Document de synthèse présenté pour obtenir

L'HABILITATION À DIRIGER DES RECHERCHES

de l'Université Pierre et Marie Curie

Spécialité :
Mathématiques Appliquées
et Application des Mathématiques

Présenté par

Rémi Munos
Centre de Mathématiques Appliquées
Ecole Polytechnique, 91128 Palaiseau Cedex

Sujet :

Contributions à l'apprentissage par renforcement et au contrôle optimal avec approximation.

Soutenue le 13 décembre 2004 devant le jury composé de

Guy Barles, Université de Tours
Hélène Frankowska, Ecole Polytechnique
Stéphane Mallat, Ecole Polytechnique
Jean-Pierre Nadal, Ecole Normale Supérieure
Gilles Pagès, Université Pierre et Marie Curie
Marc Schoenauer, INRIA Futurs

au vu des rapports de

Guy Barles, Université de Tours
Jean-Pierre Nadal, Ecole Normale Supérieure
John Tsitsiklis, Massachusetts Institute of Technology

Remerciements

Je souhaite commencer par exprimer ma profonde gratitude envers mon directeur de thèse, Paul Bourgine, dont l'ouverture et la curiosité intellectuelles sont pour moi une source d'inspiration inépuisable ; que les Sciences sont riches et belles en sa présence !

J'adresse mes plus sincères remerciements aux rapporteurs de mon habilitation : Guy Barles, qui suit mes recherches depuis des années et qui a eu le courage de se mettre dans la peau d'un cognitiviste (acte rare de la part d'un mathématicien !) pour saisir la problématique particulière de l'apprentissage par renforcement ; Jean Pierre Nadal, pour l'intérêt qu'il témoigne ainsi pour ce projet d'habilitation ; et John Tsitsiklis, dont les travaux ont inspiré certaines de mes recherches.

Je remercie vivement les autres membres du jury : Hélène Frankowska, avec qui les échanges sont toujours pleins d'intérêt et de CREAtivité ; Stéphane Mallat et Marc Schoenauer pour leur soutien chaleureux et la confiance qu'ils m'ont témoignée à mon arrivée au CMAP ; et Gilles Pagès pour avoir accepté aimablement de participer à ce jury.

Je remercie tous les autres membres du CMAP, laboratoire où il règne une ambiance joyeuse, sympathique et où il y fait bon travailler ! Je souhaite remercier en particulier Jeanne, Véronique et Liliane. J'en profite pour exprimer le plaisir que j'ai de partager mon bureau avec Randal Douc, qui m'explique toujours (même quand je ne l'écoute plus...) ses derniers résultats révolutionnaires. J'espère pouvoir travailler un jour avec lui (depuis le temps qu'on se le dit !). Je remercie Emmanuel Gobet, dont l'attitude responsable et humaine m'inspire plus que je saurais lui dire, et qui m'a encouragé à entreprendre cette habilitation.

J'exprime ma reconnaissance à Andrew Moore et tous les membres de l'équipe AUTON du Robotics Institute de CMU, en particulier Geoffroy Gordon, Jeff Schneider, Justin Boyan et Leemon Baird avec qui j'ai partagé mes années de postdoc. Il régnait alors dans ce laboratoire une dynamique particulière, où il semblait que le domaine Machine Learning se construisait là, au cours de réunions nourries de pizzas et arrosées de coca : tout paraissait réalisable. Je souhaite aussi exprimer ma gratitude envers Andy Barto, Richard Sutton et Harold Kushner, personnes dont le comportement et les recherches m'ont grandement inspiré.

Je désire aussi remercier tous ceux avec qui j'ai travaillé ou échangé des idées ; parmi eux, Guillaume Deffuant, Olivier Sigaud, Frédérick Garcia, Bruno Scherrer, Frédéric Bonnans, Hasnaa Zidani, Olivier Bokanowski, Sophie Martin, et bien d'autres encore, trop nombreux pour pouvoir tous être cités ici.

Enfin à mes amis : Guylène, dont le combat dans l'adversité brille comme un phare dans l'obscurité et me ramène à l'essentiel d'une vie. Cécile, qui m'a entraîné plus tôt que prévu dans ce projet d'habilitation, je la remercie pour son attitude courageuse et constructive. A Malai, qui se soucie de mon habilitation et de mon développement personnel. A tous mes amis qui luttent pour construire une citadelle de paix en eux-même et dans leur environnement.

A Daisaku Ikeda, qui, par ses écrits et son combat, me donne l'espoir et le courage de devenir un être un peu plus humain, chaque jour.

A mes parents, pour leurs encouragements (que je n'ai pas toujours écoutés d'ailleurs !) à travailler dur et à poursuivre des études scientifiques, je leur en suis aujourd'hui très reconnaissant.

Introduction

Ce document présente mes activités de recherche depuis la soutenance de mon Doctorat en 1997. L'essentiel de ces travaux a été réalisé de 1998 à 2000 au *Robotics Institute* de *Carnegie Mellon University* (Pittsburgh, Etats-Unis) au cours d'années postdoctorales, et depuis 2000, au *Centre de Mathématiques Appliquées* de l'*Ecole Polytechnique,* où je suis actuellement Professeur Chargé de Cours.

Mes recherches se situent au croisement des disciplines *Mathématiques Appliquées* et *Intelligence Artificielle :* elles concernent des modèles d'apprentissage pour la prise de décisions en milieu complexe et incertain.

Les travaux résultants sont, à l'image de mon parcours personnel, pluridisciplinaires. Si mes premières contributions portent sur les domaines Intelligence Artificielles et Machine Learning, depuis 4 ans, j'oriente mes recherches vers les Mathématiques Appliquées, direction que je désire poursuivre.

Les contributions présentées sont regroupées en 4 thèmes :

1. Algorithmes d'apprentissage par renforcement sur maillage uniforme.
2. Méthodes de raffinement de maillage et d'allocation de ressources pour la résolution numérique de problèmes de contrôle optimal.
3. Programmation dynamique avec approximation.
4. Analyse de sensibilité par rapport à des paramètres de contrôle.

Les publications concernées sont les suivantes.

Liste des travaux présentés

Revues avec comité de lecture

[1] E. Gobet et R. Munos. *Sensitivity analysis using Itô-Malliavin calculus and martingales. Application to stochastic optimal control.* 39 pages. A paraître dans **SIAM Journal on Control and Optimization**, 2004.

[2] R. Munos. *Algorithme d'itération sur les politiques avec approximation linéaire.* 12 pages. **Journal Electronique d'Intelligence Artificielle**, 4-37, 2004.

[3] R. Munos et A. Moore. *Variable Resolution Discretization in Optimal Control*[1]. **Journal of Machine Learning**, 49, p. 291-323, 2002.

[4] R. Munos. *A study of Reinforcement Learning in the Continuous case by the means of Viscosity Solutions.* **Journal of Machine Learning**, 40, p. 265-299, 2000.

Conférences internationales avec comité de lecture

[5] R. Munos. *Error Bounds for Approximate Policy Iteration.* **Proceedings of the Twentieth International Conference on Machine Learning**, AAAI Press, p. 560-567, 2003.

[6] R. Munos. *Efficient Resources Allocation for Markov Decision Processes*[2]. **Advances in Neural Information Processing Systems 14**, MIT Press, p. 1571-1578, 2001.

[7] R. Munos et A. Moore. *Rates of Convergence for Variable Resolution Schemes in Optimal Control.* 8 pages. **Proceedings of the Seventeenth International Conference on Machine Learning**, Morgan Kaufmann Publishers, p. 647-654, 2000.

[8] R. Munos et A. Moore. *Influence and Variance of a Markov Chain : Application to Adaptive Discretization in Optimal Control.* **IEEE Conference on Decision and Control**, p. 1464-1470, 1999.

[1] Cet article contient des extrait de [8] et [9].
[2] Cet article est une version courte (sans les preuves) de [19].

[9] R. Munos et A. Moore. *Variable resolution discretization for high-accuracy solutions of optimal control problems.* **Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence**, p. 1348-1356, 1999.

[10] R. Munos, L. Baird et A. Moore. *Gradient Descent Approaches to Neural-Net-Based Solutions of the Hamilton-Jacobi-Bellman Equation.* 6 pages. **Proceeding of the International Joint Conference on Neural Networks**, 1999.

[11] R. Munos et A. Moore. *Barycentric Interpolator for Continuous Space and Time Reinforcement Learning.* **Advances in Neural Information Processing Systems 11**, MIT Press, p. 1024-1030, 1998.

[12] R. Munos. *A general convergence method for Reinforcement Learning in the continuous case*[3]. **Proceedings of the Tenth European Conference on Machine Learning**, *Lecture Notes in Artificial Intelligence* n°1398, Springer Verlag, p. 170-182, 1998.

[13] R. Munos et P. Bourgine. *Reinforcement Learning for Continuous Stochastic Control Problems.* 7 pages. **Advances in Neural Information Processing Systems 10**, MIT Press, 1997.

[14] R. Munos. *A convergent Reinforcement Learning algorithm in the continuous case based on a Finite Difference method.* **Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence**, p. 826-832, 1997.

[15] R. Munos. *Finite-Element methods with local triangulation refinement for continuous Reinforcement Learning problems.* 12 pages. **European Conference on Machine Learning,** *Lecture Notes in Artificial Intelligence* n°1224, Springer Verlag, 1997.

Articles soumis

[16] R. Munos. *Error Bounds for Approximate Value Iteration.* 13 pages. Soumis à **SIAM Journal on Control and Optimization**, 2004.

[17] C. Barrera-Esteve, F. Bergeret, E. Gobet, A. Meziou, R. Munos et D. Reboul-Salze. *Numerical methods for the pricing of Swing options : a stochastic control approach.* 24 pages. Soumis à **Methodology and Computing in Applied Probability**, 2004.

[18] R. Munos et H. Zidani. *Consistency of a Simple Multidimensional Scheme for Hamilton-Jacobi-Bellman Equations.* 4 pages. Note soumise au **C. R. Acad. Sci. Paris**, Ser. I Math, 2004.

Rapports internes

[19] R. Munos. *Decision-Making under Uncertainty : Efficiently Estimating where Extra Resources are Needed.* 15 pages. **Rapport Interne N°550**, **Ecole Polytechnique**, 2004.

[20] E. Gobet et R. Munos. *Sensitivity analysis using Itô-Malliavin calculus and martingales. Numerical implementation.* 9 pages. **Rapport Interne N°498**, **Ecole Polytechnique**, 2004.

Thèse

[T] R. Munos. *Apprentissage par Renforcement, étude du cas continu.* Thèse de doctorat de l'**Ecole des Hautes Etudes en Sciences Sociales**. 123 pages. 1997.

[3] Cet article est extrait de [4].

Table des matières

Contributions à l'apprentissage par renforcement et au contrôle optimal avec approximation

1 Introduction

Nous ne donnons ici qu'une très brève introduction au domaine de l'*apprentissage par renforcement* (A/R). Nous renvoyons le lecteur intéressé aux références usuelles : [SB98] pour une introduction historique et intuitive du domaine et [BT96] pour une présentation plus formelle. La problématique particulière du cas continu est décrite dans la thèse [T].

L'A/R aborde le problème de l'acquisition automatisée de compétences pour la prise de décisions en milieu complexe et incertain. Il s'agit d'apprendre "par l'expérience" une stratégie comportementale en fonction des échecs ou succès constatés (les *renforcements* ou *récompenses*) résultants des prises de décisions passées.

Ce domaine de recherche apparaît vers la fin des années 1970 du constat que des modèles de mise-à-jour des paramètres de décision en *Neurosciences* et en *Psychologie expérimentale* sont identiques : les poids synaptiques se "renforcent" lors des transmission neuronales selon des mêmes règles formulées dans le conditionnement animal [RW72, Wat89, SB90]. Une théorie d'un apprentissage par renforcement indépendant du substrat sur lequel il s'applique semble possible et un courant de recherche naît dans les communautés *Machine Learning* et *Intelligence Artificielle*. Une formalisation mathématique apparaît dans les années 1980 utilisant essentiellement des modèles discrets et des outils statistiques (des exemples notables sont les algorithmes *Q-learning* [WD92, Tsi94] et *Temporal Difference* [Sut88, Day92]).

Actuellement, on regroupe sous le terme A/R les méthodes visant à résoudre de manière adaptative un problème de contrôle optimal stochastique sous, au moins, une des deux problématiques suivantes :

$\mathcal{P}1$ **:** les dynamique d'état ou les récompenses sont partiellement inconnues.

$\mathcal{P}2$ **:** la grande complexité du problème interdit sa résolution exacte.

Dans le premier cas, l'agent qui apprend à commander le système peut faire des expériences afin d'explorer le domaine, mais il ne dispose pas de modèle des dynamiques. On parle d'*A/R direct* lorsqu'il est dépourvu d'un apprentissage des dynamiques d'état (par exemple le *Q-learning*), ou bien d'*A/R indirect* lorsqu'il est accompagné d'un apprentissage supervisé de ces dynamiques. Le travail de thèse [T], brièvement résumé au paragraphe suivant, entre dans cette première problématique.

Dans le deuxième cas, il devient nécessaire d'utiliser une méthode de résolution approchée. On parle de *Programmation Dynamique avec approximation* (*Approximate Dynamic Programming*) lorsque l'on cherche une représentation paramétrée de la fonction valeur [BT96], de *Recherche directe de politique* (*Direct Policy Search*) lorsque c'est la stratégie d'action (le *contrôle en boucle fermée* ou *politique*) qui est paramétrée [Wil92, BB01, MT03], ou bien de méthode *Actor-Critic*, lorsque la fonction valeur *et* la politique sont paramétrées [KB99, SMSM00]. L'essentiel du travail réalisé depuis la thèse porte sur cette problématique d'approximation.

1.1 Résumé du travail de thèse

Le travail de thèse [T] concerne le cas où le temps et l'espace sont modélisés par des variables continues. Il établit un cadre théorique pour l'A/R en continu et permet la construction d'algorithmes simples à mettre en œuvre et dont la convergence est garantie.

Ce travail part du constat que le cas continu génère deux problèmes théoriques nouveaux par rapport au cas discret :

- La discrétisation de l'espace d'état entraîne la perte du caractère markovien de la succession d'états visités, hypothèse nécessaire à la convergence des algorithmes d'A/R en discret.
- L'utilisation de représentations paramétrées de la fonction valeur afin de minimiser le résidu de Bellman peut mener à des solutions très différentes de la fonction valeur espérée. Ce problème provient de la non-unicité des solutions généralisées des équations de Hamilton-Jacobi-Bellman.

La méthode développée dans la thèse consiste à utiliser des schémas d'approximation numérique perturbés de type chaîne de Markov [KD01] –la perturbation venant de l'incertitude sur les paramètres (probabilités de transition et récompenses). Ces paramètres peuvent être estimés en-ligne en observant des bouts de trajectoires. Nous donnons des conditions sur la précision des estimations permettant de garantir la convergence des approximations. Les outils utilisés sont les *solutions de viscosité* [Bar94, BCD97] pour l'analyse de la fonction valeur et les approximations numériques [BS91].

Des algorithmes d'A/R directs et indirects s'en déduisent. Leur convergence ne découle plus de propriétés statistiques des séquences d'états observés (comme pour le cas discret) mais de propriétés géométriques des dynamiques d'état.

1.2 Travaux présentés

Les chapitres qui suivent présentent les contributions du travail réalisé depuis la thèse et sont regroupées en en quatre thèmes :

- *Algorithmes d'A/R sur maillage uniforme.* Il s'agit de développements des travaux de thèse et correspondent aux articles [4, 13, 14, 12, 11]. Le chapitre 2 les présente brièvement.
- *Méthodes de raffinement de maillage adaptées aux équations de Hamilton-Jacobi-Bellman.* Le chapitre 3 présente des procédures de raffinement à critère local et global et replace cette approche dans le cadre de l'allocation de ressources. Ces travaux correspondent aux articles [15, 3, 8, 9, 7, 19, 18].
- *Programmation dynamique avec approximation.* Les techniques de discrétisation précédentes se heurtent à *la malédiction de la dimension.* En dimension élevée il est nécessaire de considérer des méthodes de résolution approchée. Le chapitre 4 présente les travaux [10, 5, 16, 2] qui portent sur des majorations d'erreur lors d'approximations de la fonction valeur.
- *Analyse de sensibilité par rapport à des paramètres de contrôle.* Le chapitre 5 présente notre contribution [1, 20, 17] à l'estimation du gradient de la mesure de performance par rapport aux paramètres de la politique.

2 Algorithmes d'A/R sur maillage uniforme

Les articles [4, 13, 14, 12, 11] proposent des algorithmes d'A/R inspirés de schémas d'approximation numériques de type chaîne de Markov [KD01]. Nous résumons brièvement ces schémas maintenant et présentons notre contribution à la section suivante.

2.1 Schémas numériques

Nous restons volontairement imprécis dans cette présentation du problème de contrôle optimal afin de ne pas alourdir la rédaction et renvoyons le lecteur soucieux de rigueur à [FR75, Kry80, FS93, YZ99]. Soit $(X_t)_{t\geq 0}$ un processus de diffusion à valeurs dans $X \subset I\!R^d$ (l'espace d'état), solution de la dynamique stochastique contrôlée

$$X_t = x + \int_0^t f(X_s, u_s)ds + \int_0^t \sigma(X_s, u_s)dW_s, \tag{1}$$

où $(W_t)_{t\geq 0}$ est un mouvement brownien $q-$dimensionnel défini sur un espace de probabilité filtré $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbf{P})$ satisfaisant les conditions habituelles. Nous passons sous silence les hypothèses sur la régularité des coefficients f et σ garantissant l'existence d'une unique solution forte à l'équation précédente. Nous cherchons un contrôle $(u_t)_{t\geq 0}$ (mesurable et adapté à la filtration $(\mathcal{F}_t)_{t\geq 0}$, à valeurs dans $U \subset I\!R^m$, supposé compact) qui maximise un *gain* (appelé aussi *critère* ou *mesure de performance*), par exemple dans le cas actualisé avec horizon temporel infini,

$$J(x, u.) = \mathbb{E}\left[\int_0^T e^{-\beta t} r(X_t, u_t)dt + e^{-\beta T} R(X_T) \mid X_0 = x\right], \tag{2}$$

où $r : X \times U \to I\!R$ et $R : X \to I\!R$ sont les *fonctions récompense* courante et finale, $\beta > 0$ est le taux d'actualisation et T un temps de sortie du domaine X.

La *fonction valeur* est le gain maximum :

$$V(x) = \sup_{u.} J(x, u.). \tag{3}$$

Elle satisfait, au sens des solutions de viscosité [CL83, CIL92, FS93, Bar94, BCD97], une équation aux dérivées partielles (EDP) non-linéaire du second ordre : l'équation de *Hamilton-Jacobi-Bellman* (HJB),

$$-\beta V(x) + \max_{u\in U}\left[r(x, u) + \sum_{i=1}^{d} \partial_{x_i} V(x) f_i(x, u) + \frac{1}{2}\sum_{i,j=1}^{d} \partial^2_{x_i x_j} V(x)[\sigma\sigma']_{ij}(x, u)\right] = 0. \tag{4}$$

Nous considérons les schémas d'approximation numérique de type chaînes de Markov de Kushner et Dupuis [KD01]. Le problème de contrôle optimal en temps continu est discrétisé, à un pas h, en un *Processus de Décision Markovien* (PDM) [Ber87, Put94]. La propriété qui garantit la convergence des fonctions valeur V^h du PDM vers la fonction valeur (3), est *la consistance locale* du schéma.

Notons $X_h = \{x_i \in X\}$ les points de discrétisation (qui forment un *maillage*, ou *grille*). Ceux-ci définissent les états du PDM. La résolution spatiale h signifie qu'il existe, de tout point de X un point de la grille situé à une distance au plus égale à h. Notons $p(x_j|x_i, u)$ la probabilité de transition d'un état x_i vers l'état x_j pour le contrôle $u \in U$. Notons $\tau(x_i, u)$ les *pas de temps* (qui peuvent dépendre de l'état x_i et du contrôle u), et

$$\tau_h = \min_{x_i\in X_h, u\in U} \tau(x_i, u). \tag{5}$$

L'équation de *programmation dynamique* (PD) du PDM est

$$V^h(x_i) = \sup_{u\in U}\left[e^{-\beta\tau(x_i,u)} \sum_j p(x_j|x_i, u)V^h(x_j) + r(x_i, u)\tau(x_i, u)\right]. \tag{6}$$

La propriété de consistance locale exprime la conservation (à un $o(\tau_h)$ près) des deux premiers moments des dynamiques d'état : pour tout état $x_i \in X_h$ et contrôle $u \in U$, les états successeurs $x_j \in X_h$ vérifient :

$$\begin{aligned} \mathbb{E}[x_j - x_i] &= f(x_i,u)\tau(x_i,u) + o(\tau_h), \\ \mathbb{E}\big[\,(x_j - \mathbb{E}[x_j])\,(x_j - \mathbb{E}[x_j])'\,\big] &= [\sigma\sigma'](x_i,u)\tau(x_i,u) + o(\tau_h). \end{aligned} \tag{7}$$

La monotonie du schéma résulte de la positivité des probabilités $p(x_j|x_i,u)$, et le résultat général de convergence de Barles et Souganidis [BS91] s'applique (il faut néanmoins faire attention aux dynamiques sur les bords afin d'établir un principe d'unicité forte, mais nous ne détaillerons pas ce point ici, car il est abordé dans la thèse [T], et renvoyons simplement aux références [BP88, BP90, Bar94, FS93]).

2.2 Algorithmes d'Apprentissage par Renforcement

Nous nous plaçons dans la problématique $\mathcal{P}1$ de l'A/R où les dynamiques d'état et les fonctions récompenses sont partiellement inconnues. Nous résumons brièvement les articles qui s'inspirent des travaux de thèse [T] et les prolongent.

Convergence des schémas perturbés. L'article [4] résume les principaux résultats de la thèse [T] sur la convergence des schémas perturbés lorsqu'on utilise des dynamiques d'état approchées, quand les paramètres sont estimés en-ligne par l'observation de bouts de trajectoires. Ce travail traite du cas déterministe, pour lequel la dynamique s'écrit

$$\frac{dx_t}{dt} = f(x_t,u_t), \tag{8}$$

où $x_t \in X \subset I\!R^d$ est l'état et $u_t \in U$ la commande à l'instant t, et le critère à maximiser,

$$J(x,u.) = \int_0^T e^{-\beta t} r(x_t,u_t)dt + e^{-\beta T} R(x_T). \tag{9}$$

L'article [12] présente de manière simplifiée le résultat de convergence des schémas perturbés en utilisant une propriété de *contraction faible* de l'opérateur de Bellman approché (par opposition à la propriété de *contraction forte* de l'opérateur de Bellman exact) introduite dans la thèse.

Cas stochastique. L'article [13] réalisé avec Paul Bourgine traite de l'estimation directe des paramètres f et σ des dynamiques d'état (1) et des récompenses et donne des conditions de convergence pour un schéma perturbé de type Différences-Finies.

Comparaison avec Q-learning. Le problème, énoncé dans l'introduction, de la perte de la propriété de Markov lors de la discrétisation de l'espace et l'impossibilité d'utiliser les algorithmes d'A/R en discret (de type Q-learning) est détaillée dans l'article [14]. Une alternative est proposée basée sur une méthode aux Différences-Finies.

Interpolateurs barycentriques. Dans l'article [11] réalisé avec Andrew Moore, nous définissons une classe de fonctions, baptisées *interpolateurs barycentriques*, qui permet d'établir la convergence d'algorithmes d'A/R quand on utilise une approximation des dynamiques d'état. Dans le cas déterministe, le schéma numérique exprime le principe de PD sous la forme

$$V^h(x_i) = \sup_{u\in U} \left[e^{-\beta\tau(x_i,u)} V^h(x_i + \tau(x_i,u) f(x_i,u)) + r(x_i,u)\tau(x_i,u) \right].$$

Celui-ci se ramène au schéma (6) lorsque la fonction V^h aux point "décentrés" $x_i + \tau(x_i,u)f(x_i,u)$ est proche (à un $O(h)$ près) d'une interpolation barycentrique des valeurs $V^h(x_j)$ pondérées par des probabilités de transition $p(x_j|x_i,u)$. Un intérêt est de permettre l'utilisation de fonctions linéaires (ou multi-linéaires) par morceaux mais pas nécessairement continues aux interfaces, comme cela est illustré dans le chapitre suivant.

3 Maillage adaptatif et allocation de ressources

Nous considérons désormais la problématique $\mathcal{P}2$ de l'A/R où il s'agit d'affronter des problèmes de dimension élevée. Face à *la malédiction de la dimension*, nous développons ici des méthodes de maillage adaptatif.

Nous détaillons dans ce chapitre 4 contributions :

1. une heuristique globale de raffinement de maillage (section 3.3),
2. un schéma d'approximation dans le cas stochastique (section 3.4),
3. une majoration d'erreur de la fonction valeur (section 3.5),
4. un formalisme probabiliste pour l'allocation de ressources (section 3.6).

3.1 Le schéma d'approximation

On considère ici un problème déterministe défini par la dynamique (8) et le gain (9). Une généralisation au cas stochastique est proposée à la section 3.4.

3.1.1 Une implémentation efficace

Une première ébauche utilisant une triangulation adaptative de Delaunay [Mid93, Rup94] est développée dans l'article [15]. Cependant, ce type de triangulation étant difficilement manipulable en dimension supérieure à 3, cette approche est abandonnée. Les méthodes numériques plus efficaces développées dans le travail [3] réalisé avec Andrew Moore sont maintenant présentées.

Nous considérons des fonctions linéaires par morceaux sur une triangulation de l'espace d'état X (supposé rectangulaire). L'espace est partitionné, selon une arborescence de type *kd-tree* [FBF77], en $d-$rectangles sur lesquels est implémentée une triangulation de Coxeter-Freudenthal-Kuhn [Moo92].

Ces choix sont faits pour raison d'efficacité numérique : la représentation arborescente permet de déterminer rapidement le rectangle $\mathcal{R}$ contenant un point donné $x \in X$. Ensuite, le choix des $(d+1)$ sommets $(x_i)_{1\leq i\leq d+1}$ définissant le $d-$simplexe $\mathcal{S} \ni x$ parmi les 2^d sommets de $\mathcal{R}$ se fait en $O(d \ln d)$ opérations, et les coordonnées barycentriques de x dans $\mathcal{S}$ s'en déduisent immédiatement. Enfin, la valeur en x est la combinaison linéaire des valeurs $V^h(x_i)$ aux sommets du simplexe pondérées par les coordonnées barycentriques. Le gradient se calcule facilement et permet de déduire la *politique* $\pi(x)$ (le contrôle en boucle fermée) :

$$\pi(x) \in \arg\max_{u\in U}[r(x,u) + \nabla V(x) \cdot f(x,u)].$$

Remarquons que cette représentation n'assure pas la continuité des fonctions aux interfaces (côtés des rectangles). Cependant cela ne gène en rien la convergence des schémas numériques –les propriétés de consistance et de monotonie étant préservées. Ceci autorise une souplesse dans les procédures de raffinement de maillage que ne permettent pas les représentations continues.

3.1.2 Construction du PDM discret

Pour une grille X_h donnée, il reste à définir les probabilités de transition $p(x_j|x_i,u)$ et les pas de discrétisation $\tau(x_i,u)$. Pour chaque état $x_i \in X_h$ et contrôle $u \in U$, on intègre (selon une méthode d'Euler ou de Runge-Kutta) des bouts de trajectoires (8) partant de x_i et pour un contrôle constant u, jusqu'à ce que la trajectoire entre dans un nouveau simplexe $\mathcal{T}$. Le temps de parcours définit $\tau(x_i,u)$ et les coordonnées barycentriques du point d'entrée dans $\mathcal{T}$ définissent les probabilités de transitions vers ses sommets $(x_j)_{1\leq j\leq d+1}$.

Nous omettons les détails de l'implémentation (ainsi que les problèmes qui se posent aux bords du domaine) et renvoyons le lecteur intéressé à [3].

Pour résoudre le PDM nous utilisons un algorithme de PD où nous tirons profit de sa structure topologique particulière et de sa faible connectivité (chaque état possède $(d+1)$ successeurs). Nous utilisons un algorithme d'*itération sur les politiques modifié* (*Modified Policy Iteration*) [Put94]

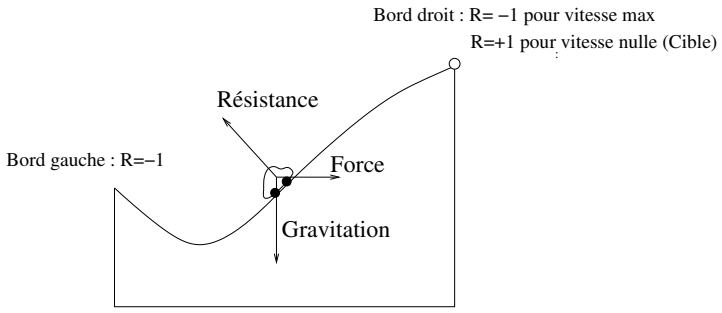


FIG. 1 – La voiture sur la colline

précédé d'un tri topologique [Ski97] où, lors de la phase d'évaluation de la politique, les valeurs sont itérées selon l'ordre obtenu par le tri en prenant en compte les cycles éventuels. Cette méthode s'apparente aux techniques de *Fast Marching* [Set99]. Le gain observé en temps CPU est de l'ordre de 100 par rapport à une méthode habituelle de type *itération sur les valeurs*.

3.2 Critère local de raffinement de maillage

Dans cette partie, nous décrivons des procédures de raffinement de maillage utilisant un critère local, c'est à dire que la subdivision d'une cellule dépend d'une mesure d'erreur locale. Nous suivons en cela l'approche habituelle de résolution adaptative des équations de HJB (voir par exemple les grilles adaptatives [Grü97] ou les méthodes de redistribution [TTZ03]).

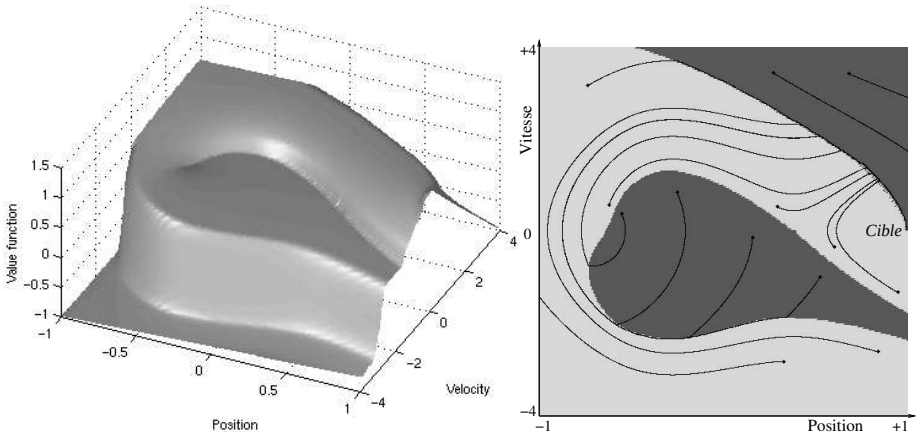
Afin de faciliter l'explication des procédures développées, nous les illustrons sur un problème de contrôle simple : *la voiture sur la colline* représentée sur la Figure 1. Il s'agit d'un problème de dimension 2 où une voiture (définie par sa position et sa vitesse) tente de monter au sommet d'une colline et de s'y arrêter. Le contrôle porte sur l'accélération de la voiture et prend 2 valeurs possibles : $U = \{-1, +1\}$. La puissance du moteur n'étant pas suffisante pour grimper directement à partir d'un état à l'arrêt, la solution consiste à partir en arrière, prendre de l'élan et repartir en avant. Les dynamiques de ce problème sont décrites dans [MA95] et les fonctions récompense sont les suivantes : la récompense courante $r = 0$. Sur le bord gauche, la récompense est $R = -1$, et sur le bord droit, R varie linéairement de $+1$ à -1 en fonction de la vitesse terminale : $R = +1$ lorsque la voiture atteint le bord droit (le sommet de la colline) avec une vitesse nulle (la cible).

Ainsi, il s'agit d'amener la voiture à la cible en temps minimum (le gain est actualisé), tout en évitant de sortir du bord gauche du domaine. La Figure 2 représente la fonction valeur (FV) et la politique optimale.

Remarquons la discontinuité de la FV le long d'une frontière qui sépare les états à partir desquels il est possible d'atteindre la cible de ceux à partir desquels, quelque soit le contrôle utilisé, la sortie par le bord gauche est inévitable.

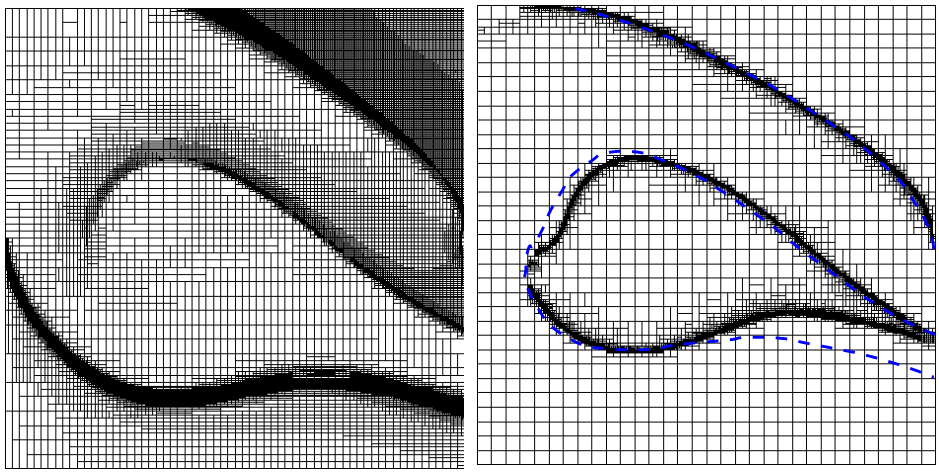
En utilisant un critère de raffinement basé sur l'erreur d'interpolation locale de la fonction valeur, nous obtenons le maillage présenté sur la Figure 3a (d'autres critères sont développés dans [3]).

Ce maillage fournit une très bonne approximation de la fonction valeur ; il présente une résolution locale très fine aux régions les plus irrégulières. Cependant, en comparant les Figures 2b et 3a, nous remarquons deux points : (1) le contrôle optimal est constant ($u = +1$) autour de la frontière de discontinuité de la FV (zone la plus fortement raffinée), (2) le maillage n'est pas particulièrement fin aux frontières de transition du contrôle optimal. Ceci nous incite à questionner la pertinence de cette procédure. Faut-il dépenser tant de ressources computationnelles (en mémoire et temps CPU de résolution) pour approcher la discontinuité de V alors que la commande optimale n'y est pas sensible ? Ne faudrait-il pas mieux allouer ces ressources pour améliorer les frontières de transition de la commande optimale ?



(a) La fonction valeur (b) La politique optimale

FIG. 2 – (a) Fonction valeur et (b) politique optimale en deux niveaux de gris (gris clair correspond à $u = +1$) ainsi que quelques trajectoires optimales.



(a) selon la fonction valeur. (b) selon la politique.

FIG. 3 – Grilles résultant d'un critère local

La FV n'est qu'un outil introduit en PD pour déterminer la politique optimale. Le but n'est pas tant la qualité d'approximation de la FV que la performance de la politique qui s'en déduit.

Une alternative consiste à raffiner le maillage selon la dissimilarité locale *de la politique*. Le maillage résultant est présenté à la Figure 3b. Ce maillage ne semble être affiné qu'aux endroits de transition de la commande. Cependant cette fine résolution ne correspond pas à la frontière de transition de la commande optimale (représentée en trait pointillé) : la mauvaise estimation de la FV (due au faible raffinement autour de sa discontinuité) entraîne une mauvaise localisation de la frontière de transition de la commande optimale.

Nous observons sur ce simple exemple l'influence non-locale de la qualité d'approximation de la FV sur les frontières de transition du contrôle optimal.

La question qui se pose est : comment répartir au mieux les ressources disponibles (par exemple le nombre maximum de points du maillage) pour maximiser la performance des politiques déduites des FV représentées ?

3.3 Heuristique globale de raffinement

Nous souhaitons calculer une bonne approximation de la FV aux frontières de transition de la commande optimale, afin de la localiser précisément. Nous introduisons deux outils :

- une *mesure d'influence* des régions de l'espace où l'erreur d'interpolation locale est la plus dommageable à l'approximation de la FV en un endroit particulier,
- la *variance* introduite lors de l'approximation numérique de la FV.

Nous combinons ces outils pour déterminer les régions où l'incertitude sur l'approximation de la FV a la plus grande influence sur les frontières de transition de la commande optimale.

3.3.1 Mesure d'influence

Nous définissons la mesure d'influence dans une chaîne de Markov. Notons X_N l'espace d'état (composé de N états), $\{p(y|x)\}_{x,y\in X_N}$ les probabilité de transition, $r : X_N \to \mathbb{R}$ la fonction récompense et $e^{-\beta\tau(x)}$ (avec $\tau(x) > 0$) le coefficient d'actualisation (qui peut dépendre de l'état x).

Définissons la matrice de probabilité actualisée Q d'éléments : $Q(x,y) = e^{-\beta\tau(x)}p(y|x)$. La valeur de la chaîne de Markov

$$V(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} e^{-\beta[\tau(x_0)+\tau(x_1)+\cdots+\tau(x_{t-1})]} r(x_t)|x_0 = x\right], \tag{10}$$

est l'unique solution de l'équation de Bellman $V = r + QV$, soit

$$V = (I - Q)^{-1} r.$$

Ainsi $V(x)$ est une combinaison linéaire des récompenses $r(y)$ pondérées par les éléments de la résolvante $(I-Q)^{-1}(x,y)$, appelés *influence de l'état y sur l'état x* et notés $\mathcal{I}(y|x)$. En considérant V comme une fonction des variables r, l'influence est la dérivée partielle $\mathcal{I}(y|x) = \frac{\partial V(x)}{\partial r(y)}$.

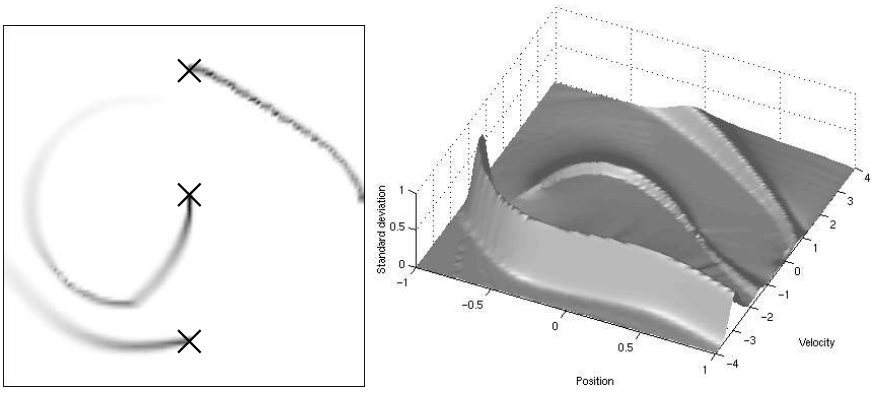
Pour un état x donné, le vecteur des influences sur x, noté $\mathcal{I}(\cdot|x)$, satisfait

$$\mathcal{I}(\cdot|x) = Q'\mathcal{I}(\cdot|x) + \mathbf{1}_x, \tag{11}$$

où $\mathbf{1}_x$ est le x^e vecteur canonique. En utilisant un argument de contraction en norme L_1, on déduit que l'influence $\mathcal{I}(\cdot|x)$ peut être calculée par itérations sur les valeurs.

On généralise la notion d'influence d'un état y sur un vecteur f : $\mathcal{I}(y|f) = \sum_x \mathcal{I}(y|x)f(x)$ et l'influence d'un vecteur g sur un vecteur f : $\mathcal{I}(g|f) = \sum_y \mathcal{I}(y|f)g(y)$.

A titre d'illustration, la Figure 4a représente l'influence sur 3 états (les croix) pour la chaîne de Markov construite par discrétisation sur un maillage donné du problème de la voiture sur la colline (pour la politique déduite de la FV approchée).



(a) Influence sur 3 points (b) Ecart-type

FIG. 4 – (a) Influence (en teintes de gris) et (b) l'écart-type σ pour la voiture sur la colline.

3.3.2 Variance d'une chaîne de Markov

La chaîne de Markov admet pour valeur (10). Sa *variance* est

$$\sigma^2(x) = \mathbb{E}\left[\Big(\sum_{t=0}^{\infty} e^{-\beta[\tau(x_0)+\tau(x_1)+\cdots+\tau(x_{t-1})]} r(x_t) - V(x) \Big)^2 | x_0 = x \right].$$

Dans l'article [8], nous montrons que la variance est solution d'une équation de Bellman

$$\sigma^2(x) = e^{-2\beta\tau(x)} \sum_y p(y|x)\sigma^2(y) + e(x)$$

où $e(x)$, contribution immédiate à la variance due à l'interpolation locale, vaut

$$e(x) = \sum_y p(y|x) \big[e^{-\beta\tau(x)} V(y) - V(x) + r(x) \big]^2.$$

Ainsi, la variance se calcule facilement par une méthode de PD usuelle. A titre d'illustration, la Figure 4b représente, pour une grille donnée, l'écart-type σ pour la voiture sur la colline.

Le système continu est déterministe ; aussi la variance de la chaîne discrète mesure la diffusion numérique introduite (à cause de l'interpolation numérique) par la discrétisation.

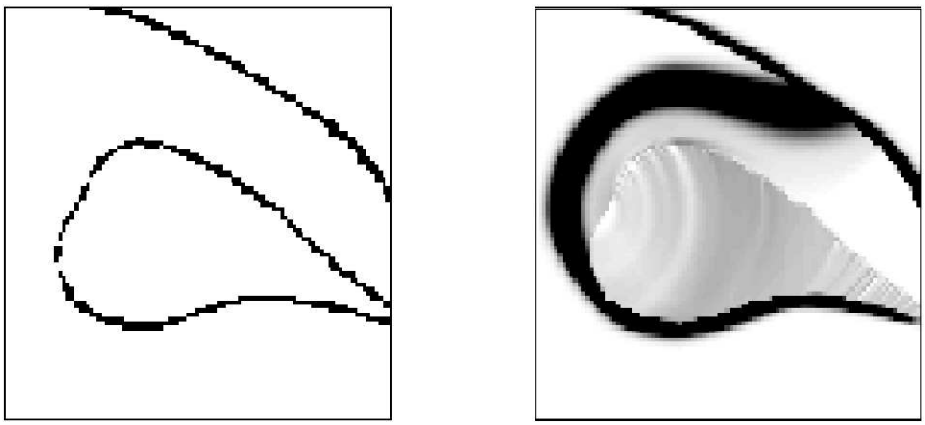
3.3.3 Heuristique de raffinement

L'heuristique de raffinement proposée dans [3] est illustrée ici sur le problème de la voiture sur la colline et quelques autres tâches de contrôle.

Afin d'améliorer la qualité d'approximation de la FV autour des frontières de transition Γ de la commande optimale, nous raffinons le maillage selon un critère qui évalue la contribution $\sigma(x)\mathcal{I}(x|\mathbf{1}_\Gamma)$, en tout état x, de l'influence $\mathcal{I}(\sigma|\mathbf{1}_\Gamma)$ de l'incertitude dans l'approximation de la FV (mesurée par σ) sur la frontière Γ. La Figure 5 représente, pour un maillage donné, ces frontières (a) et l'influence sur celles-ci (b).

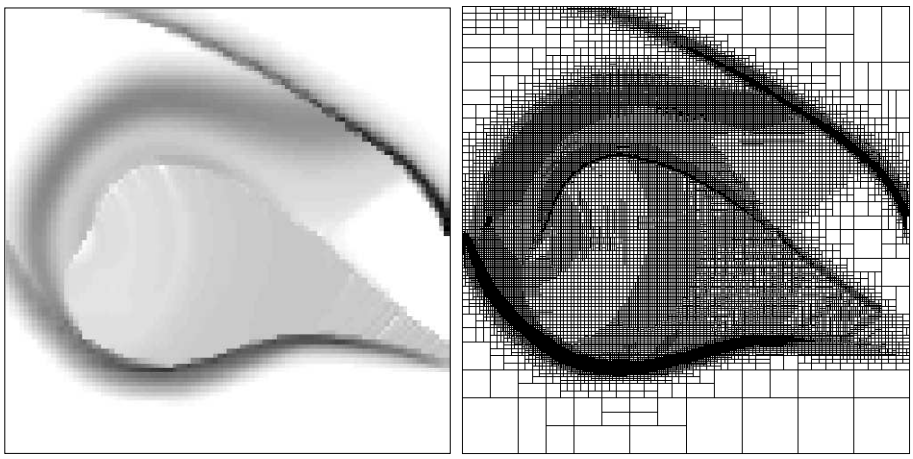
La Figure 6 représente le critère de raffinement (a) et le maillage résultant (b).

Ce critère ne raffine pas le maillage autour des discontinuités de la FV à moins que cela soit nécessaire pour améliorer la qualité du contrôleur (peut-être en d'autres endroits de l'espace d'état). Ceci est crucial pour des problèmes de dimension élevée (dont certains sont détaillés au paragraphe suivant), pour lesquels le coût de raffinement autour des discontinuités serait exorbitant.



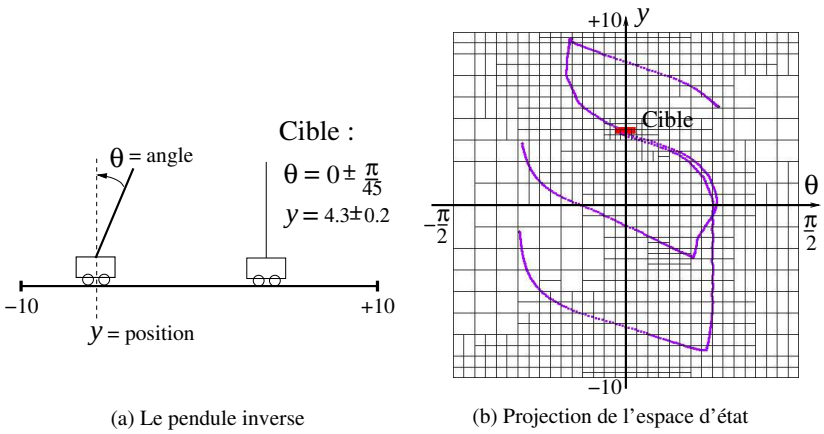
(a) Frontière de transition de la commande (b) Influence sur ces états

FIG. 5 – La frontière de transition Γ de la commande optimale et l'influence $\mathcal{I}(\cdot|\mathbf{1}_\Gamma)$ sur celle-ci.



(a) Critère $\sigma(x)\mathcal{I}(x|\mathbf{1}_\Gamma)$ (b) Le maillage résultant

FIG. 6 – Critère de raffinement et maillage résultant.



(a) Le pendule inverse (b) Projection de l'espace d'état

FIG. 7 – (a) Pendule inverse. (b) Projection du maillage (sur le plan θ, y) et plusieurs trajectoires optimales déduites.

La performance du contrôleur déduit de cette méthode est, pour un nombre global de points de discrétisation donné, bien meilleure que pour les approches locale précédemment décrites, ainsi que pour un maillage uniforme [3, Amo02].

3.3.4 Autres tâches de contrôle

Nous illustrons cette heuristique globale de raffinement sur quelques problèmes (extraits de [3]) de dimension 4.

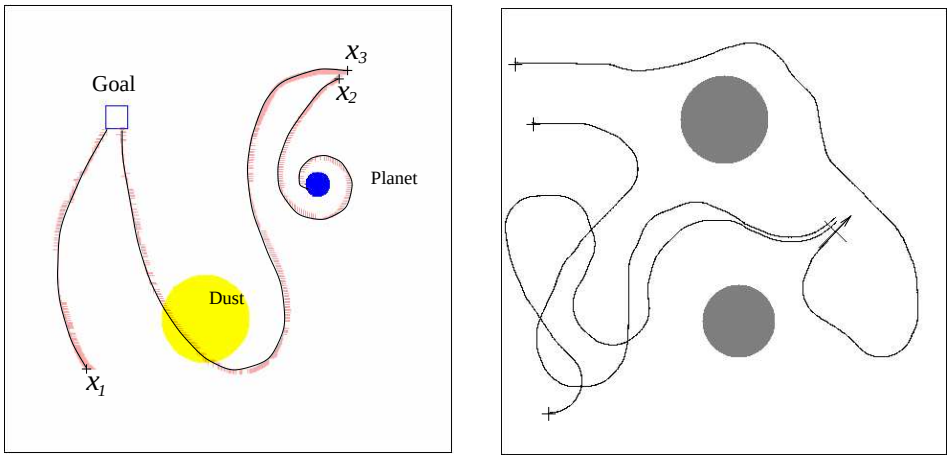
Le pendule inversé. Le système est défini par la position et la vitesse du chariot, l'angle et la vitesse angulaire du pendule. La commande est une force qui agit sur le chariot. Il s'agit d'atteindre une position d'équilibre instable en temps minimum. La Figure 7 illustre le système physique, une projection du maillage obtenu et représente quelques trajectoires déduites.

La navette spatiale. La navette est définie par sa position et sa vitesse (dans un plan) et la commande consiste à accélérer dans une des 4 directions cardinales ou à ne rien faire. Les dynamiques suivent les lois de la physique newtonienne : la navette est attirée par des objets gravitationnels présents (planète et poussière). Le but est d'atteindre une position donnée (cible) en minimisant le temps d'atteinte et la consommation de carburant (la poussée des gaz est indiquée par les petits tirés autour de la trajectoire). Quelques trajectoires sont représentées sur la Figure 8a.

Le rendez-vous d'avions. Cette application est réalisée avec Olivier Sigaud pour Dassault-Aviation. Il s'agit de commander un (ou plusieurs) avions pour atteindre un objectif (position et angle donnée) à un instant précis, tout en évitant des zones dangereuses. Dans le cas de plusieurs avions, il s'agit d'un problème de rendez-vous. Un exemple de trajectoire (pour 3 avions) est représenté sur la Figure 8b.

3.4 Le cas stochastique

L'heuristique de raffinement de maillage développée au chapitre précédent peut être généralisée au cas stochastique. On doit cependant faire attention à remplacer la mesure de variance par l'erreur d'interpolation accumulée (pour ne pas prendre en compte la diffusion intrinsèque du processus).



(a) La navette spatiale

(b) Rencontre de 3 avions

FIG. 8 – (a) Trajectoires de navette spatiale. A partir de x_1 l'objectif est directement atteignable (faible gravitation). Partant de x_2, la collision sur la planète est inévitable quelque soit la commande. A partir de x_3 le contrôle réalise un effet "fronde". (b) Le rendez-vous de 3 avions.

Ici nous décrivons un travail réalisé avec Hasna Zidani [18] qui présente un schéma d'approximation dans le cas stochastique (1).

Ce schéma est simple à implémenter et fournit une alternative intéressante aux schémas de type différences-finies [KD01] et différences finies généralisées [BZ03]. Il ne nécessite pas l'hypothèse habituelle souvent contraignante que les matrices $[\sigma\sigma'](x,u)$ sont à diagonale dominante. De plus il se construit sur des grilles non-structurées.

Soit $X_h = \{x_i\} \subset X$ une grille de résolution h. La dynamique d'état est (1). Notons $\{\alpha_j(x_i,u)\}_{1\le j\le q}$ (avec $q \le d$) les valeurs propres strictement positives de $[\sigma\sigma'](x_i,u)$ et $\{v_j(x_i,u)\}_{1\le j\le q}$ des vecteurs propres associés formant une famille orthonormée de $\mathbb{R}^d$.

Pour chaque point $x_i \in X_h$ et contrôle $u \in U$, on considère le *point décentré* $y(x_i,u) = x_i + \tau(x_i,u)f(x_i,u)$ défini par le pas de temps $\tau(x_i,u)$. Nous introduisons $(2q)$ valeurs, appelées *pas de diffusion* $\{\eta_j(x_i,u)\}_{1\le j\le q}$ et $_{-q\le j\le -1}$ qui définissent les *points diffusés* $\{z_j(x_i,u)\}_{-q\le j\le q}$ (voir Figure 9) :

$$z_j(x_i,u) = \begin{cases} y(x_i,u) + \eta_j(x_i,u)v_j(x_i,u) & \text{pour} \quad 1 \le j \le q \\ y(x_i,u) & \text{pour} \quad j = 0 \\ y(x_i,u) - \eta_j(x_i,u)v_{-j}(x_i,u) & \text{pour} \quad -q \le j \le -1 \end{cases}$$

Considérons un *interpolateur linéaire locale*, opérateur qui, appliqué à $W : X_h \to \mathbb{R}$ en $x \in X$, retourne une combinaison linéaire de W aux points $x_k \in X_h$ pondérée par des coefficients $\lambda_k(x_k|x) \ge 0$, tels que

$$\sum_k \lambda(x_k|x) = 1, \quad \sum_k \lambda(x_k|x)x_k = x$$

et tels que les points x_k dont les coefficients $\lambda_k(x_k|x)$ sont strictement positifs sont à une distance $O(h)$ de x. Une interpolation linéaire (ou multi-linéaire) locale définie sur X_h satisfait ces conditions.

Choisissons des *poids* $\{\rho_j(x_i,u)\}_{-q\le j\le q}$, coefficients positifs de somme 1.

L'approximation numérique est définie par le PDM dont l'espace d'état est X_h, l'espace de

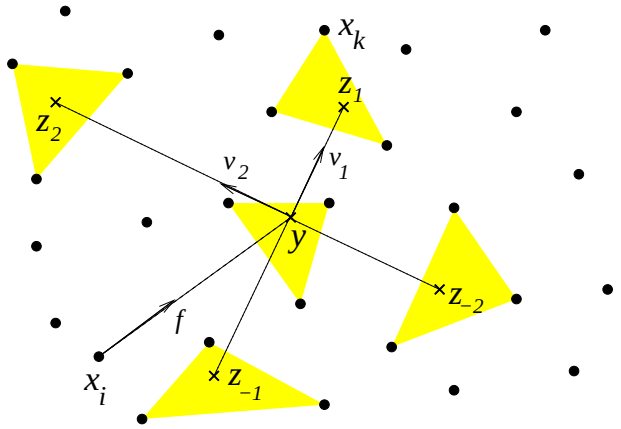


FIG. 9 – Construction du schéma numérique

contrôles U, la fonction récompense $r(x_i,u)\tau(x_i,u)$ et les probabilités de transition

$$p(x_k|x_i,u) = \sum_{j=-q}^{q} \rho_j(x_i,u)\lambda\big(x_k|z_j(x_i,u)\big).$$

La FV V^h du PDM satisfait l'équation de programmation dynamique (6). Nous montrons la proposition suivante qui garantie la convergence de V^h vers la FV V défini par (3).

Proposition 1 [18] *Sous les conditions d'équilibre entre les poids et les coefficients de diffusion :* $\forall x_i \in X_h$, $u \in U$, $1 \leq j \leq q$,

$$\begin{aligned} \rho_j(x_i,u)\eta_j(x_i,u) &= \rho_{-j}(x_i,u)\eta_{-j}(x_i,u), \\ \rho_j(x_i,u)\eta_j^2(x_i,u) + \rho_{-j}(x_i,u)\eta_{-j}^2(x_i,u) &= \alpha_j(x_i,u)\tau(x_i,u), \end{aligned} \tag{12}$$

et sous la condition CFL $h^2 = o(\tau_h)$, *avec* τ_h *défini par (5), ce schéma d'approximation est consistant localement, au sens où il satisfait (7).*

D'un point de vue numérique, il est appréciable de constater que le nombre de points utilisé dans ce schéma, dans le cas d'une interpolation linéaire locale, est au plus $2d(d+1)$.

Si l'on choisit des pas de temps constants $\tau(x_i,u) = \tau$ et des poids constants $\rho_j(x_i,u) = \frac{1}{2q}$, alors la taille du *stencil* est en $\sqrt{\tau}$. Pour satisfaire la condition CFL, on peut choisir τ équivalent à $h^{2/(1+\alpha)}$ avec $\alpha > 0$, donc la taille du stencil est en $h^{1/(1+\alpha)}$.

3.5 Erreur d'approximation de la fonction valeur

Nous approfondissons et formalisons le travail expérimental de la section 3.3 dans les articles [7, 19] présentés dans cette section et la suivante.

Le travail [7] réalisé avec Andrew Moore permet de calculer une borne sur l'erreur d'approximation de la FV en fonction des erreurs d'interpolation locale. Il s'agit d'une majoration fine qui améliore les résultats usuels de [BT96]. Appliqué à une discrétisation sur maillage adaptatif, il est possible de prédire l'effet d'un raffinement local de la résolution sur la qualité d'approximation de la FV en chaque état. Une procédure efficace de raffinement de maillage s'en déduit.

Considérons ici un PDM en temps discret, défini sur un espace continu $X \subset \mathbb{R}^d$. Le noyau de transition d'un état $x \in X$ avec un contrôle $u \in U$ vers un état dans $A \in \mathbf{A}$ (avec $\mathbf{A}$ la tribu borélienne) est notée $p(A|x,u)$. La fonction récompense est $r(x,u)$. Nous supposons un coefficient

d'actualisation constant γ. Nous définissons l'opérateur de Bellman $\mathcal{T}_u$, pour toute fonction bornée $W : X \to \mathbb{R}$, par

$$\mathcal{T}_u W(x) := \gamma \int_X p(dy|x,u)W(y) + r(x,u). \tag{13}$$

La FV satisfait alors l'équation de PD : $V(x) = \sup_{u\in U} \mathcal{T}_u V(x)$.

Ce PDM est approché par un PDM construit à partir d'une grille de N points $X_N = \{x_i\}_{1\leq i\leq N} \subset X$. Son espace d'état reste X mais les transitions se font vers l'espace fini X_N. L'espace des contrôles U est le même. Les probabilités de transition sont notées $p_N(x_i|x,u)$ et les récompenses $r_N(x,u)$. L'opérateur de Bellman approché est

$$\mathcal{T}_u^N W(x) := \gamma \sum_{i=1}^N p_N(x_i|x,u)W(x_i) + r_N(x,u),$$

et la fonction valeur V_N est solution de l'équation de PD : $V_N(x) = \sup_{u\in U} \mathcal{T}_u^N V_N(x)$. Notons que V_N est définie en tout $x \in X$ et que sa valeur en un point $x \notin X_N$ dépend des valeurs aux points de la grille, par l'intermédiaire de l'équation de PD. C'est une manière de représenter implicitement une fonction (solution d'une équation de type point-fixe) sur X à partir d'une grille X_N sans avoir à définir d'opérateur d'interpolation.

Définissons l'*erreur d'interpolation locale* de la fonction valeur :

$$e_u(x) := |\mathcal{T}_u^N V(x) - \mathcal{T}_u V(x)|$$

et l'*erreur d'approximation* de la FV : $\varepsilon(x) := |V_N(x) - V(x)|$. Il est facile de déduire une majoration uniforme sur ε en fonction de l'erreur d'interpolation globale :

$$\sup_{x\in X} \varepsilon(x) \leq \frac{1}{1-\gamma} \sup_{x\in X, u\in U} e_u(x).$$

Cependant, nous pouvons considérablement améliorer cette majoration en précisant les régions concernées. Dans [7], nous montrons qu'une borne $\bar{\varepsilon}$ sur ε satisfait l'équation de PD :

$$\bar{\varepsilon}(x) = \gamma \max_{u\in U_N(x)} \sum_{i=1}^N p_N(x_i|x,u)\bar{\varepsilon}(x_i) + \max_{u\in U'_N(x)} e_u(x) \tag{14}$$

où $U_N(x)$ et $U'_N(x)$ sont des sous-ensembles inclus dans U. Nous décrivons une méthode efficace de PD pour calculer $\bar{\varepsilon}$ en restreignant progressivement les ensembles U_N et U'_N d'une façon semblable à la procédure d'élimination d'actions de [Put94].

Pour la politique optimale déduite de l'équation (14) (l'argument du premier max), le PDM sous-jacent devient une chaîne de Markov dont on peut calculer l'influence (définie au paragraphe 3.3.1). Celle-ci mesure l'impact d'une erreur d'interpolation en x_j sur la qualité d'approximation de la FV en x_i.

Nous pouvons utiliser cette information pour une procédure de raffinement de maillage efficace. Pour cela, il faut estimer l'effet d'un raffinement de maillage sur l'erreur d'interpolation locale, effet qui dépend du procédé d'interpolation implémenté.

A titre d'exemple, considérons un problème de contrôle optimal déterministe dont la dynamique est (8) et le gain $J(x,u.) = \int_0^\infty e^{-\beta t} r(x_t,u_t)dt$. Une discrétisation temporelle (à un pas τ) est définie par l'opérateur de Bellman

$$\mathcal{T}_u W(x) := e^{-\beta\tau} W(x + \tau f(x,u)) + \tau r(x,u), \tag{15}$$

et la FV V^τ satisfait $V^\tau(x) = \sup_{u\in U} \mathcal{T}_u V^\tau(x)$. Dans le cas d'une interpolation linéaire par morceaux dans une triangulation basée sur X_N, l'opérateur de Bellman approché est

$$\mathcal{T}_u^N W(x) := e^{-\beta\tau} \sum_{i=1}^{d+1} \lambda_i(y) W(x_i) + \tau r(x,u)$$

où les $\lambda_i(y)$ sont les coordonnées barycentriques de $y := x + \tau f(x,u)$ dans le simplexe $\mathcal{S} = \{x_i\}_{1\leq i\leq d+1} \ni y$. Ici, l'erreur d'interpolation locale de V^τ dépend de sa hessienne :

$$e_u(x) = \frac{1}{2}e^{-\beta\tau}\sum_{i=1}^{d+1}\lambda_i(y)(x_i-y)'D^2V^\tau(y)(x_i-y) + o(h(y)^2) \tag{16}$$

où $h(y)$ est la résolution locale de la grille autour de y. On déduit que si l'on affine localement la grille en ajoutant ΔN nouveaux points, l'erreur d'interpolation locale décroît proportionnellement en ΔN fois une mesure de courbure de V.

Ceci se généralise au cas stochastique : par exemple, si l'on considère le schéma décrit à la section 3.4, l'erreur d'interpolation locale $e_u(x)$ dépend d'une moyenne des courbures de la FV aux points $z_j(x,u)$ pondérées par les poids $\rho_j(x,u)$.

Ces résultats se généralisent pour des approximations [Rus96, Rus97] sur grilles aléatoires ou à discrépance faible en fonction d'une mesure de la *variation* de V [Nie92].

Nous sommes alors capables d'évaluer l'effet d'un raffinement local de maillage sur l'erreur d'interpolation locale et, par le calcul de l'influence dans (14), de prédire l'amélioration de la qualité d'approximation de la FV.

Une procédure de raffinement de maillage efficace raffine la résolution aux endroits où le gain en qualité d'approximation est le plus désirable. Elle peut aussi retirer des points dans des régions de moindre impact sur l'erreur d'approximation, et les redisposer dans des zones plus pertinentes.

3.6 Allocation optimale de ressources ?

L'approche précédente présente deux inconvénients :

1. Elle s'intéresse à la qualité d'approximation de la FV mais pas à la politique déduite de cette approximation. Le gain obtenu en suivant une politique déduite d'une FV approchée peut être bien différent de cette fonction ; il s'agit pourtant de la véritable mesure de qualité.
2. Numériquement, la borne $\bar{\varepsilon}$ sur l'erreur ε calculée selon (14) est assez grossière car elle considère en chaque état la pire accumulation des erreurs d'interpolation. En pratique, ces erreurs se compensent et se neutralisent partiellement.

Le travail [19] (dont est extrait [6]) propose une réponse originale à ces problèmes ainsi qu'une procédure alternative d'allocation de ressources utilisant une fonction de croyance sur les paramètres du PDM approché.

3.6.1 Introduction

Nous modélisons un problème de prise de décision complexe en milieu incertain par un PDM fini. A cause des ressources numériques limitées, les paramètres du PDM (probabilités de transition et récompenses) sont incertains : on ne dispose que d'une fonction de croyance sur les valeurs possibles. Si nous choisissons les valeurs les plus probables, nous construisons un PDM que nous pouvons résoudre et en déduisons la politique correspondante. Cependant, à cause de l'incertitude sur les paramètres, cette politique risque de ne pas être celle qui maximise le gain du véritable (mais partiellement inconnu) problème. On peut néanmoins utiliser des techniques d'échantillonnage pour estimer la perte subie en utilisant cette politique au lieu de la politique optimale. De plus, si l'on suppose l'indépendance des paramètres (considérés comme des variables aléatoires), nous pouvons déduire la contribution de l'incertitude sur chaque paramètre à cette perte. Ainsi, nous pouvons prédire où l'ajout de ressources (permettant de réduire l'incertitude) améliore le plus vraisemblablement la performance espérée de la nouvelle politique.

Le problème de contrôle continu est approché à l'aide d'un opérateur de Bellman qui introduit une erreur d'interpolation locale que l'on modélise par une variable aléatoire. Cette approche est naturelle dans le cas de grilles aléatoires [Rus97]. Dans le cas d'une discrétisation de type de celle décrite au paragraphe 3.5, l'erreur d'interpolation locale (16) est majorée par une mesure de courbure de la FV, grandeur en partie inconnue, et qui peut avantageusement être modélisée par

une variable aléatoire. Cette étape de modélisation de l'erreur d'interpolation par des variables aléatoires n'est pas détaillée ici. On suppose simplement qu'elle découle d'une incertitude sur les paramètres due aux ressources numériques limitées du PDM discret.

L'outil développé dans l'article [19] est le calcul de sensibilité des quantités d'intérêt (la fonction valeur et la perte) par rapport aux paramètres du PDM.

3.6.2 Description du formalisme

Considérons un PDM sur un espace d'état fini X_N. L'opérateur de Bellman $\mathcal{T}_u$ est défini, pour toute fonction bornée $W : X_N \to \mathbb{R}$, par

$$\mathcal{T}_u W(x) := \gamma \sum_{y \in X_N} p(y|x,u) W(y) + r(x,u).$$

Les paramètres du PDM -les probabilités p et les récompenses r- sont partiellement inconnues : seule une fonction de croyance sur ces valeurs est connue. Notons $\alpha = \{\alpha_j\}$ l'ensemble des paramètres (les fonctions p et r) du PDM, et pour un PDM M^α défini par α, notons p^α, r^α, V^α et π^α respectivement ses fonctions probabilité, récompense, valeur et sa politique optimale.

Définissons le **gain** $J^\alpha(x,\pi)$ pour le PDM M^α :

$$J^\alpha(x,\pi) := \mathbb{E}\big[\sum_{t \geq 0} \gamma^t r^\alpha(x_t, \pi(x_t)) | x_0 = x; \pi \big].$$

Par définition, la fonction valeur de M^α est $V^\alpha(x) = J^\alpha(x, \pi^\alpha)$. Soit $\hat{\pi}$ une politique donnée (par exemple la politique optimale du PDM défini par les paramètres les plus probables). La **perte** $L^\alpha(x)$ résultant de l'usage de $\hat{\pi}$ au lieu de π^α dans M^α est

$$L^\alpha(x) := J^\alpha(x, \pi^\alpha) - J^\alpha(x, \hat{\pi}). \tag{17}$$

Si on définit une mesure de **gain global** $J^\alpha(\pi) := \sum_x \mu(x) J^\alpha(x,\pi)$ où l'on pondère les états par une distribution initiale $\mu(x)$ sur X_N, alors la **perte globale** est

$$L^\alpha := \sum_x \mu(x) L^\alpha(x). \tag{18}$$

Ainsi, connaissant la distribution de probabilité sur les paramètres α, on définit la **perte globale moyenne** $L := \mathbb{E}[L^\alpha]$ résultant de l'usage de $\hat{\pi}$ au lieu de la politique optimale.

Nous souhaitons établir l'impact sur cette perte L de l'incertitude sur chaque paramètre. Celle-ci est mesurée par un *coefficient caractéristique* (l'écart-type par exemple) de la loi (par exemple Gaussienne pour les récompenses, uniforme, bêta ou Dirichlet pour les probabilités). La contribution de l'incertitude sur un paramètre α_j à L est alors caractérisée par la dérivée partielle de L par rapport au coefficient caractéristique de la loi sur α_j.

Prenons l'exemple où la loi sur un paramètre α_j est une Gaussienne $\mathcal{N}(0, \sigma_j^2)$. Alors la contribution de cette incertitude à la perte L est estimée selon [19]

$$\frac{\partial L}{\partial \sigma_j} = \mathbb{E}\left[\frac{\partial L^\alpha}{\partial \alpha_j} \frac{\alpha_j}{\sigma_j} \right]. \tag{19}$$

L'outil développé est le calcul de sensibilité $\frac{\partial L^\alpha}{\partial \alpha_j}$ de la perte par rapport aux paramètres.

3.6.3 Calcul de sensibilité

Sensibilité de la fonction valeur. Soit un MDP M^α de paramètres α. Supprimons la référence à α pour simplifier les notations. Sous une politique optimale π, le PDM est une chaîne de Markov et la FV peut être considérée comme une fonction des variables probabilités et récompenses. Nous avons vu au paragraphe 3.3.1 que la dérivée partielle de $V(x)$ par rapport à la récompense $r(y, \pi(y))$ est l'influence $\mathcal{I}(y|x)$. Nous montrons que sa dérivée partielle par rapport à la probabilité $p(z|y, \pi(y))$ est $\gamma \mathcal{I}(y|x) V(z)$, ce qui permet d'écrire la différentielle de V.

Théorème 1 [19] *La différentielle de V est*

$$dV(x) = \sum_y \mathcal{I}(y|x)\left[\gamma \sum_z V(z) dp(z|y, \pi(y)) + dr(y, \pi(y))\right]. \tag{20}$$

Sensibilité de la perte. Pour une politique donnée $\hat{\pi}$, nous définissons la **perte immédiate** $l(x)$ résultant du choix d'un contrôle $\hat{\pi}(x)$ au lieu du contrôle optimal $\pi(x)$ en x :

$$l(x) := \mathcal{T}_{\pi(x)} V(x) - \mathcal{T}_{\hat{\pi}(x)} V(x).$$

La perte immédiate $l(x)$ peut être exprimée par une combinaison linéaire des récompenses $r(y, \pi(y))$ pondérées par les coefficient $l(y|x) := \gamma \mathcal{I}(y|p(\cdot|x, \pi(x))) - \gamma \mathcal{I}(y|p(\cdot|x, \hat{\pi}(x)))$ (où les influences sont calculées dans la chaîne de Markov déduite de la politique π) :

$$l(x) = r(x, \pi(x)) - r(x, \hat{\pi}(x)) + \sum_y l(y|x) r(y, \pi(y)).$$

La contribution de chaque paramètre à la perte immédiate et globale est donnée par le résultat suivant.

Théorème 2 [19] *Notons $d\mathcal{T}_u$ l'opérateur (formel) défini, pour tout $W : X_N \to \mathbb{R}$, par*

$$d\mathcal{T}_u W(x) := \gamma \sum_y W(y) dp(y|x, u) + dr(x, u).$$

Alors, pour tout x et u, la différentielle de la perte immédiate est

$$dl(x) = \sum_y l(y|x) d\mathcal{T}_{\pi(y)} V(y) + d\mathcal{T}_{\pi(x)} V(x) - d\mathcal{T}_{\hat{\pi}(x)} V(x).$$

En définissant les opérateurs $\mathcal{S}$ et $d\mathcal{S}$ définis, pour tout $W : X_N \to \mathbb{R}$, par

$$\begin{aligned} \mathcal{S}W(x) &:= \gamma \sum_y p(y|x, \hat{\pi}(x)) W(y) + l(x) \\ d\mathcal{S}W(x) &:= \gamma \sum_y dp(y|x, \hat{\pi}(x)) W(y) + dl(x), \end{aligned}$$

alors la perte $L(x)$ définie par (17) satisfait l'équation de Bellman $L = \mathcal{S}L$, et la différentielle de la perte $L(x)$ est

$$dL(x) = \sum_z \hat{\mathcal{I}}(z|x) d\mathcal{S}L(z). \tag{21}$$

(où l'influence $\hat{\mathcal{I}}$ est calculée dans la chaîne de Markov déduite de la politique $\hat{\pi}$)

En regroupant dans (21) les contribution de chaque paramètre, on déduit les dérivées partielles de la perte globale par rapport aux récompenses et aux probabilités de transition :

$$\begin{aligned} \frac{\partial L}{\partial r(x,u)} &= \hat{\mathcal{I}}(l(x|\cdot)|\mu(\cdot)) \mathbf{1}_{u=\pi(x)} + \hat{\mathcal{I}}(x|\mu(\cdot))(\mathbf{1}_{u=\pi(x)} - \mathbf{1}_{u=\hat{\pi}(x)}), \\ \frac{\partial L}{\partial p(y|x,u)} &= \gamma \hat{\mathcal{I}}(x|\mu(\cdot)) L(y) \mathbf{1}_{u=\hat{\pi}(x)} + \gamma V(y) \frac{\partial L}{\partial r(x,u)}. \end{aligned}$$

où $\mathbf{1}_b$ est la fonction booléenne qui vaut 1 si b est vrai, 0 sinon.

3.6.4 Guide pour la résolution numérique

Etant donnée la fonction de croyance sur le PDM, les étapes de résolution par échantillonnage sont les suivantes :

- Calcul de la politique optimale $\hat{\pi}$ du PDM le plus probable,
- K PDM $\{M^i\}_{1 \leq i \leq K}$ sont échantillonnés selon la fonction de croyance et sont résolus : fonction valeur V^i et politique π^i (l'exposant i faisant référence au PDM M^i),
- Calcul de la perte immédiate $l^i(x)$ et des influences $\hat{\mathcal{I}}(x|\mu(\cdot))$ et $\hat{\mathcal{I}}(l^i(x|\cdot)|\mu(\cdot))$ desquelles on déduit du théorème précédent la sensibilité de la perte par rapport aux récompenses : $\frac{\partial L^i}{\partial r^i(x,a)}$.
- Calcul de la perte $L^i(x)$ en résolvant l'équation de Bellman $L^i = \mathcal{S}L^i$, de laquelle on déduit la sensibilité de la perte par rapport aux probabilités : $\frac{\partial L^i}{\partial p^i(y|x,a)}$.
- Pour tout paramètre j, ces dérivées partielles $\frac{\partial L^i}{\partial \alpha_j}$ permettent d'établir l'estimateur (19).

Ainsi, la sensibilité de la FV (20) et de la perte (21) permet d'estimer les paramètres dont l'incertitude est la plus préjudiciable au gain espéré. Une procédure d'allocation de ressources efficace réduit l'incertitude (sur les paramètres) la plus dommageable.

Remarquons que l'outil de calcul de la sensibilité introduit ici peut servir à d'autres applications, par exemple pour définir une stratégie efficace d'exploration en A/R (le dilemme exploration-exploitation étant un enjeu majeur en A/R [Meu96]). Si les probabilités de transition sont initialement inconnues (problématique $\mathcal{P}1$), on peut construire un modèle bayésien [DFA99] à partir d'observations de transitions réalisées, en maintenant une fonction de croyance sur les probabilités de type distribution de Dirichlet. Ici, les ressources sont le nombre d'expériences réalisées. Le processus d'allocation de ressources détermine les régions de l'espace qui sont les plus urgent d'explorer, diminuant d'autant l'incertitude la plus préjudiciable.

4 Programmation Dynamique avec approximation

Le chapitre précédent fournit plusieurs pistes (dont certaines restent à approfondir) pour des procédures de raffinement de maillage efficaces. Cependant, ces méthodes de discrétisation ne peuvent que repousser le problème de l'explosion de la complexité de résolution -*la malédiction de la dimension*- lorsque la dimension de l'espace d'état croît. A titre d'exemple, les simulations numériques que nous avons réalisées sur un ordinateur ordinaire nous permettent de traiter efficacement des problèmes en dimension 3 sur un maillage uniforme, en dimension 4 et 5 avec les techniques développées au chapitre 3.3, et jusqu'en dimension 6 en les combinant avec des grilles aléatoires. Il est clair que ces techniques de discrétisation, même combinées à des procédures intelligentes de raffinement de maillage, sont vouées à l'échec en dimension supérieure. Nous sommes confronté au problème de représentation de fonctions en dimension élevée.

Dès lors, nous devons abandonner les méthodes de résolution exacte pour nous tourner vers une résolution approchée où les fonctions d'intérêt (fonction valeur ou politique) sont représentées à l'aide d'un jeu restreint de coefficients. Ce chapitre traite de l'approximation de la FV ; le chapitre 5 aborde le cas de politiques paramétrées.

Nous détaillons dans ce chapitre trois contributions :

1. L'illustration du problème en temps continu (section 4.1),
2. Des majorations d'erreur en norme L_1 et L_2 pour l'algorithme d'*itération sur les valeurs avec approximation* (section 4.2) en temps discret,
3. Des majorations d'erreur en norme L_2 pour l'algorithme d'*itération sur les politiques avec approximation* (section 4.3) en temps discret.

4.1 Le problème du temps continu

L'article [10] réalisé avec Leemon Baird et Andrew Moore illustre le deuxième problème mentionné dans l'introduction (et prédit dans la thèse [T]) lorsque l'on utilise des fonctions paramétrées pour minimiser le résidu de Bellman.

On considère un problème de contrôle déterministe (8) où le gain est (9) avec T un temps de sortie du domaine X. La fonction valeur $V(x) = \sup_{u.} J(x, u.)$ satisfait l'équation de HJB au sens des solutions de viscosité,

$$H(V, x) = 0,$$

où l'*Hamiltonien* (aussi appelé *résidu de Bellman*) est

$$H(V, x) := -\beta V(x) + \max_{u \in U} \left[\nabla V(x) \cdot f(x, u) + r(x, u) \right].$$

La fonction valeur est approchée par une représentation V_α à l'aide d'un réseau de neurones (les paramètres α sont les poids du réseau). La modification des poids suit une descente de gradient sur l'erreur :

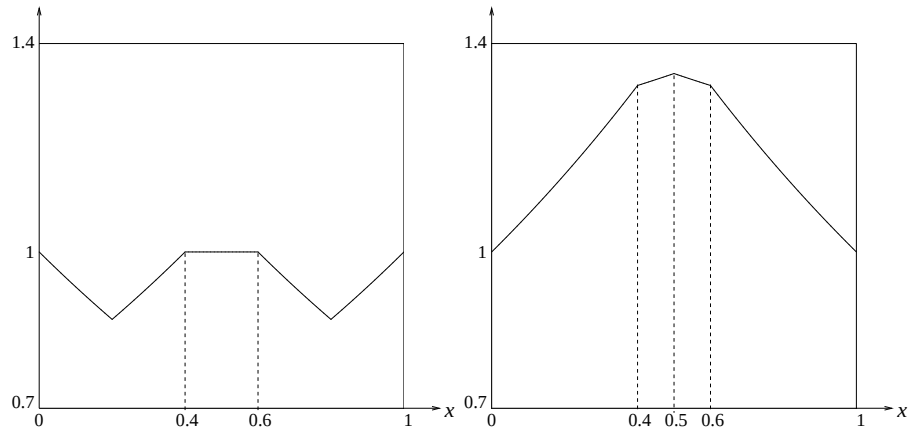
$$E(\alpha) := \frac{1}{2} \int [H(V_\alpha, x)]^2 dx. \tag{22}$$

Il est facile de déduire les équations d'évolution des poids $\partial E(\alpha)/\partial \alpha$ en fonction de l'architecture du réseau. La phase expérimentale consiste à tirer aléatoirement des états x_t (à l'intérieur du domaine ou sur le bord) et effectuer un pas de descente sur α dans la direction qui minimise $[H(V_\alpha, x_t)]^2$.

Illustrons sur un exemple simple le problème. Considérons un problème de contrôle en une dimension $x_t \in [0, 1]$ de dynamique d'état $\frac{dx}{dt} = u$ avec le contrôle $u_t \in \{-1, 1\}$ et de fonction récompense courante : $r(x) = \beta \mathbf{1}_{x \in [0.4, 0.6]}$ et récompense au bord $R(0) = 1$, $R(1) = 1$. Aux endroits où elle est différentiable, la FV satisfait l'équation de HJB :

$$-\beta V(x) + |V'(x)| + r(x) = 0. \tag{23}$$

Ici, on peut calculer la FV (représentée sur la Figure 10a pour $\beta = \ln 2$) : $V(x) = e^{-\beta x} \mathbf{1}_{x \leq 0.2} + e^{-\beta(0.4-x)} \mathbf{1}_{x \in [0.2, 0.4]} + \mathbf{1}_{x \in [0.4, 0.6]} + e^{-\beta(x-0.6)} \mathbf{1}_{x \in [0.6, 0.8]} + e^{-\beta(1-x)} \mathbf{1}_{x \geq 0.8}$.



(a) La fonction valeur V.

(b) Une solution généralisée.

FIG. 10 – La fonction valeur et une solution généralisée de HJB.

La Figure 11a représente la fonction V_α représentée par le réseau de neurones (avec une couche cachée de 100 unités) obtenue par l'algorithme de gradient sur l'erreur (22). Le résidu de Bellman (Figure 11b) résultant est proche de 0 ce qui signifie que l'algorithme a bien fonctionné en minimisant l'erreur. Cependant, la fonction apprise par le réseau diffère complètement de la fonction valeur (ainsi que le contrôle résultant, ici donné par le signe de sa dérivée). En réalité, il s'agit d'une approximation de la fonction $V_g(x) = e^{\beta x}\mathbf{1}_{x\leq 0.4} + [1 + (1 - e^{-0.4\beta})e^{\beta x}]\mathbf{1}_{x\in[0.4,0.5]} + [1 + (1 - e^{-0.4\beta})e^{-\beta(x-1)}]\mathbf{1}_{x\in[0.5,0.6]} + e^{-\beta(x-1)}\mathbf{1}_{x\geq 0.6}$ (représentée sur la figure 10b) qui est une solution généralisée (c'est à dire différentiable presque partout) de (23).

Ce problème provient de la non-unicité des solutions généralisées des équations de HJB. Etant donné que le résidu de Bellman est nul presque partout pour la FV mais aussi pour toute autre solution généralisée, le problème de minimisation de l'erreur (22) admet de nombreux (et même une infinité ici) minima *globaux* : le problème est mal posé.

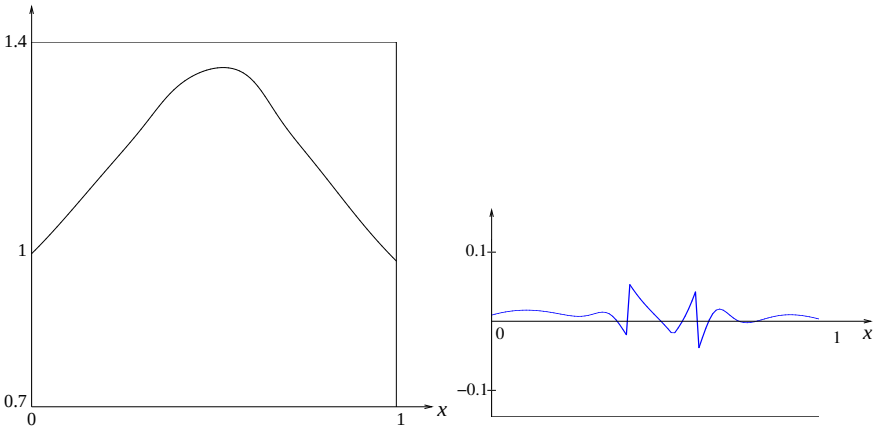
Il ne s'agit pas que d'un problème théorique apparaissant uniquement dans des cas bien choisis. Pour le problème de la voiture sur la colline, une fonction obtenue est représentée sur la Figure 12a. Celle-ci est substantiellement différente de la FV (Figure 2a) alors que le résidu de Bellman (Figure 12b) est assez proche de 0. La politique qui s'en déduit est très différente de l'optimale.

Pour tenter de surmonter ce problème, citons trois approches permettant de retrouver l'unicité de solution

1. *Introduction de stochasticité.* Si nous rendons le problème stochastique en ajoutant un terme de diffusion σ selon (1) et considérons une hypothèse d'ellipticité uniforme sur $\sigma\sigma'$ alors la FV est l'unique solution régulière de HJB (4) dans X avec la condition au bord $V = R$ [FS93, Kry95].
2. *Discrétisation temporelle.* Une discrétisation temporelle à un pas $\tau > 0$ donne un PDM dont la fonction valeur associée V^τ est l'unique solution de l'équation de PD $V^\tau(x) = \sup_u \mathcal{T}_u V^\tau(x)$, où l'opérateur $\mathcal{T}_u$ est défini par (15).
3. *Itération sur les politiques.* La non-unicité de solution de HJB vient de sa non-linéarité. Pour une politique π donnée, l'EDP linéaire

$$-\beta W(x) + \nabla W(x) \cdot f(x, \pi(x)) + r(x, \pi(x)) = 0$$

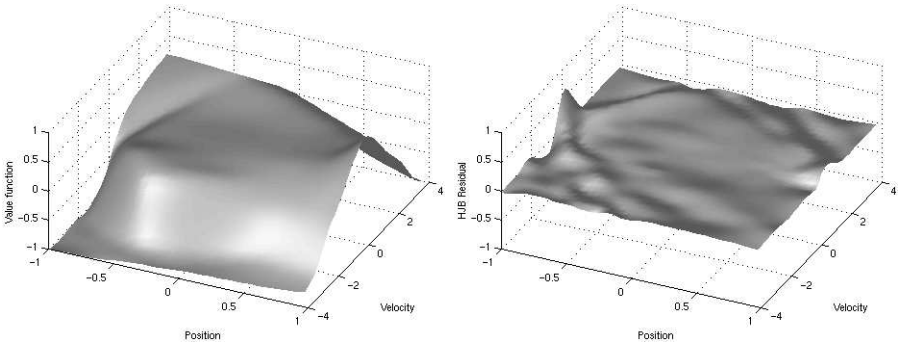
admet une unique solution V^π. Une procédure d'*itération sur les politiques* part d'une politique initiale π_0, calcule V^{π_0} puis construit une nouvelle politique π_1 vérifiant en tout x,



(a) La fonction V_α représentée par le réseau.

(b) Résidu de Bellman pour la fonction V_α.

FIG. 11 – La fonction paramétrée V_α et le résidu de Bellman associé.



(a) Fonction représentée.

(b) Résidu de Bellman.

FIG. 12 – Fonction V_α représentée par le réseau de neurones et résidu associé, pour la voiture sur la colline.

$\pi_1(x) \in \arg\max_u [\nabla V^{\pi_0}(x) \cdot f(x,u) + r(x,u)]$. La répétition de cette procédure génère des V^{π_k} qui convergent vers V [FR75, LM80].

Malgré l'unicité de solution obtenue dans ces trois approches, cela ne garantie en rien l'unicité de solution dans l'espace des fonctions V_α paramétrées, ni sa possible résolution par une méthode de gradient (sujette au problème des minima locaux).

4.2 Itération sur les valeurs avec approximation

Cette section et la suivante présentent deux algorithmes très populaires [BT96] : *itération sur les valeurs avec approximation* (*Approximate Value Iteration*) et *itération sur les politiques avec approximation* (*Approximate Policy Iteration*). Notre contribution [5, 16] est l'établissement de majoration d'erreur en normes L_1 et L_2, résultats généralisant les majorations habituelles en norme L_∞. Nous expliquons la pertinence applicative de ces résultats.

Nous considérons ici le cas où le temps et l'espace sont des variables discrètes. Le cas d'un espace continu se généralise aisément [2].

Précisons les notations. L'espace d'état (de taille N) est noté X_N. Pour une politique $\pi : X_N \to U$, notons P^π la matrice de transition, d'éléments $P^\pi(x,y) = p(y|x,\pi(x))$ et r^π le vecteur récompense, de composantes $r^\pi(x) = r(x,\pi(x))$. L'opérateur de Bellman $\mathcal{T}^\pi$ est défini, pour $W \in \mathbb{R}^N$, par $\mathcal{T}^\pi W = r^\pi + \gamma P^\pi W$. La fonction valeur V^π associée à une politique π est solution de l'équation de Bellman $V^\pi = \mathcal{T}^\pi V^\pi$.

La fonction valeur optimale, notée V^*, est le gain pour une politique optimale $\pi^* : V^* := V^{\pi^*} = \sup_\pi V^\pi$. En définissant l'opérateur de PD $\mathcal{T}$, pour $W \in \mathbb{R}^N$, par

$$\mathcal{T}W(x) := \max_{u \in U} \big[r(x,u) + \gamma \sum_y p(y|x,u) W(y) \big], \tag{24}$$

alors V^* est solution de l'équation de PD $V^* = \mathcal{T}V^*$. On appelle π une **politique déduite** de $W \in \mathbb{R}^N$ si pour tout $x \in X_N$,

$$\pi(x) \in \arg\max_{u \in U} \big[r(x,u) + \gamma \sum_y p(y|x,u) W(y) \big]. \tag{25}$$

Enfin, pour μ une distribution sur X_N, nous définissons les normes pondérées $L_1 : ||W||_{1,\mu} := \sum_x \mu(x)|W(x)|$ et $L_2 : ||W||_{2,\mu} := \big[\sum_x \mu(x)|W(x)|^2\big]^{1/2}$.

4.2.1 L'algorithme IVA

On considère la résolution d'un PDM en utilisant des représentations approchées V_n de la FV. L'algorithme d'*itération sur les valeurs avec approximation* (IVA) est défini par l'itération

$$V_{n+1} = \mathcal{A}\mathcal{T}V_n \tag{26}$$

où $\mathcal{T}$ est l'opérateur de PD (24) et $\mathcal{A}$ un *opérateur d'approximation*, ou de manière équivalente un algorithme *d'apprentissage supervisé* (AS) [HTF01].

Une implémentation simple est la suivante : supposons que l'on considère des approximations dans une classe de fonctions Φ prédéfinie. A l'étape n, on échantillonne des états $(x_k)_{1 \le k \le K}$ selon une distribution μ sur X_N, on calcule les valeurs itérées $v_k = \mathcal{T}V_n(x_k)$ et on appelle un algorithme d'AS avec les données (x_k, v_k), qui retourne une nouvelle approximation $V_{n+1} \in \Phi$, par exemple solution du problème de minimisation quadratique $V_{n+1} \in \arg\min_{W \in \Phi} \frac{1}{K} \sum_{k=1}^K (W(x_k) - v_k)^2$.

De nombreuses autres implémentations sont possibles, notamment en A/R (problématique $\mathcal{P}1$) lorsque l'on ne dispose pas de modèle des probabilités de transition [16].

Expérimentalement, on observe que la performance des politiques π_n déduites des approximations V_n s'améliore au début puis stagne. Cet algorithme ne converge pas nécessairement, néanmoins on peut estimer sa performance asymptotique en fonction des *erreurs d'approximation* $\varepsilon_n := \mathcal{T}V_n - \mathcal{A}\mathcal{T}V_n$ de l'algorithme d'AS.

Si les erreurs d'approximation sont *uniformément bornées* $||\varepsilon_n||_\infty \leq \varepsilon_\infty$, une majoration sur l'écart entre la performance des politiques π_n déduites des approximations V_n et la politique optimale est [BT96] :

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\varepsilon_\infty. \tag{27}$$

Cette borne présente l'inconvénient majeur d'utiliser une majoration *uniforme* sur ε_n, qui est difficile à fournir (et peu précise lorsque l'espace d'état est grand).

De plus, en général, un algorithme d'AS fournit une approximation (représentation compacte) d'une fonction f donnée en minimisant une norme L_1 ou L_2 de l'erreur [DeV97] (il y a cependant quelques exceptions utilisant la norme L_∞ [Gor95, GKP01]). L'approximation est dite *linéaire* lorsqu'il s'agit d'une projection sur une famille fixée de fonctions (par exemple les régressions linéaires, les décompositions sur des bases de polynômes, de cosinus). L'approximation *non-linéaire* (par exemple lorsqu'on choisit l'espace en fonction des régularités de f) est particulièrement efficace lorsque f est régulière par morceaux (voir les décompositions adaptatives sur bases d'ondelettes [Mal97]).

En *apprentissage statistique* [HTF01], d'autres algorithmes d'AS sont les *réseaux de neurones* [Hay94], les *régressions linéaires locales* et *méthodes à noyau* [ASM97], les *Support-Vectors* et *Reproducing Kernels* [VGS97]. Dans toutes ces méthodes, l'algorithme d'AS utilise une norme L_1 ou L_2.

Or, il est désirable d'exprimer la performance de l'algorithme IVA dans la même norme que celle utilisée par l'algorithme d'AS afin de garantir l'utilité et la finesse de la majoration.

4.2.2 Majorations en normes L_1 et L_2

Notre contribution généralise la borne (27) à des normes L_1 et L_2.

Théorème 3 [16] *Soit μ (vecteur ligne) une distribution sur X_N. Pour tout $n \geq 0$ et $0 \leq k \leq n-1$, définissons les matrices stochastiques*

$$S_{n,k} = \frac{1-\gamma}{2}(I - \gamma P^{\pi_n})^{-1}\left[(P^{\pi^*})^{n-k} + (\prod_{i=k+1}^{n} P^{\pi_i})\right].$$

Alors $\mu_{n,k} := \mu S_{n,k}$ est une distribution sur X_N, et pour $i = 1$ ou 2,

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_{i,\mu}^i \leq \frac{(2\gamma)^i}{(1-\gamma)^{2i-1}} \limsup_{n\to\infty} \sum_{k=0}^{n-1} \gamma^{n-1-k} ||\varepsilon_k||_{i,\mu_{n,k}}^i \tag{28}$$

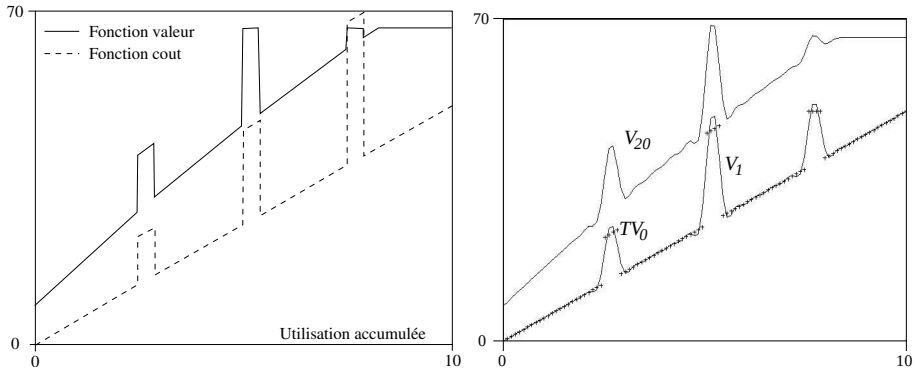
Ce résultat généralise la borne (27). En effet, pour tout x, si l'on considère la distribution $\mu = \mathbf{1}_x$ (Dirac en x), puisque $||\varepsilon_k||_{1,\mu_{n,k}} \leq ||\varepsilon_k||_\infty \leq \varepsilon_\infty$, alors $\limsup_{n\to\infty}(V^* - V^{\pi_n})(x) \leq \frac{2\gamma}{1-\gamma}\sum_{k=0}^{\infty}\gamma^k\varepsilon_\infty = \frac{2\gamma}{(1-\gamma)^2}\varepsilon_\infty$, et (27) s'en suit.

Nous pouvons questionner l'intérêt pratique de cette majoration, puisque qu'elle utilise les distributions $\mu_{n,k}$ qui dépendent de la politique optimale π^*, qui est inconnue. Cependant, sous une certaine régularité de la répartition des états futurs, nous justifions la pertinence de cette borne.

Définissons la *constante C de régularité de la répartition des d'états futurs (avec actualisation) par rapport à la distribution initiale μ*. Il s'agit de la plus petite constante telle que pour toute séquence de politiques $\pi_1, \pi_2, \ldots$ la somme (actualisées) des lois sur x_m sachant que $x_0 \sim \mu$ est majorée par C fois μ :

$$(1-\gamma)^2 \sum_{m=1}^{\infty} m\gamma^{m-1}\Pr\{x_m = y | x_0 \sim \mu,\, x_{i+1} \sim p(\cdot|x_i, \pi_i(x_i))\} \leq C\mu(y).$$

Le principal résultat de [16] est le suivant.



(a) Fonctions coût et valeur. (b) $\mathcal{T}V_0$ (les croix), V_1 et V_{20}.

FIG. 13 – Problème de remplacement optimal.

Théorème 4 [16] *Soit μ une distribution sur X_N. Soit $i = 1$ ou 2. Supposons que les erreurs d'approximation vérifient $||\varepsilon_n||_{i,\mu} \leq \varepsilon_i$. Alors*

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_{i,\mu} \leq \frac{2\gamma}{(1-\gamma)^2} C^{1/i} \varepsilon_i. \tag{29}$$

Donnons quelque intuition sur la constante C lorsque la distribution est uniforme $\mu = (\frac{1}{N} \ldots \frac{1}{N})$.

- La plus grande valeur de C est obtenue lorsque un état particulier (disons x_1) est successeur (pour une certaine politique π) de tous les états avec une probabilité 1. Alors on a $\Pr\{x_m = y\} \leq \Pr\{x_m = x_1\} = 1$, donc $C = N$. Il s'agit du pire cas. Dans ce cas, (29) se déduit de la borne L_∞ (27).
- La plus petite valeur de C correspond à un PDM où toutes les probabilités de transition sont uniformes : $\Pr\{x_m = y\} = 1/N$, donc $C = 1$.

Ainsi la constante $C \in [1, N]$ exprime la régularité de la fonction de répartition des d'états futurs (avec actualisation) par rapport à la distribution initiale μ.

Nous montrons [16] que la majoration (29) peut être beaucoup plus fine que celle en norme L_∞ (27) sur un PDM "chaînon en ligne" composé de N états. Dans cet exemple, ainsi que dans les cas (dont celui illustré au paragraphe suivant) où l'on discrétise un problème continu en espace (mais discret en temps), la constante C est indépendante de N.

Pour le PDM "chaînon en ligne", la finesse de la majoration (29) est d'ordre $O(N^{-1})$ pour la norme L_1, d'ordre $O(N^{-1/2})$ pour la norme L_2 alors que celle en norme L_∞ (27) est d'ordre $O(1)$ seulement. Le gain en précision est considérable lorsque N est grand.

4.2.3 Problème de remplacement optimal [16]

Ce problème [Rus96] modélise l'état d'usure d'un bien par une variable unidimensionnelle $x_t \in \mathbb{R}_+$ (par exemple le compteur kilométrique d'une voiture). $x_t = 0$ signifie un bien neuf. A chaque instant discret t, il y a deux décisions possibles : soit garder le bien, soit le remplacer (ce qui entraîne un coût supplémentaire de vente puis du rachat d'un nouveau bien). Nous considérons des densités de transition qui suivent une loi exponentielle. On choisit une fonction coût d'entretien qui est la somme d'une fonction lentement croissante (coût de maintenance) et d'une fonction créneau périodique (coûts de révisions par exemple). Celle-ci ainsi que la fonction valeur (calculée numériquement) sont indiquées sur la Figure 13a.

	$\|\|\mathcal{T}V_n - \mathcal{A}\mathcal{T}V_n\|\|_\infty$	$C\|\|\mathcal{T}V_n - \mathcal{A}\mathcal{T}V_n\|\|_1$	$\sqrt{C}\|\|\mathcal{T}V_n - \mathcal{A}\mathcal{T}V_n\|\|_2$
$N = 200$	12.4	0.367	1.16
$N = 2000$	12.4	0.0552	0.897

TAB. 1 – Comparaison des erreurs d'approximation en norme L_∞, L_1 et L_2.

On considère des approximations linéaires sur une base tronquée de cosinus :

$$\Phi := \left\{ V_n(x) = \sum_{j=1}^{20} \alpha_j \cos(j\pi \frac{x}{x_{\max}}) \right\}.$$

Le problème est discrétisé sur l'intervalle $X = [0, x_{\max}]$ par une grille uniforme de N points. La Figure 13b représente la première itération : les valeurs itérées $\mathcal{T}V_0$ (pour une valeur initiale $V_0 = 0$) représentées par les croix, la meilleure approximation $V_1 \in \Phi$ de $\mathcal{T}V_0$ (au sens des moindres carrés), ainsi que la FV approchée au bout de 20 itérations (lorsqu'il n'y a plus d'amélioration constatée).

Nous montrons que la constante C est indépendante de N et vaut $C = 6$ pour les valeurs numériques choisies dans [16].

Le tableau 1 compare le majorant des inégalités (27) et (29). Les majorations L_1 et L_2 sont meilleures que celle L_∞, et décroissent quand N augmente alors que ce n'est pas le cas de la majoration en norme L_∞. En effet, puisque la fonction coût est discontinue (et les approximations $V_n \in \Phi$ continues), l'erreur d'approximation L_∞ ne peut être inférieure à la demie valeur de la plus grande discontinuité.

Remarquons toutefois que la majoration $L_\infty(27)$ est plus informative que celles en L_1 ou L_2 (29) puisqu'elle fournit une majoration uniforme sur la perte asymptotique. Cependant, il est assez naturel de choisir comme critère de comparaison la perte globale (18), introduite dans la section 3.6, qui se définit par la norme $L_{1,\mu}$.

Nous voyons alors la pleine pertinence de la majoration (29) pour des discrétisations de problèmes continus en espace, surtout lorsque la résolution de la grille est fine.

4.3 Itération sur les politiques avec approximation

L'algorithme d'*itération sur les politiques avec approximation* (IPA) [BT96] est défini itérativement par les deux étapes suivantes :

- *Evaluation avec approximation de la politique* : pour une politique π_n, on calcule une approximation V_n de la fonction valeur V^{π_n}.
- *Amélioration de la politique* : on génère une nouvelle politique π_{n+1} déduite de V_n (au sens (25)).

Un résultat classique [BT96] donne une majoration sur la perte $V^* - V^{\pi_n}$ résultant de l'utilisation de la politique π_n au lieu de la politique optimale, en fonction des *erreurs d'approximation* $V_n - V^{\pi_n}$:

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{n\to\infty} ||V_n - V^{\pi_n}||_\infty. \tag{30}$$

Pour les mêmes raisons qu'au paragraphe précédent, cette majoration L_∞ est peu utile concrètement.

4.3.1 Majorations en norme L_2

Le résultat suivant établit des majorations en norme L_2 sur la perte $V^* - V^{\pi_n}$ en fonction des erreurs d'approximation $V_n - V^{\pi_n}$ et des *résidus de Bellman* $V_n - T^{\pi_n}V_n$:

Théorème 5 [5] *Définissons les matrices stochastiques : pour $n \geq 1$,*

$$\begin{aligned} S_n &= \frac{(1-\gamma)^2}{2}(I-\gamma P^{\pi^*})^{-1}[P^{\pi_{n+1}}(I-\gamma P^{\pi_{n+1}})^{-1} + P^{\pi^*}(I-\gamma P^{\pi_n})^{-1}], \\ \widetilde{S}_n &= \frac{(1-\gamma)^2}{2}(I-\gamma P^{\pi^*})^{-1}[P^{\pi_{n+1}}(I-\gamma P^{\pi_{n+1}})^{-1}(I+\gamma P^{\pi_n}) + P^{\pi^*}]. \end{aligned}$$

Soit μ une distribution sur X_N. Alors $\mu_n := \mu S_n$ et $\widetilde{\mu}_n := \mu \widetilde{S}_n$ sont des distributions sur X_N et l'on a

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_{2,\mu} \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{n\to\infty} ||V_n - T^{\pi_n} V_n||_{2,\mu_n} \tag{31}$$

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_{2,\mu} \leq \frac{2\gamma}{(1-\gamma)^2} \limsup_{n\to\infty} ||V_n - V^{\pi_n}||_{2,\widetilde{\mu}_n} \tag{32}$$

Remarquons déjà que (30) se déduit de (32). De plus, pour tout état x, $S_n(x,y)$ (respectivement $\widetilde{S}_n(x,y)$) est une majoration de la contribution du résidu de Bellman $V_n - T^{\pi_n}V_n$ (resp. l'erreur d'approximation $V_n - V^{\pi_n}$) en y à la perte $V^* - V^{\pi_n}$ en x. Cette information peut être exploitée pour définir les régions où il est utile de réduire le résidu de Bellman pour diminuer la perte globale. En approximation non-linéaire, une procédure d'allocation de ressources sélectionnant des fonctions d'un dictionnaire (de type *Matching Pursuit* [DMA97]) pourrait s'en inspirer.

Remarquons que les erreurs d'approximation $V_n - V^{\pi_n}$ définies ici sont différentes de celles $\mathcal{T}V_n - \mathcal{A}\mathcal{T}V_n$ considérées précédemment pour l'algorithme IVA : l'approximation de V^{π_n} (inconnue) est une opération difficile en général, alors que celle considérée pour IVA est "élémentaire" (simple appel à un algorithme d'AS).

Dans le cas d'une approximation linéaire, nous pouvons établir une majoration de l'erreur d'approximation et du résidu de Bellman en fonction de la capacité représentationnelle de l'architecture d'approximation ε_Φ, défini plus bas (35).

4.3.2 Approximation linéaire

Nous considérons une classe de fonctions $\Phi = \{V_\alpha(x) := \sum_{k=1}^K \alpha_k \phi_k(x)\}_{\alpha \in \mathbb{R}^K}$ paramétrées linéairement par un paramètre $\alpha \in \mathbb{R}^K$, où les $(\phi_k)_{1\leq k\leq K}$ sont des fonctions données.

La phase d'évaluation d'une politique π détermine le paramètre α pour que V_α soit une "bonne" approximation de V^π. Choisissons une distribution μ_π sur X_N. Nous détaillons deux approches [Sch02] de type *méthodes de projection* [Jud98] permettant d'établir la valeur du paramètre :

- α réalise le minimum de la norme du résidu de Bellman

$$\min_\alpha ||V_\alpha - \mathcal{T}^\pi V_\alpha||^2_{2,\mu_\pi}$$

  Remarquons que la fonction minimisée est quadratique en α. La solution, dite du **résidu quadratique (RQ)**, satisfait donc un système linéaire de taille K :

$$A\alpha = b, \text{ avec } \left\{ \begin{array}{rcl} A_{ij} &=& \langle \phi_i - \gamma P^\pi \phi_i, \phi_j - \gamma P^\pi \phi_j \rangle_{\mu_\pi}, \text{ pour } 1 \leq i,j \leq K \\ b_i &=& \langle \phi_i - \gamma P^\pi \phi_i, r^\pi \rangle_{\mu_\pi} \end{array} \right. \tag{33}$$

  (où $\langle f, g\rangle_\mu := \sum_{x\in X_N} \mu(x) f(x) g(x)$). Ce système possède une unique solution lorsque la famille $\{\phi_k\}_{1\leq k\leq K}$ est libre et $\mu_\pi > 0$.
- α est tel que V_α est le point fixe de l'opérateur composé $\Pi_{\mu_\pi}\mathcal{T}^\pi$ où $\mathcal{T}^\pi$ est l'opérateur de Bellman et Π_{μ_π} est l'opérateur de projection sur Φ (selon la norme $||\cdot||_{2,\mu_\pi}$). Cette solution, dite des **différences temporelles (DT)**, satisfait aussi un système linéaire de taille K :

$$A\alpha = b, \text{ avec } \left\{ \begin{array}{rcl} A_{ij} &=& \langle \phi_i, \phi_j - \gamma P^\pi \phi_j \rangle_{\mu_\pi}, \text{ pour } 1 \leq i,j \leq K \\ b_i &=& \langle \phi_i, r^\pi \rangle_{\mu_\pi} \end{array} \right. \tag{34}$$

  Ce système n'est pas nécessairement inversible.

Notons ε_Φ la **capacité de représentation des fonctions valeurs par l'architecture d'approximation Φ** :

$$\varepsilon_\Phi := \max_\pi \inf_\alpha ||V_\alpha - V^\pi||_{2,\mu_\pi}. \tag{35}$$

Nous souhaitons exprimer les majorants de (31) et (32) en fonction de ε_Φ.

Régularité uniforme des probabilités de transition. Pour l'algorithme IVA nous avons défini une constante de répartition de présence des états futurs, afin d'établir des majorations d'erreur sur la perte $V^* - V^{\pi_n}$ en norme L_1 et L_2. Ici, nous énonçons un résultat de majoration plus fort (majoration uniforme sur la perte) en contrepartie duquel nous utilisons une *constante $\widetilde{C}$ de majoration uniforme sur les probabilités de transition*, i.e. qui vérifie, pour μ une distribution sur X_N, pour tous $x, y \in X_N$, $u \in U$,

$$p(y|x,u) \leq \widetilde{C}\mu(y). \tag{36}$$

Une petite constante $\widetilde{C}$ signifie que la masse de la distribution de présence des états après une itération (pour tout contrôle possible) ne s'accumule pas sur quelques états particuliers. En particulier, les PDM déterministes fournissent la plus grande valeur de $\widetilde{C}$. Si l'espace est continu et si les lois de transition admettent une densité par rapport à μ, alors $\widetilde{C}$ vaut la valeur maximale de ces densités.

Solution du résidu quadratique. Si à chaque étape n, nous choisissons l'approximation $V_n = V_{\alpha_n}$ avec α_n solution de (33), alors la performance de l'algorithme IPA est majorée selon le résultat suivant.

Théorème 6 [5] *Supposons que la distribution utilisée dans (35) est $\mu_\pi = \mu$, alors*

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\sqrt{\widetilde{C}}\left(1+\gamma\sqrt{\widetilde{C}}\right)\varepsilon_\Phi.$$

D'autres résultats lorsque la distribution μ_π (introduite dans la définition de ε_Φ) est différente de μ (définissant la constante $\widetilde{C}$) sont indiqués dans [5].

Solution des différences temporelles. Pour s'assurer que le système (34) possède une solution unique, nous définissons ε_Φ en utilisant la distribution stationnaire $\overline{\mu_\pi}$ de la chaîne de Markov induite par la politique π (une telle distribution existe, par exemple sous l'hypothèse que la chaîne de Markov correspondante est irréductible et apériodique [Put94]), que nous supposons minorée.

En choisissant, à chaque étape n, le paramètre α_n solution de (34), la performance de l'algorithme IPA est majorée selon le résultat suivant.

Théorème 7 [5] *Supposons que les distributions stationnaires $\overline{\mu_\pi}$ soient minorées par $\frac{1}{\kappa}\mu$ (avec une constante $\kappa > 0$), alors*

$$\limsup_{n\to\infty} ||V^* - V^{\pi_n}||_\infty \leq \frac{2\gamma}{(1-\gamma)^3}\sqrt{\kappa\widetilde{C}}\,\varepsilon_\Phi.$$

Apprentissage par renforcement. Mentionnons simplement qu'il existe des algorithmes d'A/R inspirés de la méthode IPA qui répondent ainsi aux deux problématiques $\mathcal{P}1$ et $\mathcal{P}2$ [2]. La phase d'évaluation d'une politique π se fait en estimant la matrice A et le vecteur b à partir de données recueillies par l'observation des transitions [BB96, Boy99], puis en résolvant le système (33) ou (34) avec ces estimateurs. Par exemple, des données $(x_l, y_l, r_l)_{1\leq l\leq L}$ qui résultent des transitions :

état x_l, contrôle $\pi(x_l)$ vers un état $y_l \sim p(\cdot|x_l, \pi(x_l))$ avec une récompense r_l, donnent l'estimateur de A et b du système (34) :

$$\begin{aligned} \widehat{A}_{ij} &= \frac{1}{L}\sum_{l=1}^{L} \phi_i(x_l)\left[\phi_j(x_l) - \gamma\phi_j(y_l)\right], \\ \widehat{b}_i &= \frac{1}{L}\sum_{l=1}^{L} \phi_i(x_l) r_l. \end{aligned}$$

4.4 Minimisation du résidu de Bellman

Enfin, pour clore ce chapitre, nous illustrons comment les idées précédentes peuvent s'appliquer à d'autres méthodes de programmation dynamique avec approximation. Il semble que toute l'analyse L_∞ usuelle en PD se généralise en normes L_1 et L_2, et fera l'objet de travaux futurs.

Par exemple, considérons l'approximation de la FV (qui satisfait l'équation de PD $\mathcal{T}V = V$, ou $\mathcal{T}$ est défini par (24)) par la résolution du problème de minimisation du résidu de Bellman [Bai95] :

$$\inf_{\alpha} ||\mathcal{T}V_\alpha - V_\alpha||. \tag{37}$$

Un résultat de majoration en norme L_∞ de la performance d'une politique π_α déduite de V_α par le résidu de Bellman de V_α est [WB93] :

$$||V^* - V^{\pi_\alpha}||_\infty \leq \frac{2}{1-\gamma}||\mathcal{T}V_\alpha - V_\alpha||_\infty. \tag{38}$$

Ici, à nouveau, le problème de minimisation (37) s'exprime habituellement en norme L_1 ou L_2 (par exemple si on utilise une méthode de gradient sur l'erreur quadratique du résidu de Bellman $||\mathcal{T}V_\alpha - V_\alpha||^2_{2,\mu}$), rendant la borne (38) inexploitable. De la majoration (4.1) dans [16] appliquée à la politique optimale π^* et en utilisant la relation $(V^{\pi_\alpha} - V_\alpha) = (I - \gamma P^{\pi_\alpha})^{-1}(\mathcal{T}V_\alpha - V_\alpha)$, il vient

$$V^* - V^{\pi_\alpha} \leq \left[(I - \gamma P^{\pi^*})^{-1} - (I - \gamma P^{\pi_\alpha})^{-1}\right](\mathcal{T}V_\alpha - V_\alpha),$$

composante par composante. Le résultat suivant en découle.

Théorème 8 *Soit μ une distribution sur X_N. Définissons la matrice stochastique*

$$S_\alpha = \frac{1-\gamma}{2}\left[(I - \gamma P^{\pi^*})^{-1} + (I - \gamma P^{\pi_\alpha})^{-1}\right]$$

Alors $\mu_\alpha := \mu S_\alpha$ est une distribution sur X_N, et pour $i = 1$ ou 2,

$$||V^* - V^{\pi_\alpha}||_{i,\mu} \leq \frac{2}{1-\gamma}||\mathcal{T}V_\alpha - V_\alpha||_{i,\mu_\alpha}$$

Cette majoration est plus fine que (38). Comme dans la section 4.2, en définissant une constante de régularité de la répartition des états futurs, on peut déduire des majorations d'erreur sous une norme L_1 ou L_2 identique à celle utilisée pour le problème de minimisation (37), garantissant ainsi son utilité applicative.

5 Analyse de sensibilité par rapport à des paramètres de contrôle

Dans ce chapitre, nous cherchons à résoudre le problème de contrôle optimal stochastique en paramétrant directement la politique et en utilisant une méthode de gradient stochastique pour résoudre le problème d'optimisation paramétrique.

Le travail [1] réalisé avec Emmanuel Gobet propose des estimateurs de la sensibilité du gain par rapport aux paramètres du contrôle et les évalue numériquement.

Pour une politique (ou un contrôle en boucle ouverte) paramétrée par α (supposé unidimensionnel pour simplifier les notations), notons la dynamique d'état

$$X_t^\alpha = x + \int_0^t f(s, X_s^\alpha, \alpha) ds + \sum_{j=1}^q \int_0^t \sigma_j(s, X_s^\alpha, \alpha) dW_s^j, \tag{39}$$

(où σ_j est la j^e colonne de σ) et considérons ici un gain à horizon temporel fixé T : $\mathbb{E}\left[r(X_T^\alpha)\right]$.

L'extension à une fonctionnelle plus générale de type (2), où T est fixé, se déduit aisément des résultats qui suivent par linéarité. Le cas où T est un temps de sortie n'est pas abordé ici.

Nous posons des hypothèses de régularité (non détaillées ici) sur les coefficients f et σ.

Une première approche pour estimer la sensibilité d'espérance $\partial_\alpha \mathbb{E}[r(X_T^\alpha)]$ est la *méthode de resimulation* [LP94], qui évalue le gain pour des valeurs proches de α et estime la dérivée par un quotient aux différences finies. Cette méthode présente l'inconvénient d'être coûteuse lorsque α est multi-dimensionnel ; de plus l'estimateur fourni est biaisé.

Une deuxième approche, qui s'applique lorsque le renforcement r est dérivable, considère une *sensibilité trajectorielle.* Il s'agit de passer la dérivation sous l'espérance en introduisant la dérivée trajectorielle de X_t par rapport à α :

$$\partial_\alpha X_t^\alpha = \int_0^t (\partial_\alpha f_s + \nabla_x f_s \partial_\alpha X_s^\alpha) ds + \sum_{j=1}^q \int_0^t (\partial_\alpha \sigma_{j,s} + \nabla_x \sigma_{j,s} \partial_\alpha X_s^\alpha) dW_s^j,$$

en utilisant les notations simplifiées $f_s = f(s, X_s^\alpha, \alpha)$ et $\sigma_{j,s} = \sigma_j(s, X_s^\alpha, \alpha)$. L'estimateur résultant est le suivant :

Proposition 2 [YK91] *Si r est dérivable, alors*

$$\partial_\alpha \mathbb{E}[r(X_T^\alpha)] = \mathbb{E}\big[\nabla r(X_T^\alpha) \partial_\alpha X_T^\alpha\big]. \tag{40}$$

Cependant pour de nombreux problèmes la fonction récompense n'est pas régulière (par exemple pour un problème de cible, r est une fonction indicatrice).

On souhaiterait utiliser une formule d'intégration par parties dans (40) pour exprimer la sensibilité sous la forme $\mathbb{E}[r(X_T^\alpha) H]$ pour une certaine variable aléatoire H. Celle-ci pourrait être $\partial_\alpha \log P(X_T^\alpha)$ où $P(X_T^\alpha)$ est la densité de la loi de X_T^α, mais celle-ci est malheureusement inconnue habituellement.

Dans le cas particulier où le terme de diffusion σ ne dépend pas du contrôle, un changement de probabilité (théorème de Girsanov) permet de fournir une autre variable aléatoire H simulable numériquement. Cette approche [YK91], qui s'appelle la méthode du *score* ou du *rapport de vraisemblance* [Gly86, RW86, Gly87], fournit l'estimateur suivant :

Proposition 3 [YK91] *Si σ est inversible et indépendant de α, alors*

$$\partial_\alpha \mathbb{E}[r(X_T^\alpha)] = \mathbb{E}\left[r(X_T) \int_0^T [\sigma_t^{-1} \partial_\alpha f_t]' dW_t \right]. \tag{41}$$

Mentionnons que cette méthode est appliquée en temps discret pour la recherche directe de politiques paramétrées [Wil92, BB01, MT03].

Notre contribution [1] porte sur le cas général où r n'est pas régulière et où σ peut dépendre de α. Nous proposons trois nouvelles approches pour exprimer la sensibilité sous la forme d'une espérance.

5.1 Approche calcul de Malliavin

L'idée consiste à utiliser dans (40) une formule d'intégration par parties au sens du calcul de Malliavin [Nua95]. Dans le cadre elliptique, cette approche est développée dans le domaine de la finance [FLL+99] pour calculer des sensibilités de prix d'option. Notre contribution est une généralisation de ce résultat à des conditions plus faibles que l'ellipticité (i.e. ellipticité partielle ou hypoellipticité [CM02]).

Théorème 9 [1]. *Supposons que la matrice de covariance de Malliavin de X_T^α, définie par $\Gamma_T := \int_0^T \mathcal{D}_t X_T^\alpha [\mathcal{D}_t X_T^\alpha]' dt$, soit inversible avec des inverses dans tous les L_p, $p \geq 1$. Alors,*

$$\partial_\alpha \mathbb{E}[r(X_T^\alpha)] = \frac{1}{T}\mathbb{E}\left[r(X_T)\delta\big([\partial_\alpha X_T^\alpha]'\Gamma_T^{-1}\mathcal{D}_\cdot X_T^\alpha\big)\right]. \tag{42}$$

Ici δ est l'intégrale de Skorohod et $\mathcal{D}_\cdot X_T^\alpha$ est la dérivée de Malliavin, définie par : $\mathcal{D}_t X_T^\alpha = Y_T Y_t^{-1}\sigma_t \mathbf{1}_{t\leq T}$, où Y_t, la sensibilité de X_t^α par rapport à la condition initiale $Y_t := \nabla_x X_t^\alpha$, satisfait [Kun84] :

$$Y_t = I_d + \int_0^t \nabla_x f_s Y_s ds + \sum_{j=1}^q \int_0^t \nabla_x \sigma_{j,s} Y_s dW_s^j.$$

La formule (42) est compacte mais cache une importante complexité computationnelle (voir [20] pour les détails d'une implémentation dans le cas elliptique). Lorsque σ est inversible, d'autres approches plus simples sont les suivantes.

5.2 Approche par les états adjoints

Sous des conditions de régularité de f et σ, V^α est la solution régulière de l'EDP (linéaire) :

$$\partial_t V^\alpha(t,x) + \sum_{i=1}^d f_i(t,x,\alpha)\partial_{x_i} V^\alpha(t,x) + \frac{1}{2}\sum_{i,j=1}^d [\sigma\sigma']_{ij}(t,x,\alpha)\partial^2_{x_i x_j} V^\alpha(t,x) = 0,$$

pour $t < T$ et $V^\alpha(T,x) = r(x)$. En différentiant (formellement) cette EDP et en réinterprétant la dérivée en termes d'espérance, on obtient [1] la formule de Feynman-Kac :

$$\partial_\alpha \mathbb{E}[r(X_T^\alpha)] = \int_0^T \mathbb{E}\left[\sum_{i=1}^d \partial_\alpha f_{i,t}\partial_{x_i} V^\alpha(t,X_t^\alpha) + \frac{1}{2}\sum_{i,j=1}^d \partial_\alpha [\sigma\sigma']_{ij,t}\partial^2_{x_i x_j} V^\alpha(t,X_t^\alpha)\right] dt.$$

Remarquons que cette formulation est liée au *Principe du maximum de Pontryagin* dans le cas stochastique : les processus $[\partial_{x_i} V^\alpha(t,X_t^\alpha)]_{0\leq t\leq T}$ et $[\partial^2_{x_i x_j} V^\alpha(t,X_t^\alpha)]_{0\leq t\leq T}$ étant les *états adjoints* qui satisfont des EDS rétrogrades [Ben88, Pen90, YZ99].

Pour calculer $\partial_\alpha \mathbb{E}[r(X_T^\alpha)]$, nous rendons explicites les quantités $\partial_{x_i} V^\alpha$ et $\partial^2_{x_i x_j} V^\alpha$ par une formule d'intégration par parties [Pic02] plus simple que précédemment (utilisant uniquement des techniques de calcul stochastique de Bismut [Bis84] et évitant ainsi les intégrales de Skorohod). De la propriété de martingale de $[V^\alpha(t,X_t^\alpha)]_{0\leq t\leq T}$, on déduit que le processus $[\nabla_x V^\alpha(t,X_t^\alpha)Y_t]_{0\leq t\leq T}$ est aussi une martingale, ce qui permet de déduire la dérivée première :

$$\nabla_x V^\alpha(t,X_t^\alpha)Y_t = \frac{1}{T-t}\mathbb{E}\big[r(X_t^\alpha)\big(\int_t^T [\sigma_s^{-1}Y_s]' dW_s\big)' | \mathcal{F}_t\big], \tag{43}$$

où l'on a utilisé le théorème de représentation prévisible

$$r(X_T^\alpha) = V^\alpha(t,X_t^\alpha) + \int_t^T \nabla_x V^\alpha(s,X_s^\alpha)\sigma_s dW_s.$$

La dérivée seconde s'obtient en généralisant ce type de calcul. Nous en déduisons le résultat suivant :

Théorème 10 [1]. *Supposons σ inversible, alors*

$$\partial_\alpha \mathbb{E}[r(X_T^\alpha)] = \frac{1}{T}\mathbb{E}\Big[r(X_T)(H_T^f + H_T^\sigma)\Big], \quad où$$

$$\begin{aligned}
H_T^f &= \int_0^T dt\, \partial_\alpha f_t \cdot \frac{(Y_t^{-1})'}{T-t} \int_t^T [\sigma_s^{-1} Y_s]' dW_s, \\
H_T^\sigma &= \int_0^T dt \sum_{i,j=1}^d \partial_\alpha [\sigma\sigma']_{ij,t} \left(\frac{2e_j}{T-t} \cdot \big[(Y_t^{-1})' \int_{\frac{t+T}{2}}^T [\sigma_s^{-1} Y_s]' dW_s\big] \right. \\
&\quad \times \frac{e_i}{T-t} \cdot \big[(Y_t^{-1})' \int_t^{\frac{t+T}{2}} [\sigma_s^{-1} Y_s]' dW_s\big] \\
&\quad \left. + \frac{e_i}{T-t} \cdot \left\{ \nabla_x \big[(Y_t^{-1})' \int_t^{\frac{t+T}{2}} [\sigma_s^{-1} Y_s]' dW_s\big] Y_t^{-1} e_j \right\} \right)
\end{aligned}$$

où e_i est le i^e vecteur de la base canonique.

Cette formule est (malgré les apparences !) beaucoup plus simple à implémenter que (42) (voir le guide [20] pour une implémentation numérique efficace). Elle est d'autant plus intéressante numériquement que le nombre de paramètres est grand : les termes en facteur de $\partial_\alpha f$ et $\partial_\alpha[\sigma\sigma']$ se calculant une seule fois quelque soit le nombre de paramètres.

5.3 Approche Martingale

Cette approche est la plus simple et n'utilise que des propriétés de martingale des processus : $[V^\alpha(t, X_t^\alpha)]_{0\le t\le T}$ et $[\nabla_x V^\alpha(t, X_t^\alpha)\partial_\alpha X_t^\alpha]_{0\le t\le T}$. Ainsi, en dérivant l'égalité $\mathbb{E}[V^\alpha(t, X_t^\alpha)] = \frac{1}{T-t}\int_t^T \mathbb{E}[V^\alpha(s, X_s^\alpha)]ds$ par rapport à α, il vient l'équation intégrale

$$g(t) = \frac{1}{T-t}\int_t^T g(s)ds + h(t) \tag{44}$$

avec $g(t) := \mathbb{E}[\partial_\alpha V^\alpha(t, X_t^\alpha)]$ et

$$h(t) := \frac{1}{T-t}\int_t^T \mathbb{E}\big[\nabla_x V^\alpha(s, X_s^\alpha)(\partial_\alpha X_s^\alpha - Y_s Y_t^{-1}\partial_\alpha X_t^\alpha)\big]ds.$$

De manière similaire à (43), on déduit une formule de représentation pour h :

$$h(t) = \frac{1}{T-t}\mathbb{E}\left[r(X_T^\alpha)\int_t^T \big(\sigma_s^{-1}(\partial_\alpha X_t^\alpha - Y_s Y_t^{-1}\partial_\alpha X_t^\alpha)\big)' dW_s\right].$$

On conclut en résolvant l'équation intégrale (44) : $\partial_\alpha \mathbb{E}[r(X_T^\alpha)] = g(0) = h(0) - \int_0^T \frac{h(t)}{T-t}dt$.

Théorème 11 [1]. *Supposons σ inversible, alors*

$$\begin{aligned}
\partial_\alpha \mathbb{E}[r(X_T^\alpha)] &= \mathbb{E}\left[r(X_T)\left(\frac{1}{T}\int_0^T [\sigma_s^{-1}\partial_\alpha X_t^\alpha]' dW_s \right.\right. \\
&\quad \left.\left. + \int_0^T \frac{dt}{(T-t)^2}\int_t^T \big[\sigma_s^{-1}(\partial_\alpha X_t^\alpha - Y_s Y_t^{-1}\partial_\alpha X_t^\alpha)\big]' dW_s\right)\right]
\end{aligned} \tag{45}$$

Notons que cette intégrale converge pour des fonctions r indicatrices [1]. Cette formule est intéressante car très simple à mettre en œuvre numériquement et sa complexité est du même ordre que celle de l'estimateur trajectorielle (40).

L'article [1] propose une comparaison des estimateurs de sensibilité déduits des approches *trajectorielle*, *Malliavin*, *adjoints* et *martingale*, en termes de variance, de complexité numérique et de temps CPU de résolution pour des problèmes où r est régulière ou non-régulière. L'erreur de discrétisation pour les approches Malliavin et adjoints est aussi analysée : essentiellement, si les coefficients de la dynamique f et σ sont réguliers, l'erreur de discrétisation à un pas Δt est d'ordre $O(\Delta t)$.

Un guide pour l'implémentation numérique de ces méthodes est détaillé dans le rapport [20].

5.4 Application à la valorisation des options swing

Cette méthode de résolution approchée d'un problème de contrôle optimal stochastique par optimisation paramétrique de la politique est illustrée dans le travail [17] réalisé pour GDF.

Il s'agit de valoriser un contrat dit "*swing*" qui donne à son acheteur l'équivalent d'un gain financier $\Psi(q_t, S_t)$ où q_t est la quantité de gaz demandée et S_t le cours du gaz, à l'instant t. Par exemple, pour un *contrat d'approvisionnement* avec un prix d'achat fixe égal à K, on a $\Psi(q, S) = q(S - K)$. D'autres types de contrats (par exemple *de stockage*) sont possibles. Le vendeur du contrat détermine le prix d'un tel produit selon la consommation la plus défavorable :

$$\text{Prix} = \sup_{(q_t)_{0\leq t\leq T}\, admissible} \mathbb{E}\left[\int_0^T e^{-\beta t}\Psi(q_t, S_t)dt\right]$$

où l'espérance est prise sous la probabilité *neutre au risque*. Ainsi le prix est calculé en résolvant un problème de contrôle optimal stochastique. Les contraintes sur la consommation (le contrôle) $(q_t)_{0\leq t\leq T}$ sont, dans le cas du contrat d'approvisionnement, de type local ($q_{\min} \leq q_t \leq q_{\max}$) et global ($Q_{\min} \leq \int_0^T q_t \leq Q_{\max}$). Pour ce contrat, nous montrons [17] que la consommation optimale est de type bang-bang.

Nous testons deux paramétrisations du contrôleur : la première considère une représentation générale de $q_t = q(t, S_t, Q_t)$ (où $Q_t = \int_0^t q_s ds$ est la consommation accumulée) sous la forme d'un réseau de neurones (de paramètre α). La seconde paramétrisation tire profit du comportement bang-bang pour représenter la frontière (dans l'espace (t, S, Q)) de transition de la consommation de $q_{\min}$ à $q_{\max}$. En notant V^α le gain correspondant à la loi de consommation paramétrée par α, la sensibilité $\partial_\alpha V^\alpha$ est calculée selon l'estimateur trajectorielle (40), et permet d'effectuer un pas de gradient stochastique [BMP90, KY97] :

$$\alpha_{n+1} = \alpha_n + \eta_n \partial_\alpha V^\alpha.$$

Ces méthodes d'optimisation paramétrique du contrôleur se comparent favorablement [17] aux méthodes de programmation dynamique plus classiques : avec construction d'une *forêt d'arbres* [JRT04] ou utilisant une méthode de régression de type Longstaff-Schwartz [LS01].

5.5 Algorithmes d'A/R ?

Nous souhaitons conclure ce document en présentant un élément de réflexion (qui n'a pas fait l'objet de publication) à la question suivante. Peut-on définir des estimateurs de sensibilité du type (40), (41) ou (45) lorsque les coefficients f et σ de la dynamique d'état sont inconnus ? Cela serait appréciable dans des situations où il est coûteux de modéliser ces dynamiques, ou lorsque celles-ci ne sont pas accessibles (problématique $\mathcal{P}1$ de l'A/R).

Une astuce (usuelle en temps discret [Wil92, BB01, SMSM00, MT03]) consiste à utiliser une *politique stochastique*, notée π_α, définie par un choix aléatoire du contrôle u en tout état x selon une probabilité $\pi_\alpha(u|x)$, paramétrée par α. Cette approche permet de faire porter la dérivée des coefficients de la dynamique par rapport à α sur un rapport de vraisemblance $\partial_\alpha \log \pi_\alpha$ de la politique. Indiquons brièvement les étapes de cette démarche.

Considérons pour simplifier une dynamique déterministe de type (8) avec un critère à optimiser à horizon temporel fini : $r(x_T)$. Discrétisons, à un pas Δt, cette dynamique en utilisant la politique stochastique : le processus stochastique discret $X_t^{\Delta t}$, initié en x_0, est défini en tout $t \in \{j\Delta t\}_{0 \leq j < T/\Delta t}$, selon

$$\left\{ \begin{array}{rcl} u_t & \sim & \pi_\alpha(\cdot | X_t^{\Delta t}), \\ X_{t+\Delta t}^{\Delta t} & = & X_t^{\Delta t} + f(X_t^{\Delta t}, u_t)\Delta t. \end{array} \right.$$

Lorsque Δt est proche de 0, ce processus se comporte comme le système déterministe

$$\frac{dx_t^\alpha}{dt} = f(x_t^\alpha, \alpha),$$

(pour la même condition initiale $x_0^\alpha = x_0$) où $f(x, \alpha)$ est la dynamique moyenne :

$$f(x, \alpha) = \sum_{u \in U} \pi_\alpha(u|x) f(x, u).$$

Afin d'utiliser une formule de sensibilité trajectorielle (40) du gain par rapport à α :

$$\partial_\alpha r(x_T^\alpha) = \nabla r(x_T^\alpha) \partial_\alpha x_T^\alpha, \tag{46}$$

nous calculons la sensibilité de x_t^α par rapport à α, notée $z_t = \partial_\alpha x_t^\alpha$. Celle-ci suit la dynamique

$$\frac{dz_t}{dt} = \partial_\alpha f(x_t^\alpha, \alpha) + \nabla_x f(x_t^\alpha, \alpha) z_t \tag{47}$$

avec la condition initiale $z_0 = 0$. Nous discrétisons z_t par le processus stochastique discret $Z_t^{\Delta t}$, défini pour tout $t \in \{j\Delta t\}_{0 \leq j < T/\Delta t}$, par

$$\left\{ \begin{array}{rcl} u_t & \sim & \pi_\alpha(\cdot | X_t^{\Delta t}), \\ Z_{t+\Delta t}^{\Delta t} & = & Z_t^{\Delta t} + \big(\partial_\alpha \log \pi_\alpha(u_t | X_t^{\Delta t}) f(X_t^{\Delta t}, u_t) \\ & & \quad + \left[\nabla_x \log \pi_\alpha(u_t | X_t^{\Delta t}) f(X_t^{\Delta t}, u_t) + \nabla_x f(X_t^{\Delta t}, u_t) \right] Z_t^{\Delta t} \big) \, \Delta t. \end{array} \right.$$

Le processus couplé $(X_t^{\Delta t}, Z_t^{\Delta t})$ est une approximation de type chaîne de Markov du couple (x_t^α, z_t^α) qui vérifie la propriété de consistance locale (7) :

$$\begin{array}{rcl} \mathbb{E}\big[X_{t+\Delta t}^{\Delta t} - X_t^{\Delta t} | X_t^{\Delta t} = x \big] & = & f(x, \alpha)\Delta t, \\ \mathbb{E}\big[Z_{t+\Delta t}^{\Delta t} - Z_t^{\Delta t} | X_t^{\Delta t} = x, Z_t^{\Delta t} = z \big] & = & [\partial_\alpha f(x, \alpha) + \nabla_x f(x, \alpha) z] \Delta t, \end{array}$$

et les covariances sont d'ordre $O(\Delta t^2)$. En appliquant le résultat d'approximation de [KD01] au critère $f(x, z) := \nabla r(x) z$, on déduit que la valeur moyenne $\mathbb{E}[f(X_T^{\Delta t}, Z_T^{\Delta t})]$ du processus discret $(X_t^{\Delta t}, Z_t^{\Delta t})_{0 \leq t \leq T}$ tend vers la valeur $f(x_T^\alpha, z_T^\alpha)$ du système continu $(x_t^\alpha, z_t^\alpha)_{0 \leq t \leq T}$, lorsque $\Delta t \to 0$. Ainsi, (en supposant r régulière) la sensibilité (46) du gain par rapport à α est obtenue à la limite :

$$\partial_\alpha r(x_T^\alpha) = \lim_{\Delta t \to 0} \mathbb{E}\big[\nabla r(X_T^{\Delta t}) Z_T^{\Delta t} \big]. \tag{48}$$

L'intérêt de cette formule est que $Z_t^{\Delta t}$ peut être approché par des grandeurs estimées le long des trajectoires. Notons l'écart $\Delta X_t^{\Delta t} = X_{t+\Delta t}^{\Delta t} - X_t^{\Delta t}$, et considérons un estimateur $\widehat{\nabla_x f}(X_t^{\Delta t}, u)$ de $\nabla_x f(X_t^{\Delta t}, u)$, qui peut, par exemple être déduit d'une régression linéaire locale :

$$\widehat{\nabla_x f}(X_t^{\Delta t}, u) := \frac{1}{\Delta t} \big(\overline{\Delta X \, X'} - \overline{\Delta X} \, \overline{X'} \big) \big(\overline{X \, X'} - \overline{X} \, \overline{X'} \big)^{-1} \tag{49}$$

où les moyennes $\overline{X}$, $\overline{\Delta X}$, $\overline{X \, X'}$, $\overline{\Delta X \, X'}$ sont calculées localement (sur une petite fenêtre de temps $t - k\Delta t \leq s \leq t$) à partir des états $X_s^{\Delta t}$, écarts $\Delta X_s^{\Delta t}$, et leur produit, lorsque le contrôle $u_s = u$.

Ainsi, si l'estimateur $\widehat{\nabla_x f} \to \nabla_x f$ lorsque $\Delta t \to 0$, le calcul de $Z_t^{\Delta t}$ selon

$$\begin{aligned} Z_{t+\Delta t}^{\Delta t} &= Z_t^{\Delta t} + \partial_\alpha \log \pi_\alpha(u_t|X_t^{\Delta t})\Delta X_t^{\Delta t} \\ &\quad + \nabla_x \log \pi_\alpha(u_t|X_t^{\Delta t})\Delta X_t^{\Delta t} Z_t^{\Delta t} + \widehat{\nabla_x f}(X_t^{\Delta t}, u) Z_t^{\Delta t} \Delta t, \end{aligned}$$

(qui ne nécessite que la connaissance de la politique π_α et de la trajectoire $X_t^{\Delta t}$, si l'on utilise l'estimateur (49)) vérifie encore la propriété de consistance locale (7) et la sensibilité du gain est obtenue par la formule (48).

L'utilisation de ces politiques stochastiques, permettant de pallier l'absence de connaissance des dynamiques d'état, est peut-être liée à l'observation d'un "comportement oscillatoire" lors de l'apprentissage de certaines tâches motrices chez l'homme. Par exemple, lorsqu'il apprend à maintenir en équilibre un pendule inversé, il effectue des petits déplacements en alternance, où il semble "tester" les dynamiques en jeu pour ainsi mieux contrôler le système.

La généralisation à une dynamique stochastique (39) et à d'autres estimateurs de sensibilité (41) ou (45) fera l'objet de travaux futurs.

Références

[Amo02] Gideon Amos. *Solving the Hamilton-Jacobi-Bellman Equation for Animation.* PhD thesis, Centre for Advanced Instrumentation Systems, University College London, 2002.

[ASM97] C. G. Atkeson, S. A. Schaal, and Andrew W. Moore. Locally weighted learning. *AI Review*, 11, 1997.

[Bai95] Leemon C. Baird. Residual algorithms : Reinforcement learning with function approximation. *Machine Learning : proceedings of the Twelfth International Conference*, 1995.

[Bar94] Guy Barles. *Solutions de viscosité des équations de Hamilton-Jacobi*, volume 17 of *Mathématiques et Applications.* Springer-Verlag, 1994.

[BB96] S. Bradtke and A. Barto. Linear least-squares algorithms for temporal difference learning. *Journal of Machine Learning*, 22 :33–57, 1996.

[BB01] J. Baxter and P.L. Bartlett. Infinite-horizon gradient-based policy search. *Journal of Artificial Intelligence Research*, 15 :319–350, 2001.

[BCD97] Martino Bardi and Italo Capuzzo-Dolcetta. *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations.* Birkhauser Boston, 1997.

[Ben88] A. Bensoussan. *Perturbation methods in optimal control.* Wiley/Gauthier-Villars Series in Modern Applied Mathematics. John Wiley & Sons Ltd., Chichester, 1988. Translated from the French by C. Tomson.

[Ber87] Dimitri P. Bertsekas. *Dynamic Programming : Deterministic and Stochastic Models.* Prentice Hall, 1987.

[Bis84] J.M. Bismut. *Large deviations and the Malliavin calculus.* Birkhäuser Boston Inc., Boston, MA, 1984.

[BMP90] A. Benveniste, M. Metivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximations.* Springer-Verlag, New York, 1990.

[Boy99] Justin Boyan. Least-squares temporal difference learning. *Proceedings of the 16th International Conference on Machine Learning*, pages 49–56, 1999.

[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.

[BS91] Guy Barles and P.E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Analysis*, 4 :271–283, 1991.

[BT96] Dimitri P. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming.* Athena Scientific, 1996.

[BZ03] F. Bonnans and H. Zidani. Consistency of generalized finite difference schemes for the stochastic hjb equation. *SIAM J. on Numerical Analysis*, 41-3, 2003.

[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.

[CM02] P. Cattiaux and L. Mesnager. Hypoelliptic non-homogeneous diffusions. *Probab. Theory Related Fields*, 123(4) :453–483, 2002.

[Day92] Peter Dayan. The convergence of td(λ) for general λ. *Machine Learning*, 8 :341–362, 1992.

[DeV97] R. DeVore. *Nonlinear Approximation.* Acta Numerica, 1997.

[DFA99] Richard Dearden, Nir Friedman, and David Andre. Model based bayesian exploration. *Proceeding of Uncertainty in Artificial Intelligence*, 1999.

[DMA97] G.M. Davies, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *J. of Constr. Approx.*, 13 :57–98, 1997.

[FBF77] J. H. Friedman, J. L. Bentley, and R. A. Finkel. An Algorithm for Finding Best Matches in Logarithmic Expected Time. *ACM Trans. on Mathematical Software*, 3(3) :209–226, September 1977.

[FLL+99] E. Fournié, J.M. Lasry, J. Lebuchoux, P.L. Lions, and N. Touzi. Applications of Malliavin calculus to Monte Carlo methods in finance. *Finance and Stochastics*, 3(4) :391–412, 1999.

[FR75] Wendell H. Fleming and R.W. Rishel. *Deterministic and Stochastic Optimal Control.* Springer-Verlag, New York, 1975.

[FS93] Wendell H. Fleming and H. Mete Soner. *Controlled Markov Processes and Viscosity Solutions.* Applications of Mathematics. Springer-Verlag, 1993.

[GKP01] Carlos Guestrin, Daphne Koller, and Ronald Parr. Max-norm projections for factored mdps. *Proceedings of the International Joint Conference on Artificial Intelligence*, 2001.

[Gly86] P.W. Glynn. Stochastic approximation for Monte Carlo optimization. In J. Wilson, J. Henriksen, and S. Roberts, editors, *Proceedings of the 1986 Winter Simulation Conference*, pages 356–365, 1986.

[Gly87] P.W. Glynn. Likelihood ratio gradient estimation : an overview. In A. Thesen, H. Grant, and W.D. Kelton, editors, *Proceedings of the 1987 Winter Simulation Conference*, pages 366–375, 1987.

[Gor95] G. Gordon. Stable function approximation in dynamic programming. *Proceedings of the International Conference on Machine Learning*, 1995.

[Grü97] Lars Grüne. An adaptive grid scheme for the discrete hamilton-jacobi-bellman equation. *Numerische Mathematik*, 75-3, 1997.

[Hay94] S. Haykin. *Neural Networks : A Conprehensive Foundation.* McMillan, New York, 1994.

[HTF01] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics, 2001.

[JRT04] P. Jaillet, E.I. Ronn, and S. Tompaidis. Valuation of commodity based swing options. *Management Science*, 50(7) :909–921, 2004.

[Jud98] Kenneth Judd. *Numerical Methods in Economics.* MIT Press, 1998.

[KB99] V. R. Konda and V. S. Borkar. Actor-critic-type learning algorithms for markov decision processes. *SIAM Journal of Control and Optimization*, 38 :1 :94–123, 1999.

[KD01] Harold J. Kushner and Paul Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time. Second Edition.* Applications of Mathematics. Springer, 2001.

[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.

[Kun84] H. Kunita. Stochastic differential equations and stochastic flows of diffeomorphisms. *Ecole d'Eté de Probabilités de St-Flour XII, 1982 - Lecture Notes in Math. 1097 - Springer Verlag*, pages 144–305, 1984.

[KY97] H. J. Kushner and G. Yin. *Stochastic Approximation Algorithms and Applications.* Springer-Verlag, Berlin and New York, 1997.

[LM80] P.-L. Lions and B. Mercier. Approximation numérique des équations de hamilton-jacobi-bellman. *R.A.I.R.O. Analyse numérique*, 14(4) :369–393, 1980.

[LP94] P. L'Ecuyer and G. Perron. On the convergence rates of IPA and FDC derivative estimators. *Oper. Res.*, 42(4) :643–656, 1994.

[LS01] F. Longstaff and E.S. Schwartz. Valuing american options by simulation : A simple least squares approach. *The Review of Financial Studies*, 14 :113–147, 2001.

[MA95] Andrew W. Moore and C.G. Atkeson. The parti-game algorithm for variable resolution reinforcement learning in multidimensional state space. *Machine Learning Journal*, 21, 1995.

[Mal97] Stephane Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, 1997.

[Meu96] Nicolas 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.

[Moo92] D. W. Moore. *Simplical Mesh Generation with Applications.* PhD thesis, Cornell University, 1992.

[MT03] P. Marbach and J. N. Tsitsiklis. Approximate gradient methods in policy-space optimization of markov reward processes. *Journal of Discrete Event Dynamical Systems*, 13 :111–148, 2003.

[Nie92] H. Niederreiter. Random number generation and quasi-monte carlo methods. *SIAM CBMS-NSF Conference Series in Applied Mathematics, Philadelphia*, 63, 1992.

[Nua95] D. Nualart. *Malliavin calculus and related topics.* Springer Verlag, 1995.

[Pen90] S.G. Peng. A general stochastic maximum principle for optimal control problems. *SIAM J. Control Optim.*, 28(4) :966–979, 1990.

[Pic02] J. Picard. Gradient estimates for some diffusion semigroups. *Probab. Theory Related Fields*, 122 :593–612, 2002.

[Put94] Martin L. Puterman. *Markov Decision Processes, Discrete Stochastic Dynamic Programming.* A Wiley-Interscience Publication, 1994.

[Rup94] Jim Ruppert. A delaunay refinement algorithm for quality 2-dimensional mesh generation. *Journal of Algorithms*, 1994.

[Rus96] John Rust. *Numerical Dynamic Programming in Economics.* In Handbook of Computational Economics. Elsevier, North Holland, 1996.

[Rus97] John Rust. *Using Randomization to Break the Curse of Dimensionality.* Computational Economics. 1997.

[RW72] R.A. Rescorla and A.R. Wagner. A theory of pavlovian conditioning : variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II : Current research and theory*, 1972.

[RW86] M.I. Reiman and A. Weiss. Sensitivity analysis via likelihood ratios. In J. Wilson, J. Henriksen, and S. Roberts, editors, *Proceedings of the 1986 Winter Simulation Conference*, pages 285–289, 1986.

[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.

[SB98] Richard S. Sutton and Andrew G. Barto. Reinforcement learning : An introduction. *Bradford Book*, 1998.

[Sch02] R. Schoknecht. Optimality of reinforcement learning algorithms with linear function approximation. *Proceedings of the 15th Neural Information Processing Systems conference*, 2002.

[Set99] J.A. Sethian. *Level Set Methods and Fast Marching Methods.* Cambridge University Press, 1999.

[Ski97] Steven Skiena. *Algorithm Design Manual.* Springer Verlag, 1997.

[SMSM00] R.S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. *Neural Information Processing Systems. MIT Press*, pages 1057–1063, 2000.

[Sut88] R.S. Sutton. Learning to predict by the method of temporal differences. *Machine Learning*, pages 9–44, 1988.

[Tsi94] J.N. Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine Learning*, 16 :185–202, 1994.

[TTZ03] H.-Z. Tang, T. Tang, and P. Zhang. An adaptive mesh redistribution method for non-linear hamilton-jacobi equations in two- and three- dimensions. *Journal of Computational Physics*, 188 :543–572, 2003.

[VGS97] Vladimir Vapnik, Steven E. Golowich, and Alex Smola. Support vector method for function approximation, regression estimation and signal processing. *In Advances in Neural Information Processing Systems*, pages 281–287, 1997.

[Wat89] Christopher J.C.H. Watkins. *Learning from delayed reward.* PhD thesis, Cambridge University, 1989.

[WB93] R.J. Williams and L.C. Baird. Tight performance bounds on greedy policies based on imperfect value functions. *Technical Report NU-CCS-93-14. Northeastern University*, 1993.

[WD92] Christopher J.C.H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8 :279–292, 1992.

[Wil92] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8 :229–256, 1992.

[YK91] J. Yang and H.J. Kushner. A Monte Carlo method for sensitivity analysis and parametric optimization of nonlinear stochastic systems. *SIAM J. Control Optim.*, 29(5) :1216–1249, 1991.

[YZ99] J. Yong and X.Y. Zhou. *Stochastic Controls : Hamiltonian Systems and HJB Equations.* Springler-Verlag, New York, 1999.