leq \frac{(1-\lambda)}{\lambda}\frac{L_f}{m_f}\left(1+\frac{1}{2k_\rho}\right)\delta^2.$$

Donc, η est aussi proche de y_1' que souhaité : il existe Δ_0 tel que pour tout $\delta \leq \Delta_0$, on a y', y_1' et $\eta \in T_{out}$.

D.2.2 Comparaison de τ_ξ and τ

Ici nous établissons une comparaison (majoration (D.7) ci-dessous) entre le temps τ_ξ du schéma défini par (D.1) et l'approximation (3.21) par $\frac{\tau}{\lambda_\xi(x)}$ faite dans l'algorithme .

Le théorème de Thales donne $\frac{||\xi'-x||}{||\xi'-\xi||} = \frac{||y_2-x||}{||\eta-\xi||}$ donc :

$$\lambda_\xi(x) = \frac{\tau_2}{\tau_\xi}. \tag{D.5}$$

De plus des considérations géométriques donnent :

$$\frac{\tau_2}{\tau_1} = \frac{\|y_2-x\|}{\|f(\xi,u)\|}\frac{\|f(x,u)\|}{\|y_1-x\|} = \frac{f(x,u)\vec{n}}{f(\xi,u)\vec{n}}. \tag{D.6}$$

A partir de la majoration (D.4) nous obtenons :

$$|\tau_1-\tau|.||f(x,u)|| = ||y_1-y_3|| \leq ||y_1-y|| + ||y-y_3|| \leq \frac{1}{2}L_f\tau^2\left(1+\frac{1}{2k_\rho}\right).$$

En remplaçant τ_1 par l'expression équivalente obtenue par (D.5) et (D.6), on déduit de ce qui précède que :

$$\begin{aligned}
\left|\lambda_\xi(x)\frac{f(x,u)\vec{n}}{f(\xi,u)\vec{n}}\tau_\xi - \tau\right| ||f(x,u)|| &\leq \frac{1}{2}L_f\tau^2\left(1+\frac{1}{2k_\rho}\right) \\
|\lambda_\xi(x)\tau_\xi - \tau|.||f(x,u)|| &\leq \frac{1}{2}L_f\tau^2\left(1+\frac{1}{2k_\rho}\right) \\
&\quad +\lambda_\xi(x)\tau_\xi||f(\xi,u)-f(x,u)||\frac{||f(x,u)||}{f(x,u)\vec{n}}
\end{aligned}$$

et d'après la propriété de Lipschitz de f, nous obtenons :

$$\begin{aligned}
\left|\tau_\xi - \frac{\tau}{\lambda_\xi(x)}\right| &\leq k_\tau\delta^2 \qquad (D.7)\\
\text{avec } k_\tau &= \frac{L_f}{m_f^2}\left[\frac{1}{2k_\rho}+\frac{1}{2m_f}\left(1+\frac{1}{2k_\rho}\right)\right]
\end{aligned}$$

On déduit de cette majoration et de la propriété de la fonction exponentielle : $\gamma^x \geq 1 - x \ln \frac{1}{\gamma}$, que :

$$\left| \gamma^{\tau_\xi} - \gamma^{\frac{\tau_x}{\lambda_\xi(x)}} \right| \leq k_\tau \ln \frac{1}{\gamma} . \delta^2$$

D.2.3 Majoration de $V^\delta(\eta) - V^\delta(y')$

Partant de l'idée sous-jacente à l'algorithme d'approcher $\eta(\xi, u)$ par $y' = \xi + \frac{y-x}{\lambda_\xi(x)}$ (approximation (3.22)), nous établissons ici une majoration ((D.12) ci-dessous) de la différence des valeurs du schéma V^δ en η et en y'.

La fonction linéaire V^δ définie sur le triangle T_{sortie} de dimension $d-1$ a un gradient

$$\sup_{\substack{\zeta \in T_{sortie}, \\ \xi_i \in Sommet(T_{sortie})}} \frac{|V^\delta(\xi_i) - V^\delta(\zeta)|}{||\xi_i - \zeta||}$$

qui est majoré, en définissant

$$\begin{aligned} \xi_{\text{sup}} &= \arg \sup_{\xi_i \in Sommet(T_{sortie})} V^\delta(\xi_i) \\ \text{et } \xi_{\text{inf}} &= \arg \inf_{\xi_i \in Sommet(T_{sortie})} V^\delta(\xi_i) \end{aligned}$$

- soit par $\frac{V^\delta(\xi_{\text{sup}}) - V^\delta(\xi_{\text{inf}})}{||\xi_{\text{sup}} - \zeta||}$ avec un point ζ sur le côté de T_{sortie} opposé à ξ_{sup},
- soit par $\frac{V^\delta(\xi_{\text{sup}}) - V^\delta(\xi_{\text{inf}})}{||\xi_{\text{inf}} - \zeta||}$avec un point ζ sur le côté de T_{sortie} opposé à ξ_{inf}.

Dans les deux cas, d'après (3.19), le dénominateur de ces majorations est minoré par $2k_\rho\delta$. Nous en déduisons que :

$$\frac{\left|V^\delta(y\prime) - V^\delta(\eta)\right|}{||y\prime - \eta||} \leq \frac{V^\delta(\xi_{\text{sup}}) - V^\delta(\xi_{\text{inf}})}{2k_\rho\delta}.$$

De plus,

$$\left|V^\delta(\xi_{\text{sup}}) - V^\delta(\xi_{\text{inf}})\right| \leq \left|V^\delta(\xi_{\text{sup}}) - V(\xi_{\text{sup}})\right| + \left|V(\xi_{\text{sup}}) - V(\xi_{\text{inf}})\right| + \left|V(\xi_{\text{inf}}) - V^\delta(\xi_{\text{inf}})\right|.$$

Or d'après le théorème 5, l'erreur d'approximation $\sup_\Omega \left|V^\delta - V\right|$ du schéma tend vers 0 lorsque $\delta \downarrow 0$ pour tout compact $\Omega \subset O$ et grâce à la

continuité de V (théorème 2), $\sup_{\substack{z\in\Omega \\ \|h\|\leq\delta}} |V(z)-V(z+h)|$ tend vers 0 lorsque $\delta \downarrow 0$. On en déduit la majoration :

$$\left|V^\delta(\xi_{\sup})-V^\delta(\xi_{\inf})\right| \leq \varepsilon(\delta)$$

avec $\varepsilon(\delta) = 2\sup_{z\in\Omega}\left|V^\delta(z)-V(z)\right| + \sup_{\substack{z\in\Omega \\ \|h\|\leq\delta}} |V(z)-V(z+h)|$, fonction qui tend vers 0 lorsque $\delta \downarrow 0$. Nous déduisons :

$$\left|V^\delta(y')-V^\delta(\eta)\right| \leq ||y'-\eta||\frac{\varepsilon(\delta)}{2k_\rho\delta} \tag{D.8}$$

Etablissons maintenant une **estimation de** $||y'-\eta||$**.**

D'après la propriété (utilisant (D.6)) que $\frac{\tau\xi}{\tau_1'} = \frac{\tau_2}{\tau_1} = \frac{f(x,u)\vec{n}}{f(\xi,u)\vec{n}}$, nous établissons :

$$||\eta-y_1'|| = ||\tau_\xi f(\xi,u)-\tau_1' f(x,u)|| = \tau_1'\left|\left|\frac{f(x,u).\vec{n}}{f(\xi,u).\vec{n}}f(\xi,u)-f(x,u)\right|\right|$$

et puisque f est lipschitzienne et $\tau_1' \leq \frac{\delta}{m_f}$, on obtient :

$$\begin{aligned} ||\eta-y_1'|| &\leq \tau_1'||\xi-x||L_f\left(1+\frac{\|f(\xi,u)\|}{f(\xi,u).\vec{n}}\right) \\ &\leq \tau_1'||\xi-x||L_f\left(1+\frac{1}{2k_\rho}\right) \leq \frac{L_f}{m_f}\left(1+\frac{1}{2k_\rho}\right)\delta^2 \end{aligned} \tag{D.9}$$

Remarque 19 *Ces majorations ont été établies dans le cas $y \in O$. Cependant, dans le cas $y \in \partial O$ (décrit par la figure (D.3)), de l'hypothèse (3.18) on déduit que $||\eta-\eta'|| \leq \frac{k_\Sigma\delta^2}{2k_\rho}$ et la majoration (D.9) devient :*

$$||\eta-y_1'|| \leq ||\eta-\eta'|| + ||\eta'-y_1'|| \leq \frac{k_\Sigma\delta^2}{2k_\rho} + \frac{L_f}{m_f}\left(1+\frac{1}{2k_\rho}\right)\delta^2 \tag{D.10}$$

De plus, nous avons d'après (D.4) :

$$||y_1'-y'|| = \frac{||y_1-y||}{\lambda_\xi(x)} \leq \frac{L_f\tau^2}{4\lambda_\xi(x)k_\rho}$$

or d'après (D.7) $\frac{\tau}{\lambda_\xi(x)} \leq \tau_\xi + k_\tau\delta^2$, d'où :

$$||y_1'-y'|| \leq \frac{L_f}{4k_\rho}\tau(\tau_\xi+k_\tau\delta^2) \leq \frac{L_f(1+k_\tau m_f\delta)}{4m_f^2k_\rho}\delta^2. \tag{D.11}$$

En regrouppant (D.8), (D.10) et (D.11), nous obtenons :

$$|V^\delta(\eta)-V^\delta(y')| \leq \left[\frac{L_f}{m_f}\left(1+\frac{1}{2k_\rho}\right) + \frac{L_f(1+k_\tau m_f\delta)}{4m_f^2k_\rho} + \frac{k_\Sigma}{2k_\rho}\right]\frac{\varepsilon(\delta)\delta}{2k_\rho} \tag{D.12}$$

D.3 Convergence de l'algorithme

Preuve du théorème 10. Soit $E_n^\delta = \sup_{\xi\in(\Sigma^\delta\cup\partial\Sigma^\delta)\cap O, u\in U^\delta} |Q_n^\delta(\xi,u) - Q^\delta(\xi,u)|$.

Condition (3.2) : La Q-valeur $Q_n^\delta(\xi,u)$ est itérée avec la règle (3.25) dont la différence au schéma d'approximation (2.37), notée $\Lambda = Q_{n+1}^\delta(\xi,u) - Q^\delta(\xi,u)$, vaut :

$$\Lambda = \gamma^{\frac{\tau}{\lambda_\xi(x)}} \left[\frac{V_n^\delta(y)-V_n^\delta(x)}{\lambda_\xi(x)} + V_n^\delta(\xi)\right] - \gamma^{\tau_\xi} V^\delta(\eta) + \frac{\tau_x}{\lambda_\xi(x)} r(x,u) - \tau_\xi r(\xi,u)$$

D'après la définition de y' qui vaut $\frac{y-x}{\lambda_\xi(x)} + \xi$ et la linéarité de V_n^δ sur le simplexe T, on a :

$$\begin{aligned}
\Lambda &= \gamma^{\frac{\tau}{\lambda_\xi(x)}} V_n^\delta(y') - \gamma^{\tau_\xi} V^\delta(\eta) + \frac{\tau}{\lambda_\xi(x)} r(x,u) - \tau_\xi r(\xi,u) \\
&= \gamma^{\frac{\tau}{\lambda_\xi(x)}} \left[V_n^\delta(y') - V^\delta(y')\right] + \gamma^{\tau_\xi} \left[V^\delta(y') - V^\delta(\eta)\right] + \tau_\xi \left[r(x,u) - r(\xi,u)\right] \\
&\quad + V^\delta(y') \left(\gamma^{\frac{\tau}{\lambda_\xi(x)}} - \gamma^{\tau_\xi}\right) + \left(\frac{\tau}{\lambda_\xi(x)} - \tau_\xi\right) r(x,u).
\end{aligned}$$

On en déduit la majoration suivante :

$$|\Lambda| \le \gamma^{\frac{\tau}{\lambda_\xi(x)}} E_n^\delta + \left|V^\delta(y') - V^\delta(\eta)\right| + \left(\frac{L_r}{m_f} + M_{V^\delta} k_\tau \ln\frac{1}{\gamma} + M_r k_\tau\right)\delta^2 \quad \text{(D.13)}$$

D'après la propriété de la fonction exponentielle, on a : $\gamma^{\frac{\tau}{\lambda_\xi(x)}} \le 1 - \frac{\tau}{2\lambda_\xi(x)} \ln\frac{1}{\gamma}$ pour $\frac{\tau}{\lambda_\xi(x)}$ assez petit. De plus, (D.7) implique $\frac{\tau}{\lambda_\xi(x)} \ge \tau_\xi - k_\tau\delta^2$, or $\tau_\xi \ge \frac{2k_\rho\delta}{M_f}$, donc :

$$\gamma^{\frac{\tau}{\lambda_\xi(x)}} \le 1 - \left(\frac{k_\rho\delta}{M_f} - \frac{k_\tau\delta^2}{2}\right)\ln\frac{1}{\gamma} \le 1 - \frac{k_\rho\delta}{2M_f}\ln\frac{1}{\gamma}$$

pour $\delta \le \Delta_2 = \frac{k_\rho}{k_\tau M_f}$. En utilisant la majoration (D.12), on déduit de (D.13) que :

$$|\Lambda| \le (1 - k.\delta)E_n^\delta + e(\delta).\delta \quad \text{(D.14)}$$

avec :

$$\begin{aligned}
k &= \frac{k_\rho}{2M_f}\ln\frac{1}{\gamma} \\
e(\delta) &= \left[\frac{L_f}{m_f}\left(1 + \frac{1}{2k_\rho}\right) + \frac{L_f(1+k_\tau m_f\delta)}{4m_f^2 k_\rho} + \frac{k_\Sigma}{2k_\rho}\right]\frac{\varepsilon(\delta)}{2k_\rho} \\
&\quad + \left(\frac{L_r}{m_f} + M_{V^\delta} k_\tau \ln\frac{1}{\gamma} + M_r k_\tau\right)\delta
\end{aligned}$$

Puisque $\varepsilon(\delta) \downarrow 0$ lorsque $\delta \downarrow 0$, $e(\delta)$ aussi tend vers 0, et d'après la définition de $V^{\delta}_{n+1}(\xi)$ $(= \sup_{u \in U^{\delta}} Q^{\delta}_{n+1}(\xi, u))$, (D.14) donne :

$$\left|V^{\delta}_{n+1}(\xi) - V^{\delta}(\xi)\right| \leq (1 - k.\delta) \sup_{\xi \in \Sigma^{\delta} \cup \partial\Sigma^{\delta}} \left|V^{\delta}_{n}(\xi) - V^{\delta}(\xi)\right| + e(\delta).\delta$$

et la condition (3.2) est satisfaite pour $\delta \leq \min\{\Delta_0, \Delta_1, \Delta_2\}$.

Condition (3.3) : lorsque $y \in \partial O$, le sommet $\xi' \in \partial\Sigma^{\delta}$ le plus proche de y est itéré selon :

$$V^{\delta}_{n+1}(\xi') \leftarrow R(y).$$

donc :

$$\left|V^{\delta}_{n+1}(\xi') - V^{\delta}(\xi')\right| = |R(y) - R(\xi')| \leq L_R \|y - \xi'\| \leq k_R.\delta$$

et la condition (3.3) est satisfaite avec $k_R = L_R$.

Ainsi le théorème 7 s'applique et l'algorithme direct d'A/R aux éléments finis est convergent. ■

Annexe E

Convergence algorithme A/R direct, aux différences finies

Le paragraphe E.1 établit une comparaison entre le temps $\frac{\delta}{||f(\xi,u)||_1}$ utilisé par le schéma $F^{\delta}_{DF_\infty}$ dans l'équation de PD (2.28) et le temps approché $\frac{\delta.\tau}{\|y-x\|_1}$ utilisé par l'algorithme, ainsi qu'une estimation de l'erreur entre les probabilités de transition $\frac{f_i(\xi,u)}{||f(\xi,u)||}$ données par (2.27) et celles approchées $\frac{y_i-x_i}{||y-x||_1}$.

Enfin, le paragraphe E.2 évalue la différence entre la valeur itérée $Q^{\delta}_{n+1}(\xi,u)$ donnée par (3.29) et la valeur du schéma $Q^{\delta}(\xi,u)$ décrit au paragraphe 3.3.2 démontrant ainsi (3.2) et permettant de prouver la convergence de l'algorithme (théorème 11).

Notations : m_f minorant de $||f||_1$, L_r la constant de Lipschitz de r, M_{V^δ} un majorant des V^δ (propriété (2.43)), M_r un majorant de r. Notons

$$p_i = \frac{|f_i(\xi,u)|}{||f(\xi,u)||} \text{ et } \widetilde{p_i} = \frac{|y_i - x_i|}{||y-x||_1}$$

les probabilités de transition de la case contenant ξ aux cases voisines.

E.1 Quelques propriétés

E.1.1 Comparaison des temps $\frac{\delta}{||f(\xi,u)||_1}$ et $\frac{\delta.\tau}{\|y-x\|_1}$

Pour une trajectoire $x(t)$ traversant la case X de centre ξ, nous notons x le point d'entrée, y le point de sortie et τ le temps de passage de la trajectoire dans X. D'après le développement de Taylor de f, on déduit :

$$||y - x - f(x,u).\tau||_1 \leq \frac{1}{2} L_f.\tau^2.$$

Puisque la dynamique d'état f est bornée inférieurement par m_f, le temps τ est majoré par $\frac{d.\delta}{m_f}$. A partir du caractère lipschitzien de f,

$$||f(x,u) - f(\xi,u)||_1 \leq L_f.||x-\xi||_1 \leq \frac{d.\delta}{2} L_f.$$

Or $||y-x-f(\xi,u).\tau||_1 = ||y-x-f(x,u).\tau+\tau[f(x,u)-f(\xi,u)]||_1$, donc :

$$||y-x-f(\xi,u).\tau||_1 \leq \frac{d^2\delta^2}{2m_f} L_f.(1+\frac{1}{m_f}).$$

Et puisque $||y-x||_1 \geq \lambda.\delta$ (hypothèse (3.27)), nous obtenons :

$$\left| \frac{\delta}{\|f(\xi,u)\|_1} - \frac{\delta.\tau}{\|y-x\|_1} \right| \leq k_\tau \delta^2 \tag{E.1}$$

avec : $k_\tau = \frac{d^2}{2\lambda.m_f^2} L_f.(1+\frac{1}{m_f})$. On en déduit, en utilisant une propriété de la fonction exponentielle que :

$$\left| \gamma^{\frac{\delta}{\|f(\xi,u)\|_1}} - \gamma^{\frac{\delta}{\|y-x\|_1}.\tau} \right| \leq k_\tau \ln\frac{1}{\gamma}.\delta^2 \tag{E.2}$$

E.1.2 Comparaison des probabilités p_i et $\widetilde{p_i}$:

Nous avons :

$$\left| \frac{\frac{|y_i-x_i|}{\tau}}{\frac{||y-x||_1}{\tau}} - \frac{|f_i(\xi,u)|}{||f(\xi,u)||_1} \right| \leq \frac{\left| \frac{||y-x||_1}{\tau} - ||f(\xi,u)||_1 \right| + \left| |f_i(\xi,u)| - \frac{|y_i-x_i|}{\tau} \right|}{||f(\xi,u)||_1}$$

or d'après le paragraphe précédent,

$$\left\| \frac{y-x}{\tau} - f(\xi,u) \right\|_1 \leq \frac{d}{2} L_f (1+\frac{1}{m_f})\delta$$

d'où nous déduisons :

$$\left| \frac{|y_i-x_i|}{||y-x||_1} - \frac{|f_i(\xi,u)|}{||f(\xi,u)||_1} \right| \leq \frac{dL_f}{m_f}(1+\frac{1}{m_f})\delta \tag{E.3}$$

E.2 Convergence de l'algorithme

La Q-valeur $Q_n^\delta(\xi,u)$ est itérée selon la règle (3.29) dont la différence au schéma d'approximation (2.26), notée $\Lambda = Q^\delta(\xi,u) - Q_{n+1}^\delta(\xi,u)$, vaut :

$$\begin{aligned}\Lambda &= Q^\delta(\xi,u) - \sum_{i=1}^{d} \frac{|y_i - x_i|}{\|y-x\|_1} Q_{n+1}^\delta(\xi,u,i) \\ &= Q^\delta(\xi,u) - \sum_{i=1}^{d} \frac{|y_i - x_i|}{\|y-x\|_1} \left[V_n^\delta(\xi^i) + \frac{\delta.\tau^i}{\|y^i - x^i\|_1} r(z^i,u) \right]\end{aligned}$$

pour des trajectoires $x^i(t)$ traversant la case X vers les cases voisines X^i (de centre $\xi^i = \xi \pm \delta e_i$ selon que $|f_i(\xi,u)|$ est positif ou négatif) avec un contrôle u, dont les points d'entrée, de sortie, le temps de passage dans X et un point quelconque de la trajectoire dans X sont respectivement x^i, y^i, τ^i et z^i.

Dans cette équation, il faut faire attention aux notations : x_i et y_i représentent les $i^{ième}$ coordonnées de x et y, provenant d'une trajectoire $x(t)$ tandis que x^i et y^i sont les points d'entrée et de sortie des trajectoires $x^i(t)$.

D'après le schéma d'approximation (2.29), on obtient :

$$\begin{aligned}\Lambda &= \sum_{i=1}^{d} \left[\gamma^{\frac{\delta}{||f(\xi,u)||_1}} . \frac{|f_i(\xi,u)|}{||f(\xi,u)||_1} V^\delta(\xi^i) - \gamma^{\frac{\delta.\tau^i}{\|y^i-x^i\|_1}} \frac{|y_i - x_i|}{\|y-x\|_1} V_n^\delta(\xi^i) \right] \\ &\quad + \sum_{i=1}^{d} \frac{|y_i - x_i|}{\|y-x\|_1} \left[\frac{\delta}{||f(\xi,u)||_1} r(\xi,u) - \frac{\delta.\tau^i}{\|y^i-x^i\|_1} r(z^i,u) \right] \\ &= \gamma^{\frac{\delta}{||f(\xi,u)||_1}} . \sum_{i=1}^{d} \left(\frac{|f_i(\xi,u)|}{||f(\xi,u)||_1} - \frac{|y_i - x_i|}{\|y-x\|_1} \right) V^\delta(\xi^i) \\ &\quad + \sum_{i=1}^{d} \frac{|y_i - x_i|}{\|y-x\|_1} \left(\gamma^{\frac{\delta}{||f(\xi,u)||_1}} - \gamma^{\frac{\delta.\tau^i}{\|y^i-x^i\|_1}} \right) V^\delta(\xi^i) \\ &\quad + \sum_{i=1}^{d} \gamma^{\frac{\delta.\tau^i}{\|y^i-x^i\|_1}} \frac{|y_i - x_i|}{\|y-x\|_1} \left[V^\delta(\xi^i) - V_n^\delta(\xi^i) \right] \\ &\quad + \sum_{i=1}^{d} \frac{|y_i - x_i|}{\|y-x\|_1} \frac{\delta.\tau^i}{\|y^i-x^i\|_1} \left[r(\xi,u) - r(z^i,u) \right] \\ &\quad + \sum_{i=1}^{d} \frac{|y_i - x_i|}{\|y-x\|_1} \left| \left(\frac{\delta}{||f(\xi,u)||_1} - \frac{\delta.\tau^i}{\|y^i-x^i\|_1} \right) r(\xi,u) \right|\end{aligned}$$

Or d'après (E.2), (E.1) et du caractère lipschtzien de r, on déduit la majoration suivante :

$$\begin{aligned} |\Lambda| \quad \leq \quad & \gamma^{\frac{\delta}{||f(\xi,u)||_1}} . \left| \sum_{i=1}^{d} \left[\frac{|f_i(\xi,u)|}{||f(\xi,u)||_1} - \frac{|y_i - x_i|}{\|y-x\|_1} \right] V^\delta(\xi^i) \right| \\ & + \sum_{i=1}^{d} \gamma^{\frac{\delta.\tau^i}{\|y^i - x^i\|_1}} \frac{|y_i - x_i|}{\|y-x\|_1} \left| V^\delta(\xi^i) - V_n^\delta(\xi^i) \right| \qquad \text{(E.4)} \\ & + k_\tau \ln \frac{1}{\gamma} . M_{V^\delta} . \delta^2 + \frac{L_r}{m_f} \delta^2 + k_\tau L_r \delta^3 + k_\tau M_r \delta^2 . \end{aligned}$$

E.2.1 Majoration de $\sum_{i=1}^{d} [p_i - \widetilde{p}_i] V^\delta(\xi^i)$

Nous avons : $V^\delta(\xi') = V^\delta(\xi) + \left[V^\delta(\xi') - V^\delta(\xi) \right]$. Or d'après les propriétés des probabilités p_i et $\widetilde{p}_i$, nous déduisons :

$$\left| \sum_{i=1}^{d} [p_i - \widetilde{p}_i] V^\delta(\xi') \right| = \left| \sum_{i=1}^{d} [p_i - \widetilde{p}_i] \left[V^\delta(\xi') - V^\delta(\xi) \right] \right| \qquad \text{(E.5)}$$

De plus,

$$|V^\delta(\xi') - V^\delta(\xi)| \leq |V^\delta(\xi') - V(\xi')| + |V(\xi') - V(\xi)| + |V(\xi) - V^\delta(\xi)|$$

Or d'après le théorème 5, l'erreur d'approximation $\sup_\Omega \left| V^\delta - V \right|$ du schéma tend vers 0 lorsque $\delta \downarrow 0$ pour tout compact $\Omega \subset O$ et grâce à la continuité de V (théorème 2), $\sup_{\substack{z \in \Omega \\ \|h\| \leq \delta}} |V(z) - V(z+h)|$ tend vers 0 lorsque $\delta \downarrow 0$. On en déduit la majoration :

$$\left| V^\delta(\xi') - V^\delta(\xi) \right| \leq \varepsilon(\delta)$$

avec $\varepsilon(\delta) = 2 \sup_{z \in \Omega} \left| V^\delta(z) - V(z) \right| + \sup_{\substack{z \in \Omega \\ \|h\| \leq \delta}} |V(z) - V(z+h)|$, fonction qui tend vers 0 lorsque $\delta \downarrow 0$.

De (E.5) et de (E.3), nous obtenons :

$$\left| \sum_{i=1}^{d} \left[\frac{|f_i(\xi,u)|}{||f(\xi,u)||_1} - \frac{|y_i - x_i|}{\|y-x\|_1} \right] V^\delta(\xi_i) \right| \leq \frac{dL_f}{m_f} (1 + \frac{1}{m_f}) \delta \varepsilon(\delta) \qquad \text{(E.6)}$$

Nous pouvons maintenant établir la preuve de convergence de l'algorithme.

E.2.2 Preuve de convergence

Soit $E_n^\delta = \sup_{\xi \in (\Sigma^\delta \cup \partial\Sigma^\delta) \cap O} |V_n^\delta(\xi, u) - V^\delta(\xi, u)|$.

Condition (3.2) :

De la propriété de la fonction exponentielle $\gamma^{\frac{\delta.\tau}{\|y-x\|_1}} \leq 1 - \frac{\delta.\tau}{2\|y-x\|_1} \ln \frac{1}{\gamma}$ pour $\frac{\delta.\tau}{\|y-x\|_1}$ suffisamment petit, et de (E.1), on obtient :

$$\gamma^{\frac{\delta.\tau}{\|y-x\|_1}} \leq 1 - (\frac{\delta}{2M_f} - k_\tau \delta^2) \ln \frac{1}{\gamma} \leq 1 - \frac{\delta}{4M_f} \ln \frac{1}{\gamma}$$

pour $\delta \leq \Delta_1 = \frac{1}{4k_\tau M_f}$ et nous déduisons de (E.4) et (E.6) :

$$|\Lambda| \leq (1 - k.\delta) E_n^\delta + e(\delta).\delta \tag{E.7}$$

avec :

$$\begin{aligned} k &= \frac{1}{4M_f} \ln \frac{1}{\gamma} \\ e(\delta) &= \frac{dL_f}{m_f}(1 + \frac{1}{m_f})\varepsilon(\delta) + k_\tau \ln \frac{1}{\gamma}.M_{V^\delta}.\delta + \frac{L_r}{m_f}\delta + k_\tau M_r \delta + k_\tau L_r \delta^2 \end{aligned}$$

Puisque $\varepsilon(\delta) \downarrow 0$ lorsque $\delta \downarrow 0$, $e(\delta)$ aussi tend vers 0, et (E.7) donne :

$$\left|V_{n+1}^\delta(\xi) - V^\delta(\xi)\right| \leq (1 - k.\delta) \sup_{\xi \in \Sigma^\delta \cup \partial\Sigma^\delta} \left|V_n^\delta(\xi) - V^\delta(\xi)\right| + e(\delta).\delta$$

et la condition (3.2) est satisfaite pour $\delta \leq \min\{\Delta_0, \Delta_1\}$.

Condition (3.3) : lorsque $y^i \in \partial O$, le point ξ est itéré selon :

$$V_{n+1}^\delta(\xi) \leftarrow R(y^i).$$

donc :

$$\left|V_{n+1}^\delta(\xi) - R(\xi)\right| = \left|R(y^i) - R(\xi)\right| \leq L_R \|y - \xi\|_1 \leq k_R.\delta$$

et la condition (3.3) est satisfaite avec $k_R = L_R$.

Ainsi le théorème 7 s'applique et l'algorithme direct d'A/R aux différences finies est convergent. ■

Index

Action, 12
Agent, 12
Algorithme, 59
Algorithme d'apprentissage par renforcement, 60
Algorithme direct
  aux DF, 72
  aux EF, 69
Algorithme indirect
  cas déterministe, 67
  cas stochastique, 68
Algorithmes indirects, 66

Comportement, 12
Continuité fonction valeur, 35
Contrôle, 12, 30
Contrôle optimal, 35
  cas discret, 16
Convergence d'algorithmes d'A/R, 62

Décision, 12
Dérive, 52
Descente de gradient (méthode), 38
Diffusion, 52
Double dynamique, 75
Dynamique d'état, 12, 30
Dynamique d'apprentissage, 75
Dynamique structurelle, 75

Environnement, 12
Equation de Bellman, 16
Equation de la Programmation Dynamique
  cas discret, 16
Equation différentielle stochastique, 52
Espace d'état, 30
  discrétisé, 42
Espace de contrôle, 30
Espace de contrôle approché, 41
Etat, 30
Exploitation, 19
Exploration, 19

Facteur de décroissance, 32
Fonction valeur
  cas déterministe, 32, 33
  cas stochastique, 53
fonction valeur
  cas discret, 15
Fonctionnelle de renforcement
  actualisé, 32
  cas déterministe, 31
  cas stochastique, 53
Frontière
  (de l'espace discrétisé), 42
  latérale, 30
  terminale, 30

Grille, 42

Hamilton-Jacobi-Bellman, 33
  cas stochastique, 54
Hamiltonien, 34
Horizon temporel
  fini, 30
  infini, 32

Intégrale stochastique, 51
Itération sur les stratégies, 17
Itération sur les valeurs, 17

Méthode de PD
    asynchrone, 18
    de Gauss-Seidel, 18
    synchrone, 18
Méthodes aux
    différences finies, 42
Méthodes de PD
    en ligne, 18
    hors-ligne, 18

Parabolicité uniforme, 53
Partie
    négative, 42
    positive, 42
Pas de discrétisation
    spatial, 42
    temporel, 43
Plan de la thèse, 27
Probabilités de transition, 56, 57
probabilités de transition, 43
Probabilités de transition, 45
Processus de Décision Markovien, 14
Processus de Wiener, 52
Programmation Dynamique, 15
Programmation dynamique
    en temps réel, 18
propriété de "contraction faible", 61

Q-learning, 20
Q-valeurs, 17
Qualités, 17
Quotient aux différences finies, 42

Récompense, 13
Renforcement, 13
    courant, 31
    latéral, 31
    terminal, 31

Schéma aux DF
    cas déterministe
        horizon temporel fini, 44
        horizon temporel infini, 45
    cas stochastique
        horizon infini, 57
Schéma aux EF, 48
Schéma d'algorithme, 59
Schémas d'approximation
    convergence, 49
Schémas numériques, 41
Solution de viscosité, 36, 39
    propriétés, 39
Stratégie, 16
Système artificiel, 12

Taux d'actualisation, 32
Temps de sortie, 30
Trajectoire, 30

Unicité forte (principe), 41

Bibliographie

[Aki90] Marianne Akian. *Méthodes multigrilles en contrôle stochastique.* PhD thesis, University Paris IX Dauphine, 1990.

[Bai93] Leemon C. Baird. Advantage updating. *Report WL-TR-93-1146, Wright Patterson Air Force Base*, 1993.

[Bai95] Leemon C. Baird. Residual algorithms : Reinforcement learning with function approximation. *Machine Learning : proceedings of the Twelfth International Conference*, 1995.

[Bar90] Andrew G. Barto. *Neural networks for control.* W.T. Miller, R.S.Sutton, P.J. Werbos editors. MIT press, Cambridge, Massachussetts, 1990.

[Bar94] Guy Barles. *Solutions de viscosité des équations de Hamilton-Jacobi*, volume 17 of *Mathématiques et Applications.* Springer-Verlag, 1994.

[BBS91] Andrew G. Barto, Steven J. Bradtke, and Satinder P. Singh. Real-time learning and control using asynchronous dynamic programming. Technical Report 91-57, Computer Science Department, University of Massachusetts, 1991.

[BBS95] Andrew G. Barto, Steven J. Bradtke, and Satinder P. Singh. Learning to act using real-time dynamic programming. *Artificial Intelligence*, (72): 81–138, 1995.

[Bel57] R. Bellman. *Dynamic Programming.* Princeton Univ. Press, 1957.

[Ber87] Dimitri P. Bertsekas. *Dynamic Programming : Deterministic and Stochastic Models.* Prentice Hall, 1987.

[BG97] Hugues Bersini and Vittorio Gorrini. A simplification of the back-propagation-through-time algorithm for optimal neurocontrol. *IEEE Transaction on Neural Networks*, 8: 437–441, 1997.

[BM95] J.A. Boyan and A.W. Moore. Generalization in reinforcement learning: Safely approximating the value function. *Advances in Neural Information Processing Systems*, 7, 1995.

[BP88] Guy Barles and B. Perthame. Exit time problems in optimal control and vanishing viscosity solutions of hamilton-jacobi equations. *SIAM Control Optimization*, 26: 1133–1148, 1988.

[BP90] Guy Barles and B. Perthame. Comparison principle for dirichlet-type hamilton-jacobi equations and singular perturbations of degenerated elliptic equations. *Applied Mathematics and Optimization*, 21: 21–44, 1990.

[Bre95] P.C. Bressloff. Stochastic dynamics of reinforcement learning. *Network: Computation in Neural Systems*, 6, 1995.

[BS91] Guy Barles and P.E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Analysis*, 4: 271–283, 1991.

[BSA83] Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. *Neuronlike adaptive elements that can solve difficult learning control problems.* IEEE Transactions on Systems, Man and Sybernetics, 1983.

[BT96] Dimitri P. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming.* Athena Scientific, 1996.

[CB96] B. Crites and A. Barto. Improving elevator performance using reinforcement learning. *Advances in Neural Information Processing Systems*, 8, 1996.

[CHMW94] D. Cliff, P. Husbands, J.A. Meyer, and S.W. Wilson. *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior.* The MIT Press/Bradford Books, 1994.

[CIL92] M.G. Crandall, Hitoshi Ishii, and P.L. Lions. User's guide to viscosity solutions of second order partial differential equations. *Bulletin of the American Mathematical Society*, 27(1), 1992.

[CL83] M.G. Crandall and P.L. Lions. Viscosity solutions of hamilton-jacobi equations. *Trans. of the American Mathematical Society*, 277, 1983.

[Dav96] Scott Davies. Multidimensional triangulation and interpolation for reinforcement learning. *Advances in Neural Information Processing Systems*, 8, 1996.

[Day92] Peter Dayan. The convergence of td(λ) for general λ. *Machine Learning*, 8: 341–362, 1992.

[DB94] Marco Dorigo and Hugues Bersini. A comparison of q-learning and classifier systems. *Simulation of Adaptive Behavior*, 1994.

[DS94] Peter Dayan and T.J. Sejnowski. Td(λ) converges with probability 1. *Machine Learning*, 14: 295–301, 1994.

[FS93] Wendell H. Fleming and H. Mete Soner. *Controlled Markov Processes and Viscosity Solutions*. Applications of Mathematics. Springer-Verlag, 1993.

[Gol89] D.E. Goldberg. *Genetic Algorithms in search, optimization and machine learning*. Addisson Wesley, 1989.

[Gor95] G. Gordon. Stable function approximation in dynamic programming. *International Conference on Machine Learning*, 1995.

[Gul90] Vijay Gullapalli. A stochastic reinforcement learning algorithm for learning real-valued functions. *Neural Networks*, 3: 671–692, 1990.

[Gul92] Vijay Gullapalli. *Reinforcement Learning and its application to control*. PhD thesis, University of Massachussetts, Amherst., 1992.

[IW89] N. Ikeda and S. Watanabe. *Stochastic Differential Equations and Diffusion Processes*. North Holland, Amsterdam, 1989.

[JJS94] Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6: 1185–1201, 1994.

[KBW96] Leslie P. Kaelbling, A.G. Barto, and R.J. Williams. Reinforcement learning: a survey. *Journal of AI Research*, 4, 1996.

[KD92] Harold J. Kushner and Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time*. Applications of Mathematics. Springer-Verlag, 1992.

[KP95] Peter E. Kloeden and Eckhard Platen. *Numerical Solutions of Stochastic Differential Equations*. Springer-Verlag, 1995.

[Kry80] N.V. Krylov. *Controlled Diffusion Processes*. Springer-Verlag, New York, 1980.

[Kry95] N.V. Krylov. *Introduction to the Theory of Diffusion Processes*, volume 142. American Mathematical Society, 1995.

[Kus90] Harold J. Kushner. Numerical methods for stochastic control problems in continuous time. *SIAM J. Control and Optimization*, 28: 999–1048, 1990.

[Lin93] Long-Ji Lin. *Reinforcement Learning for Robots using Neural Networks*. PhD thesis, Carnegie Mellon University, Pittsburg, Pennsylvania, 1993.

[MC92] Sridhar Mahadevan and Jonathan Connell. Automatic programming of behavior-based robots using reinforcement learning. *Artificial Intelligence*, 5: 311–365, 1992.

[Meu96] N. Meuleau. *Le dilemme Exploration/Exploitation dans les systèmes d'apprentissage par renforcement*. PhD thesis, Université de Caen, 1996.

[Mid93] Terje Midtbo. *Spatial Modelling by Delaunay Networks of Two and Three Dimensions*. PhD thesis, Norwegian Institute of Technology, 1993.

[Moo91] Andrew W. Moore. Variable resolution dynamic programming: Efficiently learning action maps in multivariate real-valued state-spaces. *Machine Learning : Poceedings of the Eight International Workshop*, pages 333–337, 1991.

[Moo94] Andrew W. Moore. The parti-game algorithm for variable resolution reinforcement learning in multidimensional state space. *Advances in Neural Information Processing Systems*, 6, 1994.

[MP94] R. Munos and J. Patinel. Reinforcement learning with dynamic covering of the state-action space : Partitioning q-learning. *Simulation of Adaptive Behavior*, 1994.

[Mun92] Rémi Munos. Algorithmes génétiques et apprentissage par renforcement. *Rapport de DEA Sciences Cognitives*, 1992.

[Mun96a] Rémi Munos. A convergent reinforcement learning algorithm in the continuous case : the finite-element reinforcement learning. *International Conference on Machine Learning*, 1996.

[Mun96b] Rémi Munos. Using finite differences methods for approximating the value function of continuous reinforcement learning problems. *International Symposium on Multi-Technology Information Processing*, 1996.

[Mun97a] Rémi Munos. Catégorisation adaptative de donnés sensorimotrices pour un système d'apprentissage par renforcement. *Journées de Rochebrune, rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels*, 1997.

[Mun97b] Rémi Munos. A convergent reinforcement learning algorithm in the continuous case based on a finite difference method. *International Joint Conference on Artificial Intelligence*, 1997.

[Mun97c] Rémi Munos. Finite-element methods with local triangulation refinement for continuous reinforcement learning problems. *European Conference on Machine Learning*, 1997.

[Mun97d] Rémi Munos. Reinforcement learning for continuous stochastic control problems. *Neural Information Processing Systems*, 1997.

[Mun97e] Rémi Munos. A study of reinforcement learning in the continuous case by the means of viscosity solutions. *Submitted to Machine Learning Journal*, 1997.

[Now95] Ann Nowé. Fuzzy reinforcement learning. an overview. *Advances in Fuzzy Theory and Technology*, 1995.

[Oks95] Bernt Oksendal. *Stochastic Differential Equations, an Introduction with Applications.* Springer-Verlag, 1995.

[Put94] Martin L. Puterman. *Markov Decision Processes, Discrete Stochastic Dynamic Programming.* A Wiley-Interscience Publication, 1994.

[Rez97] Mourad Rezgui. Méthodes d'apprentissage par renforcement dans le domaine continu. *Rapport de DEA Intelligence Artificielle*, 1997.

[Rup94] Jim Ruppert. A delaunay refinement algorithm for quality 2-dimensional mesh generation. *Journal of Algorithms*, 1994.

[RW72] R.A. Rescorla and A.R. Wagner. A theory of pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. *Classical conditioning II: Current research and theory*, 1972.

[SA94] S. Schaal and C. Atkeson. Robot juggling: an implementation of memory-based learning. *Control Systems Magazine*, 14, 1994.

[Sau92] Pierre Saurel. Q-learning et apprentissage par renforcement. *Rapport de DEA Sciences Cognitives*, 1992.

[SB81] Richard S. Sutton and Andrew G. Barto. Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88(2), 1981.

[SB90] Richard S. Sutton and Andrew G. Barto. Time-derivative models of pavlovian reinforcement. *Learning and Computational Neuroscience: Foundations of Adaptive Networks, M. Gabriel and J. Moore Eds.*, 1990.

[SJJ94] Satinder P. Singh, Tommi Jaakkola, and Michael I. Jordan. Reinforcement learning with soft state aggregation. *Advances in Neural Information Processing Systems*, 6, 1994.

[Sut88] R.S. Sutton. Learning to predict by the method of temporal differences. *Machine Learning*, pages 9–44, 1988.

[Sut96] Richard S. Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding. *Advances in Neural Information Processing Systems*, 8, 1996.

[SW93] R.S. Sutton and S.D. Whitehead. Online learning with random representations. *International Conference on Machine Learning*, 1993.

[Tes92] G.J. Tesauro. Practical issues in temporal difference learning. *Machine Learning*, 8: 257–277, 1992.

[Tes95] G.J. Tesauro. Temporal difference learning and td-gammon. *Communication of the ACM*, 38: 58–68, 1995.

[Tsi94] J.N. Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine Learning*, 16: 185–202, 1994.

[Wat89] Christopher J.C.H. Watkins. *Learning from delayed reward.* PhD thesis, Cambridge University, 1989.

[WD92] Christopher J.C.H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8: 279–292, 1992.

[Wer90] P. Werbos. Backpropagation through time : What it does and how to do it. *Proceedings of the IEEE*, 78(10), 1990.

[Wil87] Stewart W. Wilson. Classifier systems and the animat problem. *Machine Learning*, 2: 199–228, 1987.

[Wil92] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8: 229–256, 1992.

[ZD95] W. Zhang and T. Dietterich. A reinforcement learning approach to job-shop scheduling. *International Joint Conference on Artificial Intelligence*, 1995.

[Zom94] Albert Y. Zomaya. Reinforcement learning for the adaptive control of nonlinear systems. *IEEE Transaction on Systems, Man, and Cybernetics*, 24(2): 357–363, 1994.