Généralisation Min Max pour l'Apprentissage par Renforcement Batch et Déterministe : Schémas de Relaxation

Raphael Fonteneau, Damien Ernst, Bernard Boigelot, Quentin Louveaux

Département d'Electricité, Electronique et Informatique
Université de Liège, Belgique {raphael.fonteneau, dernst, bernard.boigelot,q.louveaux}@ulg.ac.be

Résumé : On s'intéresse au problème de généralisation min max dans le cadre de l'apprentissage par renforcement batch et déterministe. Le problème a été originellement introduit par Fonteneau *et al.* (2011). Dans un premier temps, on montre que le problème est NP-dur. Dans le cas où l'horizon d'optimisation vaut 2, on développe deux schémas de relaxation. Le premier schéma fonctionne en éliminant des contraintes de telle sorte qu'on obtienne un problème soluble en temps polynomial. Le deuxième schéma est une relaxation Lagrangienne conduisant à un problème conique-quadratique. On montre théoriquement et empiriquement que ces deux schémas permettent d'obtenir de meilleurs résultats que ceux proposés par Fonteneau *et al.* (2011).
Mots-clés : Apprentissage par renforcement, Généralisation min max, Optimisation non convexe, Complexité algorithmique

1 Introduction

Les recherches menées en apprentissage par renforcement (RL) ont pour but principal de développer des agents intelligents capables d'apprendre comment interagir avec leur environnement afin de maximiser un critère de récompense. Les techniques issues de ces recherches se sont exportées vers d'autres champs d'applications, notamment la finance (Ingersoll (1987)), la médecine (Murphy (2003, 2005)) ou l'ingénierie (Riedmiller (2005)). Depuis la fin des années 90, une communauté de recherche s'est constituée autour de la résolution d'un sous-problème du RL : calculer des politiques de décision performantes lorsque l'on ne connaît qu'un ensemble fini de trajectoires du système que l'on souhaite contrôler (Bradtke & Barto (1996); Ernst *et al.* (2005); Lagoudakis & Parr (2003); Ormoneit & Sen (2002); Riedmiller (2005); Fonteneau (2011)). Ce cas particuler du RL est communément appelé RL en mode "batch" (BMRL).

Les algorithmes BMRL sont mis en difficulté par les espaces d'état continus ou de grandes tailles, pour lesquels ils doivent mettre en oeuvre des procédés de généralisation de l'information (potentiellement éparse) contenue dans l'échantillon de trajectoires. La parade la plus courante consiste à combiner les algorithmes BMRL avec des approximateurs de fonction (Bertsekas & Tsitsiklis (1996); Lagoudakis & Parr (2003); Ernst *et al.* (2005); Busoniu *et al.* (2010)). Ces approximateurs (réseaux de neurones, ensembles d'arbres de régression, machines à supports vectoriels, etc) généralisent l'information contenue dans l'échantillon de trajectoires en prolongeant les propriétés du système depuis les zones connues (via l'échantillon) vers leurs voisinages inconnus. Ce procédé induit inévitablement une dégradation des garanties de performance des algorithmes BMRL lorsque de vastes zones de l'espace d'état ne sont pas décrites par l'échantillon. En effet, dans de telles situations, le calcul des garanties prend en compte le fait que la politique de décision inférée peut mener le système dans des zones inconnues mais supposées intéressantes par le procédé de généralisation, alors que ces dernières sont potentiellement catastrophiques. Cette constatation est également corroborée par des résultats théoriques montrant que les garanties de performances des politiques inférées par des algorithmes BMRL se dégradent avec la dispersion de l'échantillon (la dispersion pouvant être décrite comme l'étendue de la plus grande zone de l'espace non décrite par l'échantillon de trajectoires) (Fonteneau *et al.* (2010b)).

Dans l'optique de contourner ce problème, Fonteneau *et al.* (2011) proposent de généraliser selon une stratégie de type min max pour les environnements déterministes, continus et Lipschitziens avec un espace de décision fini et un horizon d'optimisation fini. La stratégie min max consiste à identifier une séquence de décisions menant à la maximisation du pire retour que l'on pourrait obtenir en considérant n'importe quel système compatible avec l'échantillon de trajectoires et la connaissance a priori de bornes supérieures sur les constantes de Lipschitz du système. Cependant, l'approche proposée par Fonteneau *et al.* (2011) est loin d'être facile à mettre en oeuvre, même après plusieurs reformulations permettant d'éviter la recherche dans un espace fonctionnel. Dans la suite de Fonteneau *et al.* (2011), la valeur du pire retour possible est remplacée par une borne inférieure, et un algorithme permettant de déterminer une séquence de décisions maximisant cette borne est proposé : l'algorithme CGRL ("Cautious approach to Generalization in Reinforcement Learning"). Cette borne inférieure, provenant de leurs travaux précédents (Fonteneau *et al.* (2009, 2010a)), se montre malheureusement empiriquement assez imprécise.

Dans cet article, on investigue de manière plus approfondie le problème de génération min max initialement proposé par Fonteneau *et al.* (2011). Tout d'abord, on montre que ce problème est NP-dur. On s'intéresse ensuite au cas "deux-étapes", qui est toujours NP-dur. Puisqu'il semble difficile de résoudre de façon exacte ce problème, on propose deux schémas de relaxation qui préservent la nature du problème de généralisation min max, c'est à dire menant à des politiques de décision dont les performances sont garanties. Le premier schéma de relaxation fonctionne par élimination de certaines contraintes afin de déboucher sur un problème soluble en temps polynomial. On obtient alors un problème de type *région de confiance* (Conn *et al.* (2000)). Le second schéma de relaxation, une relaxation Lagrangienne dans laquelle toutes les contraintes sont dualisées, débouche sur un problème conique-quadratique, également soluble en temps polynomial. On monte que ces deux schémas permettent d'obtenir des bornes plus précises que CGRL. Par ailleurs, ces bornes convergent vers le retour des politiques associées lorsque la dispersion de l'échantillon de trajectoires converge vers 0, et les séquences de décisions maximisant ces bornes convergent également vers des séquence de décisions optimales.

La suite du papier est organisée de la manière suivante : en section 2, on donne un bref exposé des travaux connexes à celui-ci. La section 3 formalise le problème de généralisation min max dans le contexte BMRL déterministe et Lipschitzien. En section 4, on se focalise sur le cas "deux-étapes", et on donne la preuve du caractère NP-dur du problème. La section 5 présente deux schémas de relaxation, ainsi que leurs propriétés théoriques, tandis que la section 6 en illustre quelques propriétés empiriques sur un problème-jouet. Quelques perspectives d'amélioration sont proposées en section 7. On trouvera toutes les démonstrations des résultats théoriques donnés dans ce papier dans Fonteneau *et al.* (2012).

2 Travaux connexes

Différentes approches de calcul de politiques de décision en RL ont été bâties suivant le paradigme min max. Dans un contexte stochastique, le paradigme min max a été utilisé afin de calculer des politiques robustes vis-à-vis des incertitudes dans l'identification des paramètres des distributions de probabilités associées à l'environnement (Delage & Mannor (2010)). Quand plusieurs agents interagissent les uns avec les autres dans un environnement commun, l'approche min max se révèle être efficace pour mettre au point des politiques maximisant les récompenses obtenues par un agent étant donné les comportements les moins favorables des autres agents (Littman (1994); Rovatous & Lagoudakis (2010)). Ces approches ont également été utilisées dans le cadre de la résolution de processus de décision markoviens partiellement observables (Littman (2009); Koenig (2001)).

L'approche min max en généralisation, originellement introduite par Fonteneau *et al.* (2011), repose implicitement sur des techniques de calcul de bornes inférieures sur le retour de politiques de décisions dans un contexte déterministe pour lequel l'environnement est très peu connu. De ce point de vue, ce travail est connexe à toute autre approche visant à calculer des garanties de performances sur le retour de politiques de décision (Mannor *et al.* (2004); Qian & Murphy (2009); Paduraru *et al.* (2011)).

D'autres domaines de recherche ont proposé des approches de type min max pour calculer des politiques de décision, notamment en théorie du contrôle (Hansen & Sargent (2001)) avec le contrôle

robuste H_∞ (Başar & Bernhard (1995)), ou en commande prédictive (Camacho & Bordons (2004); Ernst *et al.* (2009)) pour laquelle le paradigme min max vise à déterminer une séquence de décisions optimale par rapport aux pires aléas possibles (Scokaert & Mayne (1998); Bemporad & Morari (1999)). Enfin, il y a également une littérature assez fournie en programmation stochastique (Birge & Louveaux (1997)) s'intéressant à la planification prudente sous incertitudes (Defourny *et al.* (2008); Shapiro (2011a,b); Nemirovski *et al.* (2009)). Dans ce contexte, le cas "deux-étapes" a été particulièrement étudié (Frauendorfer (1992); Darby-Dowman *et al.* (2000)).

3 Formalisation du problème

On formalise le BMRL en section 3.1 et le problème de généralisation min max en section 3.2.

3.1 Apprentissage par renforcement en mode batch

On considère un système à temps discret déterministe dont la dynamique sur T pas de temps est décrite par l'équation :

$$x_{t+1} = f(x_t, u_t) \quad t = 0, \ldots, T-1,$$

où pour tout t, l'état x_t est un élément de l'espace d'état $\mathcal{X} \subset \mathbb{R}^d$ où $\mathbb{R}^d$ désigne l'espace Euclidien de dimension d et où u_t est un élément de l'espace de décision fini $\mathcal{U} = \left\{u^{(1)}, \ldots, u^{(m)}\right\}$ que l'on identifie (par abus de notation) à $\{1, \ldots, m\}$. $T \in \mathbb{N} \setminus \{0\}$ est l'horizon d'optimisation fini. Une récompense instantanée

$$r_t = \rho(x_t, u_t) \in \mathbb{R}$$

est associée à la prise de décision u_t dans l'état x_t. Pour un état initial donné $x_0 \in \mathcal{X}$ et une séquence de décisions $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$, le retour cumulé sur T pas de temps (retour) s'écrit :

Définition 1 (Retour)
$\forall (u_0, \ldots, u_{T-1}) \in \mathcal{U}^T, J_T^{(u_0,\ldots,u_{T-1})} \triangleq \sum_{t=0}^{T-1} \rho(x_t, u_t)$ *où* $x_{t+1} = f(x_t, u_t)$ $, \forall t \in \{0, \ldots, T-1\}$.

Une séquence de décisions optimale est une séquence menant à la maximisation du retour :

Définition 2 (Retour optimal)

$$J_T^* \triangleq \max_{(u_0,\ldots,u_{T-1}) \in \mathcal{U}^T} J_T^{(u_0,\ldots,u_{T-1})} .$$

On effectue également les hypothèses suivantes, caractéristiques du *mode batch* :

1. La dynamique f et la fonction de récompense ρ sont *inconnues* ;
2. Pour toute décision $u \in \mathcal{U}$, un ensemble de $n^{(u)} \in \mathbb{N}$ transitions

$$\mathcal{F}^{(u)} = \left\{\left(x^{(u),k}, r^{(u),k}, y^{(u),k}\right)\right\}_{k=1}^{n^{(u)}}$$

est connu. Chaque transition est telle que :

$$y^{(u),k} = f\left(x^{(u),k}, u\right) \text{ et } r^{(u),k} = \rho\left(x^{(u),k}, u\right) .$$

3. Chaque ensemble $\mathcal{F}^{(u)}$ contient au moins un élément : $\forall u \in \mathcal{U}, \quad n^{(u)} > 0$.

Par la suite, on désigne par $\mathcal{F}$ l'ensemble de toutes les transitions : $\mathcal{F} = \mathcal{F}^{(1)} \cup \ldots \cup \mathcal{F}^{(m)}$. Sous ces hypothèses, l'apprentissage par renforcement en mode batch (BMRL) a pour but d'inférer à partir de l'échantillon $\mathcal{F}$ une séquence de décisions performante, c'est à dire une séquence $\left(\tilde{u}_0^*, \ldots, \tilde{u}_{T-1}^*\right) \in \mathcal{U}^T$ telle que $J_T^{(\tilde{u}_0^*,\ldots,\tilde{u}_{T-1}^*)}$ est le plus proche possible de J_T^*.

3.2 Généralisation min max sous hypothèses de continuité Lipschitzienne

On énonce ici le problème de généralisation min max étudié dans ce papier. La formalisation originelle est donnée par Fonteneau *et al.* (2011).

On fait tout d'abord l'hypothèse que la dynamique du système f et la fonction de récompense ρ sont Lipschitziennes, c'est à dire qu'il existe deux constantes $L_f, L_\rho \in \mathbb{R}$ telles que :

$$\begin{aligned} \forall (x,x') \in \mathcal{X}^2, \forall u \in \mathcal{U}, \qquad \|f(x,u) - f(x',u)\| &\leq L_f \|x - x'\| \ , \\ |\rho(x,u) - \rho(x',u)| &\leq L_\rho \|x - x'\| \ , \end{aligned}$$

où $\|.\|$ la norme Euclidienne sur l'espace $\mathcal{X}$. On suppose également que deux constantes L_f et L_ρ satisfaisant les inégalités ci-dessus sont connues.

Etant donnée une séquence de décisions, on peut définir le pire retour qui pourrait être obtenu par un système dont la dynamique f' et la fonction de récompense ρ' seraient L_f et L_ρ Lipschitziennes tout en coïncidant avec les valeurs de f et ρ données par l'échantillon $\mathcal{F}$. D'après Fonteneau *et al.* (2011), ce pire retour possible peut-être obtenu en résolvant un problème d'optimisation dans l'espace $\mathcal{X}^{T-1} \times \mathbb{R}^T$ correspondant intuitivement à identifier la trajectoire la plus pessimiste possible tout en restant compatible avec les inégalités de Lipschitz et les données de $\mathcal{F}$. Concrètement, étant donnée une séquence de décisions $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$ et un état initial $x_0 \in \mathcal{X}$, le problème d'optimisation s'écrit :

$(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1}))$:

$$\min_{\substack{\hat{\mathbf{r}}_\mathbf{0} \ \ldots \ \hat{\mathbf{r}}_{\mathbf{T-1}} \in \mathbb{R} \\ \hat{\mathbf{x}}_\mathbf{0} \ \ldots \ \hat{\mathbf{x}}_{\mathbf{T-1}} \in \mathcal{X}}} \sum_{t=0}^{T-1} \hat{\mathbf{r}}_\mathbf{t},$$

sous contraintes

$$\begin{aligned} &\left|\hat{\mathbf{r}}_\mathbf{t} - r^{(u_t),k_t}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_\mathbf{t} - x^{(u_t),k_t}\right\|^2 \ , \forall (t,k_t) \in \{0,\ldots,T-1\} \times \left\{1,\ldots,n^{(u_t)}\right\}, \\ &\left\|\hat{\mathbf{x}}_\mathbf{t+1} - y^{(u_t),k_t}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_\mathbf{t} - x^{(u_t),k_t}\right\|^2, \forall (t,k_t) \in \{0,\ldots,T-1\} \times \left\{1,\ldots,n^{(u_t)}\right\}, \\ &|\hat{\mathbf{r}}_\mathbf{t} - \hat{\mathbf{r}}_\mathbf{t'}|^2 \leq L_\rho^2 \|\hat{\mathbf{x}}_\mathbf{t} - \hat{\mathbf{x}}_\mathbf{t'}\|^2 \ , \forall t,t' \in \{0,\ldots,T-1 | u_t = u_{t'}\}, \\ &\|\hat{\mathbf{x}}_\mathbf{t+1} - \hat{\mathbf{x}}_\mathbf{t'+1}\|^2 \leq L_f^2 \|\hat{\mathbf{x}}_\mathbf{t} - \hat{\mathbf{x}}_\mathbf{t'}\|^2 \ , \forall t,t' \in \{0,\ldots,T-2 | u_t = u_{t'}\}, \\ &\hat{\mathbf{x}}_\mathbf{0} = x_0 \ . \end{aligned}$$

Tout au long du papier, les variables d'optimisation seront écrites en caractères gras.

L'approche min max en généralisation consiste à identifier une séquence de décision pour laquelle le pire retour possible est maximisé, c'est à dire une séquence de décisions maximisant $(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1}))$.

On s'intéresse dans ce papier à la mise au point de schémas de relaxation pour le problème $(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1}))$. Ces schémas peuvent ensuite être utilisés comme oracles pour aborder le problème de généralisation min max.

Plus tard dans ce papier, on analyse également la complexité algorithmique du problème de généralisation min max. Lors de cette analyse, on fait l'hypothèse que toutes les données du problème (c'est à dire $T, \mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1}$) sont sous forme de nombres rationnels. On rappelle que toutes les démonstrations des résultats théoriques énoncés dans ce papier sont détaillées dans Fonteneau *et al.* (2012).

4 Le cas “deux-étapes"

Dans cette section, on se focalise sur le cas “deux-étapes”, c'est-à-dire le cas $T = 2$, qui est un cas particulier important du problème original $\left(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1})\right)$. Beaucoup de travaux ont spécialement abordé le cas “deux-étapes” (Frauendorfer (1992); Darby-Dowman *et al.* (2000)), ceci en raison d'un nombre important d'applications liées à ce dernier, notamment dans le contexte

médical avec l'inférence de traitements médicaux "sûrs" à partir de données issues de protocoles cliniques (Banerjee & Tsiatis (2006); Lokhnygina & Tsiatis (2008); Lunceford *et al.* (2002); Wahed & Tsiatis (2004)).

En section 4.1, on montre que ce problème peut-être décomposé en deux sous-problèmes indépendants. Le premier de ces problèmes se résout de manière analytique, tandis que le second est NP-dur (cf. section 4.2), ce qui implique que le problème "deux-étapes" et le problème général $(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1}))$ sont également NP-durs.

Etant donné une séquence de décisions $(u_0, u_1) \in \mathcal{U}^2$, le problème "deux-étapes" s'écrit :

$\big(\mathcal{P}_2(\mathcal{F}, L_f, L_\rho, x_0, u_0, u_1)\big)$:

$$\min_{\substack{\hat{\mathbf{r}}_\mathbf{0}, \hat{\mathbf{r}}_\mathbf{1} \in \mathbb{R} \\ \hat{\mathbf{x}}_\mathbf{0}, \hat{\mathbf{x}}_\mathbf{1} \in \mathcal{X}}} \quad \hat{\mathbf{r}}_\mathbf{0} + \hat{\mathbf{r}}_\mathbf{1},$$

sous contraintes

$$\left|\hat{\mathbf{r}}_\mathbf{0} - r^{(u_0),k_0}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_\mathbf{0} - x^{(u_0),k_0}\right\|^2 , \forall k_0 \in \left\{1, \ldots, n^{(u_0)}\right\} , \quad (1)$$

$$\left|\hat{\mathbf{r}}_\mathbf{1} - r^{(u_1),k_1}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_\mathbf{1} - x^{(u_1),k_1}\right\|^2 , \forall k_1 \in \left\{1, \ldots, n^{(u_1)}\right\} , \quad (2)$$

$$\left\|\hat{\mathbf{x}}_\mathbf{1} - y^{(u_0),k_0}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_\mathbf{0} - x^{(u_0),k_0}\right\|^2 , \forall k_0 \in \left\{1, \ldots, n^{(u_0)}\right\} , \quad (3)$$

$$|\hat{\mathbf{r}}_\mathbf{0} - \hat{\mathbf{r}}_\mathbf{1}|^2 \leq L_\rho^2 \|\hat{\mathbf{x}}_\mathbf{0} - \hat{\mathbf{x}}_\mathbf{1}\|^2 \text{ if } u_0 = u_1 , \quad (4)$$

$$\hat{\mathbf{x}}_\mathbf{0} = x_0 . \quad (5)$$

Par soucis de concision, on oubliera les arguments dans la définition du problème et on désignera $\big(\mathcal{P}_2(\mathcal{F}, L_f, L_\rho, x_0, u_0, u_1)\big)$ par $(\mathcal{P}_2^{(u_0,u_1)})$. On nomme $B_2^{(u_0,u_1)}(\mathcal{F})$ la valeur de l'objectif associé à la solution de $(\mathcal{P}_2^{(u_0,u_1)})$:

Définition 3 (Valeur optimale $B_2^{(u_0,u_1)}(\mathcal{F})$)
Soit $(u_0, u_1) \in \mathcal{U}^2$ et $(\hat{\mathbf{r}}_\mathbf{0}^*, \hat{\mathbf{r}}_\mathbf{1}^*, \hat{\mathbf{x}}_\mathbf{0}^*, \hat{\mathbf{x}}_\mathbf{1}^*)$ une solution optimale de $\left(\mathcal{P}_2^{(u_0,u_1)}\right)$. Alors,

$$B_2^{(u_0,u_1)}(\mathcal{F}) \triangleq \hat{\mathbf{r}}_\mathbf{0}^* + \hat{\mathbf{r}}_\mathbf{1}^* .$$

4.1 Découpler le problème "deux étapes"

Soit $(\mathcal{P}_2'^{(u_0,u_1)})$ et $(\mathcal{P}_2''^{(u_0,u_1)})$ les deux sous-problèmes suivants :

$\left(\mathcal{P}_2'^{(u_0,u_1)}\right)$:

$$\min_{\substack{\hat{\mathbf{r}}_\mathbf{0} \in \mathbb{R} \\ \hat{\mathbf{x}}_\mathbf{0} \in \mathcal{X}}} \quad \hat{\mathbf{r}}_\mathbf{0}$$

sous contraintes

$$\left|\hat{\mathbf{r}}_\mathbf{0} - r^{(u_0),k_0}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_\mathbf{0} - x^{(u_0),k_0}\right\|^2 , \forall k_0 \in \left\{1, \ldots, n^{(u_0)}\right\} ,$$

$$\hat{\mathbf{x}}_\mathbf{0} = x_0 .$$

$\left(\mathcal{P}_2''^{(u_0,u_1)}\right)$:

$$\min_{\substack{\hat{\mathbf{r}}_\mathbf{1} \in \mathbb{R} \\ \hat{\mathbf{x}}_\mathbf{1} \in \mathcal{X}}} \quad \hat{\mathbf{r}}_\mathbf{1} \tag{6}$$

sous contraintes

$$\left|\hat{\mathbf{r}}_\mathbf{1} - r^{(u_1),k_1}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_\mathbf{1} - x^{(u_1),k_1}\right\|^2 \quad ,\forall k_1 \in \left\{1,\ldots,n^{(u_1)}\right\} , \tag{7}$$

$$\left\|\hat{\mathbf{x}}_\mathbf{1} - y^{(u_0),k_0}\right\|^2 \leq L_f^2 \left\|x_0 - x^{(u_0),k_0}\right\|^2 \quad ,\forall k_0 \in \left\{1,\ldots,n^{(u_0)}\right\} . \tag{8}$$

Dans cette section, on montre tout d'abord qu'une solution optimale de $(\mathcal{P}_2^{(u_0,u_1)})$ peut être obtenue en résolvant les deux sous-problèmes $(\mathcal{P}_2'^{(u_0,u_1)})$ et $(\mathcal{P}_2''^{(u_0,u_1)})$ qui correspondent respectivement au premier et au deuxième pas de temps.

De manière immédiate, on peut remarquer que les étapes $t = 0$ et $t = 1$ sont théoriquement couplées par la contrainte (4), sauf dans le cas où les deux décisions u_0 et u_1 sont différentes pour lequel la contrainte (4) disparaît.

Par la suite, on prouve également que dans le cas $u_0 = u_1$, des solutions optimales aux deux problèmes $(\mathcal{P}_2'^{(u_0,u_1)})$ et $(\mathcal{P}_2''^{(u_0,u_1)})$ satisfont également la contrainte (4). On obtient également la solution du problème $(\mathcal{P}_2'^{(u_0,u_1)})$.

Théorème 1
Soit $(u_0, u_1) \in \mathcal{U}^2$. *Si* $(\hat{\mathbf{r}}_\mathbf{0}^*, \hat{\mathbf{x}}_\mathbf{0}^*)$ *est une solution optimale de* $\left(\mathcal{P}_2'^{(u_0,u_1)}\right)$ *et* $(\hat{\mathbf{r}}_\mathbf{1}^*, \hat{\mathbf{x}}_\mathbf{1}^*)$ *une solution optimale de* $\left(\mathcal{P}_2''^{(u_0,u_1)}\right)$, *alors* $(\hat{\mathbf{r}}_\mathbf{0}^*, \hat{\mathbf{r}}_\mathbf{1}^*, \hat{\mathbf{x}}_\mathbf{0}^*, \hat{\mathbf{x}}_\mathbf{1}^*)$ *est une solution optimale de* $\left(\mathcal{P}_2^{(u_0,u_1)}\right)$.

Dans la suite du papier, on se focalise sur les deux sous-problèmes $(\mathcal{P}_2'^{(u_0,u_1)})$ et $(\mathcal{P}_2''^{(u_0,u_1)})$ plutôt que $(\mathcal{P}_2^{(u_0,u_1)})$. A partir de la preuve du Théorème 1 donnée par Fonteneau *et al.* (2012), on obtient directement la solution du problème $(\mathcal{P}_2'^{(u_0,u_1)})$:

Corollaire 1
La solution du problème $(\mathcal{P}_2'^{(u_0,u_1)})$ *est* $\hat{\mathbf{r}}_\mathbf{0}^* = \max_{k_0 \in \left\{1,\ldots,n^{(u_0)}\right\}} r^{(u_0),k_0} - L_\rho \left\|x_0 - x^{(u_0),k_0}\right\|$.

4.2 Complexité du problème $(\mathcal{P}_2''^{(u_0,u_1)})$

Le problème $(\mathcal{P}_2'^{(u_0,u_1)})$ étant résolu, on s'intéresse maintenant à $(\mathcal{P}_2''^{(u_0,u_1)})$. En particulier, on montre que ce problème est NP-dur, même dans le cas particulier où il n'y a qu'un seul élément dans $\mathcal{F}^{(u_1)} = \left\{\left(x^{(u_1),1}, r^{(u_1),1}, y^{(u_1),1}\right)\right\}$. Dans ce cas particulier, le problème $(\mathcal{P}_2''^{(u_0,u_1)})$ consiste à maximiser la distance $\left\|\hat{\mathbf{x}}_\mathbf{1} - x^{(u_1),1}\right\|$ sous une contrainte d'intersection de boules comme on le décrit dans le lemme suivant.

Lemme 1
Si le cardinal de $\mathcal{F}^{(u_1)}$ *vaut 1 :* $\mathcal{F}^{(u_1)} = \left\{\left(x^{(u_1),1}, r^{(u_1),1}, y^{(u_1),1}\right)\right\}$, *alors la solution optimale de* $\left(\mathcal{P}_2''^{(u_0,u_1)}\right)$ *vérifie* $\hat{\mathbf{r}}_\mathbf{1}^* = r^{(u_1),1} - L_\rho \left\|\hat{\mathbf{x}}_\mathbf{1}^* - x^{(u_1),1}\right\|$ *où* $\hat{\mathbf{x}}_\mathbf{1}^*$ *maximise* $\left\|\hat{\mathbf{x}}_\mathbf{1} - x^{(u_1),1}\right\|$ *sous la contrainte*

$$\left\|\hat{\mathbf{x}}_\mathbf{1} - y^{(u_0),k_0}\right\|^2 \leq L_f^2 \left\|x_0 - x^{(u_0),k_0}\right\|^2 , \qquad \forall \left(x^{(u_0),k_0}, r^{(u_0),k_0}, y^{(u_0),k_0}\right) \in \mathcal{F}^{(u_0)} .$$

A noter que si le cardinal $n^{(u_0)}$ de $\mathcal{F}^{(u_0)}$ vaut aussi 1, alors le problème $(\mathcal{P}_2^{(u_0,u_1)})$ peut être résolu de manière exacte (cf. corollaire 3). En revanche, dans le cas général, ce problème est NP-dur comme on le montre ci-dessous. On introduit d'abord le problème de décision MNBC ("Max Norm with Ball Constraints") :

Définition 4 (Problème de décision MNBC)
Etant donné $x^{(0)} \in \mathbb{Q}^d, y^i \in \mathbb{Q}^d, \gamma_i \in \mathbb{Q}, i \in \{1,\dots,I\}, C \in \mathbb{Q}$, *le problème MNBC consiste à déterminer s'il existe* $x \in \mathbb{R}^d$ *tel que* $\left\|x - x^{(0)}\right\|^2 \geq C$ *et* $\left\|x - y^i\right\|^2 \leq \gamma_i \ , \forall i \in \{1,\dots,I\}$.

Lemme 2
MNBC est NP-dur.

Il en découle les deux résultats suivants :

Corollaire 2
$\left(\mathcal{P}_2^{\prime\prime(u_0,u_1)}\right)$ *est NP-dur.*

Théorème 2
Le problème "deux-étapes" $\left(\mathcal{P}_2^{(u_0,u_1)}\right)$ *et le problème général* $(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \dots, u_{T-1}))$ *sont NP-durs.*

5 Schémas de relaxation pour le cas "deux-étapes"

On a montré que le cas "deux-étapes" avec seulement un élément dans $\mathcal{F}^{(u_1)}$ est NP-dur (sauf dans le cas où le cardinal $n^{(u_0)}$ de $\mathcal{F}^{(u_0)}$ vaut 1, et dans ce cas $(\mathcal{P}_2^{(u_0,u_1)})$ est soluble en temps polynomial comme le montre le corollaire 3). Il est dès lors peu probable d'obtenir un algorithme capable de résoudre le problème "deux-étapes" en temps polynomial (sauf si P=NP). La philosophie du problème de généralisation min max est d'obtenir une séquence de décisions avec garanties de performances, c'est pourquoi obtenir une solution approchée qui pourrait être une borne supérieure ne serait pas pertinent ici. On propose donc par la suite de mettre au point des schémas de relaxation permettant d'obtenir des bornes inférieures sur le retour des séquences de décisions.

Le premier schéma de relaxation se base sur le principe d'éliminer certaines contraintes afin de déboucher sur une problème soluble en temps polynomial. On montre que ce schéma de relaxation mène à des bornes inférieures qui sont plus précises que celles fournies par CGRL dans Fonteneau *et al.* (2011). Le deuxième schéma de relaxation est une relaxation Lagrangienne où toutes les contraintes sont dualisées. Elle débouche sur un problème conique-quadratique qui peut être résolu en temps polynomial à l'aide de méthodes de point intérieur. On montre également que ce deuxième schéma offre une meilleure précision que le premier schéma, et par conséquent, une meilleure précision que CGRL. Enfin, on montre que les bornes calculées à partir des ces schémas convergent vers le vrai retour de la séquence (u_0, u_1) lorsque la dispersion de l'échantillon de trajectoires converge vers 0. Par conséquent, les séquences de décisions maximisant ces bornes convergent également vers des séquences de décisions optimales.

Depuis la section précédente, on sait que le problème deux-étapes $(\mathcal{P}_2^{(u_0,u_1)})$ peut être découplé en deux sous-problèmes $(\mathcal{P}_2^{\prime(u_0,u_1)})$ et $(\mathcal{P}_2^{\prime\prime(u_0,u_1)})$, où $(\mathcal{P}_2^{\prime(u_0,u_1)})$ se résout analytiquement (cf Théorème 1). On s'intéresse donc par la suite à des schémas de relaxation de $(\mathcal{P}_2^{\prime\prime(u_0,u_1)})$ que l'on rappelle ici :

$$\left(\mathcal{P}_2^{\prime\prime(u_0,u_1)}\right) : \min_{\substack{\hat{\mathbf{r}}_\mathbf{1} \in \mathbb{R} \\ \hat{\mathbf{x}}_\mathbf{1} \in \mathcal{X}}} \quad \hat{\mathbf{r}}_\mathbf{1}$$

$$\text{sous contraintes} \quad \left|\hat{\mathbf{r}}_\mathbf{1} - r^{(u_1),k_1}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_\mathbf{1} - x^{(u_1),k_1}\right\|^2 \quad \forall k_1 \in \left\{1,\dots,n^{(u_1)}\right\} \tag{9}$$

$$\left\|\hat{\mathbf{x}}_\mathbf{1} - y^{(u_0),k_0}\right\|^2 \leq L_f^2 \left\|x_0 - x^{(u_0),k_0}\right\|^2 \quad \forall k_0 \in \left\{1,\dots,n^{(u_0)}\right\} \tag{10}$$

5.1 Le schéma de relaxation par région de confiance

Une première approche de relaxation consiste à éliminer certaines contraintes. On suggère ici d'éliminer toute les contraintes (9) sauf une, indexée par k_1. De manière similaire, on laisse tomber toutes les contraintes (10) sauf une, indexée by k_0. Le problème suivant est ainsi une relaxation de $(\mathcal{P}_2^{\prime\prime(u_0,u_1)})$:

$\left(\mathcal{P}_{TR}^{\prime\prime(u_0,u_1)}(k_0,k_1)\right)$:

$$\begin{array}{c} \min \quad \hat{\mathbf{r}}_\mathbf{1} \\ \hat{\mathbf{r}}_\mathbf{1} \in \mathbb{R} \\ \hat{\mathbf{x}}_\mathbf{1} \in \mathcal{X} \end{array}$$

sous contraintes

$$\left|\hat{\mathbf{r}}_\mathbf{1} - r^{(u_1),k_1}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_\mathbf{1} - x^{(u_1),k_1}\right\|^2 , \tag{11}$$

$$\left\|\hat{\mathbf{x}}_\mathbf{1} - y^{(u_0),k_0}\right\|^2 \leq L_f^2 \left\|x_0 - x^{(u_0),k_0}\right\|^2 . \tag{12}$$

On a le résultat :

Théorème 3
Soit $B_{TR}^{\prime\prime(u_0,u_1),k_0,k_1}(\mathcal{F})$ la borne obtenue par la résolution de $(\mathcal{P}_{TR}^{\prime\prime(u_0,u_1)}(k_0,k_1))$. On a :

$$B_{TR}^{\prime\prime(u_0,u_1),k_0,k_1}(\mathcal{F}) = r^{(u_1),k_1} - L_\rho \left\|\hat{\mathbf{x}}_\mathbf{1}^*(k_0,k_1) - x^{(u_1),k_1}\right\|,$$

où

$$\hat{\mathbf{x}}_\mathbf{1}^*(k_0,k_1) \doteq y^{(u_0),k_0} + L_f \frac{\left\|x_0 - x^{(u_0),k_0}\right\|}{\left\|y^{(u_0),k_0} - x^{(u_1),k_1}\right\|} \left(y^{(u_0),k_0} - x^{(u_1),k_1}\right) \text{ if } y^{(u_0),k_0} \neq x^{(u_1),k_1}$$

et, si $y^{(u_0),k_0} = x^{(u_1),k_1}$, $\hat{\mathbf{x}}_\mathbf{1}^*(k_0,k_1)$ peut être n'importe quel point de la sphère centrée en $y^{(u_0),k_0} = x^{(u_1),k_1}$ et de rayon $L_f\|x_0 - x^{(u_0),k_0}\|$.

La preuve de ce résultat repose sur le fait que le problème $(\mathcal{P}_{TR}^{\prime\prime(u_0,u_1)}(k_0,k_1))$ est connu comme étant le sous-problème de la méthode par *région de confiance* (Conn *et al.* (2000)) soluble en temps polynomial (l'indice TR provient de l'anglais "trust region").

En considérant toutes les combinaisons (k_0,k_1) des contraintes non-éliminées, on obtient une famille de relaxations. Chaque combinaison aboutit à une valeur de borne inférieure, et on peut donc obtenir une borne inférieure maximale en cherchant la meilleur combinaison de contraintes. En notant $B_{TR}^{(u_0,u_1)}(\mathcal{F})$ le borne obtenue en sommant la solution du problème $(\mathcal{P}_2^{\prime(u_0,u_1)})$ et la borne donnée par la meilleure relaxation par région de confiance $(\mathcal{P}_2^{\prime\prime(u_0,u_1)})$ (considérant toutes les combinaisons (k_0,k_1)), on obtient :

Définition 5 (Borne région de confiance $B_{TR}^{(u_0,u_1)}(\mathcal{F})$)
$\forall (u_0,u_1) \in \mathcal{U}^2, \quad B_{TR}^{(u_0,u_1)}(\mathcal{F}) \triangleq \hat{\mathbf{r}}_\mathbf{0}^* + \max_{\substack{k_1 \in \{1,\ldots,n^{(u_1)}\} \\ k_0 \in \{1,\ldots,n^{(u_0)}\}}} B_{TR}^{\prime\prime(u_0,u_1),k_0,k_1}(\mathcal{F}).$

Dans le cas où $n^{(u_0)}$ et $n^{(u_1)}$ valent 1, la relaxation par région de confiance offre une solution exacte au problème deux-étapes $(\mathcal{P}_2^{(u_0,u_1)})$:

Corollaire 3
$\forall (u_0,u_1) \in \mathcal{U}^2, \qquad \left(\begin{cases} n^{(u_0)} = 1 \\ n^{(u_1)} = 1 \end{cases}\right) \implies B_{TR}^{(u_0,u_1)}(\mathcal{F}) = B_2^{(u_0,u_1)}(\mathcal{F}).$

5.2 La relaxation Lagrangienne

Une autre approche permettant d'obtenir une borne inférieure sur la valeur d'un problème de minimisation est de considérer une relaxation Lagrangienne. Dans cette section, on montre que la relaxation Lagrangienne du problème $(\mathcal{P}_2^{\prime\prime(u_0,u_1)})$ débouche sur un problème conique-quadratique. Concrètement, on multiplie les contraintes (9) par les variables duales $\mu_1,\ldots,\mu_{k_1},\ldots,\mu_{n^{(u_1)}} \geq 0$ et les contraintes (10) par les variables duales $\lambda_1,\ldots,\lambda_{k_0},\ldots,\lambda_{n^{(u_0)}} \geq 0$. On obtient ainsi le dual Lagrangien :

$\left(\mathcal{P}_{LD}^{\prime\prime(u_0,u_1)}\right)$:

$$\begin{array}{ll} \max & \min \quad \hat{\mathbf{r}}_{\mathbf{1}} \\ \lambda_1,\ldots,\lambda_{n^{(u_0)}} \in \mathbb{R}_+ & \hat{\mathbf{r}}_{\mathbf{1}} \in \mathbb{R} \\ \mu_1,\ldots,\mu_{n^{(u_1)}} \in \mathbb{R}_+ & \hat{\mathbf{x}}_{\mathbf{1}} \in \mathcal{X} \end{array}$$

$$+ \sum_{k_1=1}^{n^{(u_1)}} \mu_{k_1} \left(\left(\hat{\mathbf{r}}_{\mathbf{1}} - r^{(u_1),k_1} \right)^2 - L_\rho^2 \left\| \hat{\mathbf{x}}_{\mathbf{1}} - x^{(u_1),k_1} \right\|^2 \right)$$

$$+ \sum_{k_0=1}^{n^{(u_0)}} \lambda_{k_0} \left(\left\| \hat{\mathbf{x}}_{\mathbf{1}} - y^{(u_0),k_0} \right\|^2 - L_f^2 \left\| x_0 - x^{(u_0),k_0} \right\|^2 \right) . \qquad (13)$$

A noter que la valeur optimale de $(\mathcal{P}_{LD}^{\prime\prime(u_0,u_1)})$ est une borne inférieure de la valeur optimale de $(\mathcal{P}_2^{\prime\prime(u_0,u_1)})$ (Hiriart-Urruty & Lemaréchal (1996)). On a le résultat :

Théorème 4
$\left(\mathcal{P}_{LD}^{\prime\prime(u_0,u_1)}\right)$ *est un problème conique-quadratique.*

On introduit les notations suivantes :

Définition 6 (Notations)

$$M \triangleq \sum_{k_1=1}^{n^{(u_1)}} \mu_{k_1} \quad , \quad L \triangleq \sum_{k_0=1}^{n^{(u_0)}} \lambda_{k_0} \ ,$$

$$X \triangleq \left(x^{(u_1),1} \cdots x^{(u_1),n^{(u_1)}} \right) \quad , \quad Y \triangleq \left(y^{(u_0),1} \cdots y^{(u_0),n^{(u_0)}} \right) \ ,$$

$$\boldsymbol{\lambda} \triangleq \begin{pmatrix} \lambda_1 & \ldots & \lambda_{n^{(u_0)}} \end{pmatrix}^T \quad , \quad \boldsymbol{\mu} \triangleq \begin{pmatrix} \mu_1 & \ldots & \mu_{n^{(u_1)}} \end{pmatrix}^T \quad , \quad \bar{r} \triangleq \begin{pmatrix} r^{(1)} & \ldots & r^{(n^{(u_1)})} \end{pmatrix}^T \ ,$$

$\forall p \in \mathbb{N}_0, I_p$ *est une matrice identité de taille* p.

De la preuve du Théorème 4, on obtient le corollaire suivant :

Corollaire 4
$\forall (u_0,u_1) \in \mathcal{U}^2$,

$$B_{LD}^{\prime\prime(u_0,u_1)}(\mathcal{F}) \triangleq \max_{\boldsymbol{\lambda}\in\mathbb{R}_+^{n^{(u_0)}}, \boldsymbol{\mu}\in\mathbb{R}_+^{n^{(u_1)}}} \frac{-\|L_\rho^2 X\boldsymbol{\mu} - Y\boldsymbol{\lambda}\|^2}{-ML_\rho^2 + L} - \frac{\left(1 - 2\bar{r}^T\boldsymbol{\mu}\right)^2}{4M} \qquad (14)$$

$$+ \sum_{k_0=1}^{n^{(u_0)}} \lambda_{k_0} \left(\left\| y^{(u_0),k_0} \right\|^2 - L_f^2 \left\| x^{(u_0),k_0} - x_0 \right\|^2 \right) \qquad (15)$$

$$+ \sum_{k_1=1}^{n^{(u_1)}} \mu_{k_1} \left(\left(r^{(u_1),k_1} \right)^2 - L_\rho^2 \left\| x^{(u_1),k_1} \right\|^2 \right)$$

$$\begin{array}{lll} \textit{sous contraintes} & M & > \quad 0 \\ & L & > \quad ML_\rho^2 \end{array}$$

Dans la suite, on note $B_{LD}^{(u_0,u_1)}(\mathcal{F})$ la borne inférieure obtenue en sommant la solution de $(\mathcal{P}_2^{\prime(u_0,u_1)})$ et de la relaxation Lagrangienne de $(\mathcal{P}_2^{\prime\prime(u_0,u_1)})$:

Définition 7 (Borne Lagrangienne $B_{LD}^{(u_0,u_1)}(\mathcal{F})$)
$\forall (u_0,u_1) \in \mathcal{U}^2, \quad B_{LD}^{(u_0,u_1)}(\mathcal{F}) \triangleq \hat{\mathbf{r}}_{\mathbf{0}}^* + B_{LD}^{\prime\prime(u_0,u_1)}(\mathcal{F})$.

5.3 Comparaison des bornes

L'algorithme CGRL proposé par Fonteneau *et al.* (2010a, 2011) pour aborder le problème min max fait appel à une procédé initialement décrit dans Fonteneau *et al.* (2009) permettant de calculer des bornes inférieures sur le retour de politiques de décision à partir d'un échantillon de trajectoires. Concrètement, étant donnée une séquence de décisions $(u_0, u_1) \in \mathcal{U}^2$, le problème $(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1}))$ est remplacé par une borne $B_{CGRL}^{(u_0,u_1)}(\mathcal{F})$. On cherche maintenant à comparer cette borne dans le cas deux-étapes avec les deux nouvelles bornes sur $\left(\mathcal{P}_2^{(u_0,u_1)}\right)$ proposées dans ce papier : la borne région de confiance et la borne Lagrangienne.

On rappelle l'expression de la borne CGRL dans le cas "deux-étapes" (Fonteneau *et al.* (2011)) :

Définition 8 (Borne CGRL $B_{CGRL}^{(u_0,u_1)}(\mathcal{F})$)
$\forall (u_0, u_1) \in \mathcal{U}^2$,

$$B_{CGRL}^{(u_0,u_1)}(\mathcal{F}) \triangleq \max_{\substack{k_1 \in \{1,\ldots,n^{(u_1)}\} \\ k_0 \in \{1,\ldots,n^{(u_0)}\}}} r^{(u_0),k_0} - L_\rho(1+L_f)\left\|x^{(u_0),k_0} - x_0\right\| + r^{(u_1),k_1} - L_\rho\left\|y^{(u_0),k_0} - x^{(u_1),k_1}\right\| .$$

On a les résultats suivants :

Théorème 5
$\forall (u_0, u_1) \in \mathcal{U}^2, \quad B_{CGRL}^{(u_0,u_1)}(\mathcal{F}) \leq B_{TR}^{(u_0,u_1)}(\mathcal{F})$.

Théorème 6
$\forall\, (u_0, u_1) \in \mathcal{U}^2, \quad B_{TR}^{(u_0,u_1)}(\mathcal{F}) \leq B_{LD}^{(u_0,u_1)}(\mathcal{F})$.

Théorème 7
$\forall\, (u_0, u_1) \in \mathcal{U}^2, \quad B_{CGRL}^{(u_0,u_1)}(\mathcal{F}) \leq B_{TR}^{(u_0,u_1)}(\mathcal{F}) \leq B_{LD}^{(u_0,u_1)}(\mathcal{F}) \leq B_2^{(u_0,u_1)}(\mathcal{F}) \leq J_2^{(u_0,u_1)}$.

5.4 Convergence

On étudie maintenant la convergence des bornes et des séquences de décisions qui les maximisent lorsque la dispersion de l'échantillon de trajectoires décroît vers 0. On fait l'hypothèse que l'espace d'état est borné : $\exists C_{\mathcal{X}} > 0 : \forall (x, x') \in \mathcal{X}^2, \quad \|x - x'\| \leq C_{\mathcal{X}}$. On formalise la dispersion de l'échantillon de trajectoires :

Définition 9 (Dispersion de l'échantillon)
$\mathcal{X}$ étant borné, on a :

$$\exists\, \alpha > 0 : \forall u \in \mathcal{U}\ , \quad \sup_{x \in \mathcal{X}} \min_{k \in \left\{1,\ldots,n^{(u)}\right\}} \left\|x^{(u),k} - x\right\| \leq \alpha\ . \tag{16}$$

Le plus petit α vérifiant la relation (16) est nommé dispersion de $\mathcal{F}$, et noté $\alpha^*(\mathcal{F})$.

Intuitivement, la dispersion $\alpha^*(\mathcal{F})$ peut être vue comme le rayon de la plus grande zone de l'espace d'état non-couverte par l'échantillon de transitions.

Lemme 3
$\exists\, C > 0 : \forall (u_0, u_1) \in \mathcal{U}^2, \forall B^{(u_0,u_1)}(\mathcal{F}) \in \left\{B_{CGRL}^{(u_0,u_1)}(\mathcal{F}), B_{TR}^{(u_0,u_1)}(\mathcal{F}), B_{LD}^{(u_0,u_1)}(\mathcal{F})\right\}$,

$$J_2^{(u_0,u_1)} - B^{(u_0,u_1)}(\mathcal{F}) \leq C\alpha^*(\mathcal{F}).$$

S'ensuit le résultat suivant :

Théorème 8
$\forall (u_0,u_1) \in \mathcal{U}^2, \forall B^{(u_0,u1)}(\mathcal{F}) \in \left\{ B_{CGRL}^{(u_0,u_1)}(\mathcal{F}), B_{TR}^{(u_0,u_1)}(\mathcal{F}), B_{LD}^{(u_0,u_1)}(\mathcal{F}) \right\},$

$$\lim_{\alpha^*(\mathcal{F})\to 0} \quad J_2^{(u_0,u_1)} - B^{(u_0,u_1)}(\mathcal{F}) = 0 \ .$$

Dans la suite, on note $B_{CGRL}^{(*)}(\mathcal{F})$ (resp. $B_{TR}^{(*)}(\mathcal{F})$ et $B_{LD}^{(*)}(\mathcal{F})$) la borne CGRL maximale (resp. borne région de confiance maximale et borne Lagrangienne maximale) sur l'ensemble des séquences de décisions possibles, c'est-à-dire :

Définition 10 (Bornes maximales)

$$\begin{aligned} B_{CGRL}^{(*)}(\mathcal{F}) &\triangleq \max_{(u_0,u_1)\in\mathcal{U}^2} B_{CGRL}^{(u_0,u_1)}(\mathcal{F}) \ , \\ B_{TR}^{(*)}(\mathcal{F}) &\triangleq \max_{(u_0,u_1)\in\mathcal{U}^2} B_{TR}^{(u_0,u_1)}(\mathcal{F}) \ , \\ B_{LD}^{(*)}(\mathcal{F}) &\triangleq \max_{(u_0,u_1)\in\mathcal{U}^2} B_{LD}^{(u_0,u_1)}(\mathcal{F}) \ . \end{aligned}$$

On désigne également par $(u_0,u_1)_{\mathcal{F}}^{CGRL}$ (resp. $(u_0,u_1)_{\mathcal{F}}^{TR}$ et $(u_0,u_1)_{\mathcal{F}}^{LD}$) trois séquences de décisions maximisant les bornes :

Définition 11 (Séquences de décisions maximisant les bornes)

$$\begin{aligned} (u_0,u_1)_{\mathcal{F}}^{CGRL} &\in \left\{ (u_0,u_1)\in\mathcal{U}^2 | B_{CGRL}^{(u_0,u_1)}(\mathcal{F}) = B_{CGRL}^{(*)}(\mathcal{F}) \right\} \ , \\ (u_0,u_1)_{\mathcal{F}}^{TR} &\in \left\{ (u_0,u_1)\in\mathcal{U}^2 | B_{TR}^{(u_0,u_1)}(\mathcal{F}) = B_{TR}^{(*)}(\mathcal{F}) \right\} \ , \\ (u_0,u_1)_{\mathcal{F}}^{LD} &\in \left\{ (u_0,u_1)\in\mathcal{U}^2 | B_{LD}^{(u_0,u_1)}(\mathcal{F}) = B_{LD}^{(*)}(\mathcal{F}) \right\} \ . \end{aligned}$$

On donne un dernier théorème montrant la convergence des séquence de décisions $(u_0,u_1)_{\mathcal{F}}^{CGRL}$, $(u_0,u_1)_{\mathcal{F}}^{TR}$ et $(u_0,u_1)_{\mathcal{F}}^{LD}$ vers des séquences de décisions optimales - c'est-à-dire des séquences de décisions menant à un retour optimal J_2^* - lorsque la dispersion $\alpha^*(\mathcal{F})$ converge vers 0.

Théorème 9
Soit $\mathfrak{J}^*$ l'ensemble des séquences de décisions optimales : $\mathfrak{J}_2^* \triangleq \left\{ (u_0,u_1)\in\mathcal{U}^2 | J_2^{(u_0,u_1)} = J_2^* \right\}$, et supposons que $\mathfrak{J}_2^* \neq \mathcal{U}^2$ (si $\mathfrak{J}_2^* = \mathcal{U}^2$, la recherche d'une séquence de décisions optimale est triviale). On définit : $\epsilon \triangleq \min_{(u_0,u_1)\in\mathcal{U}^2\setminus\mathfrak{J}_2^*} \left\{ J_2^* - J_2^{(u_0,u_1)} \right\}$. On a alors
$\forall\, (\tilde{u}_0,\tilde{u}_1)_{\mathcal{F}} \in \left\{ (u_0,u_1)_{\mathcal{F}}^{CGRL}, (u_0,u_1)_{\mathcal{F}}^{TR}, (u_0,u_1)_{\mathcal{F}}^{LD} \right\}, \left(C\alpha^*(\mathcal{F}) < \epsilon \right) \implies (\tilde{u}_0,\tilde{u}_1)_{\mathcal{F}} \in \mathfrak{J}_2^* \ .$

Il est important de noter que la précision des bornes issues des schémas de relaxation proposés dans ce papier ne dépend pas *explicitement* de la dispersion (qui explose avec la dimension), mais dépend plutôt de l'état initial à partir duquel la séquence de décisions est prise et de la concentration locale de transitions autour de la trajectoire (inconnue) que prendrait le système piloté par la même séquence. On rencontrera donc en pratique souvent des cas de figure pour lesquels les bornes sont précises, alors que l'échantillon est très épars.

6 Résultats expérimentaux

On propose dans cette section quelques résultats expérimentaux illustrant les propriétés théoriques des trois bornes mentionnées dans ce papier (CGRL, région de confiance et Lagrangienne).

6.1 Problème-jouet

On considère le système linéaire suivant : $\forall (x,u) \in \mathcal{X} \times \mathcal{U}, \quad f(x,u) = x + 3.1416 \times u \times 1_d$, où $1_d \in \mathbb{R}^d$ désigne un vecteur de dimension d ne contenant que des 1. La fonction de récompense est définie comme suit : $\forall (x,u) \in \mathcal{X} \times \mathcal{U}, \qquad \rho(x,u) = \sum_{i=1}^{d} x(i)$, où $x(i)$ désigne la $i-$ème coordonnée de x. L'espace d'état $\mathcal{X}$ est inclus dans $\mathbb{R}^d$ et l'espace de décision fini vaut $\mathcal{U} = \{0, 0.1\}$. La dynamique du système f est $1-$Lipschitzienne et la fonction de récompense ρ est $\sqrt{d}-$Lipschitzienne. L'état initial du système est fixé à $x_0 = 0.5772 \times 1_d$. La dimension d de l'espace d'état vaut $d = 2$. Dans toutes nos expériences, la résolution de la relaxation Lagrangienne, qui est ici un problème conique-quadratique, est effectuée avec SeDuMi (Sturm (1999)).

6.2 Protocole et résultats

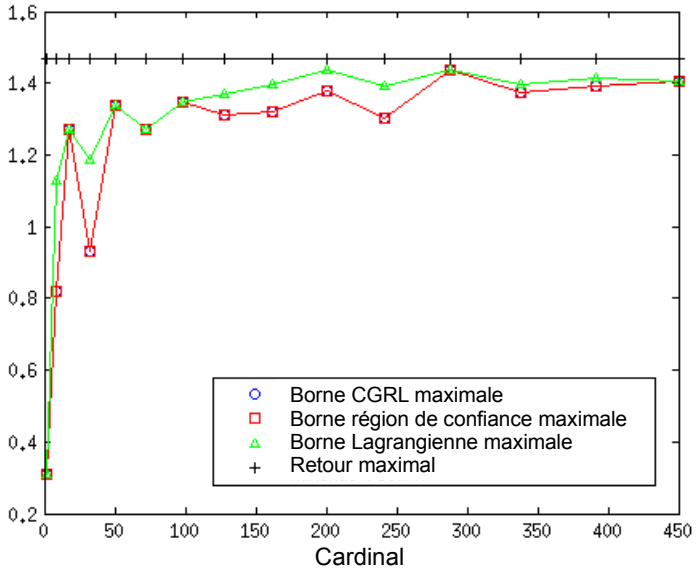

FIGURE 1 – Bornes $B^{(*)}_{CGRL}(\mathcal{F}_{c_i})$, $B^{(*)}_{TR}(\mathcal{F}_{c_i})$ et $B^{(*)}_{LD}(\mathcal{F}_{c_i})$ calculées à partir des échantillons de transitions $\mathcal{F}_{c_i} \quad i \in \{1, \ldots, 15\}$ de cardinal $c_i = 2i^2$.

6.2.1 Résultats typiques

Pour différents cardinaux $c_i = 2i^2, i = 1, \ldots, 15$, on génère un échantillon de transitions $\mathcal{F}_{c_i}$ en utilisant une grille uniforme sur l'espace $[0,1]^d \times \mathcal{U}$ comme suit : $\forall u \in \mathcal{U}$,

$$\mathcal{F}^{(u)}_{c_i} = \left\{ \left(\left[\frac{i_1}{i}; \frac{i_2}{i} \right], u, \rho\left(\left[\frac{i_1}{i}; \frac{i_2}{i} \right], u \right), f\left(\left[\frac{i_1}{i}; \frac{i_2}{i} \right], u \right) \right) \middle| (i_1, i_2) \in \{1, \ldots, i\}^2 \right\}$$

et

$$\mathcal{F}_{c_i} = \mathcal{F}^{(0)}_{c_i} \cup \mathcal{F}^{(.1)}_{c_i}$$

On donne à la figure 1 les valeurs de la borne CGRL maximale $B^{(*)}_{CGRL}(\mathcal{F}_{c_i})$, la borne région de confiance maximale $B^{(*)}_{TR}(\mathcal{F}_{c_i})$ et la borne Lagrangienne maximale $B^{(*)}_{LD}(\mathcal{F}_{c_i})$ en fonction du cardinal c_i des échantillons $\mathcal{F}_{c_i}$. On donne à la figure 2 les retours $J_2^{(u_0,u_1)^{CGRL}_{\mathcal{F}_{c_i}}}$, $J_2^{(u_0,u_1)^{TR}_{\mathcal{F}_{c_i}}}$ et $J_2^{(u_0,u_1)^{LD}_{\mathcal{F}_{c_i}}}$ des séquences de décisions maximisant les bornes $(u_0,u_1)^{CGRL}_{\mathcal{F}_{c_i}}$, $(u_0,u_1)^{TR}_{\mathcal{F}_{c_i}}$ et $(u_0,u_1)^{LD}_{\mathcal{F}_{c_i}}$.

On observe comme attendu que les bornes Lagrangiennes sont toujours supérieures ou égales aux bornes région de confiance, qui sont elle-même supérieure ou égales aux bornes CGRL, comme cela

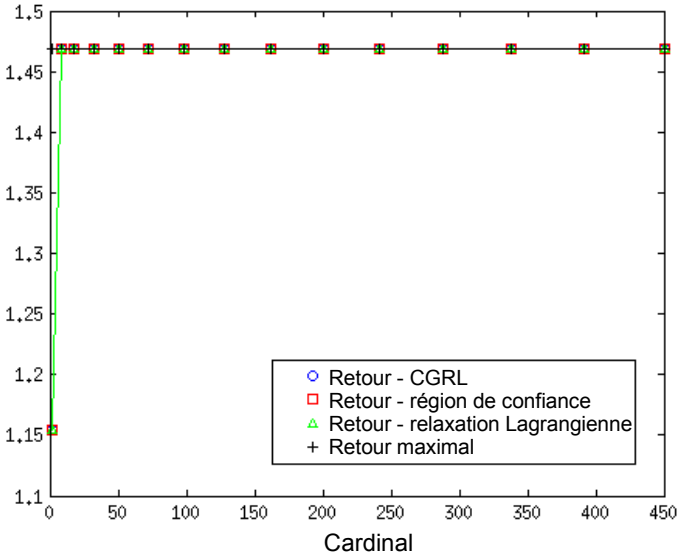

FIGURE 2 – Retours des séquences $(u_0, u_1)^{CGRL}_{\mathcal{F}_{c_i}}$, $(u_0, u_1)^{TR}_{\mathcal{F}_{c_i}}$ et $(u_0, u_1)^{LD}_{\mathcal{F}_{c_i}}$ calculées à partir des échantillons de transitions $\mathcal{F}_{c_i}$ $i \in \{1, \ldots, 15\}$ de cardinal $c_i = 2i^2$.

est décrit par le Théorème 7. En revanche, aucune différence n'est observée concernant les retours des séquences de décisions maximisant les bornes.

6.2.2 Échantillons de transitions tirés aléatoirement uniformément

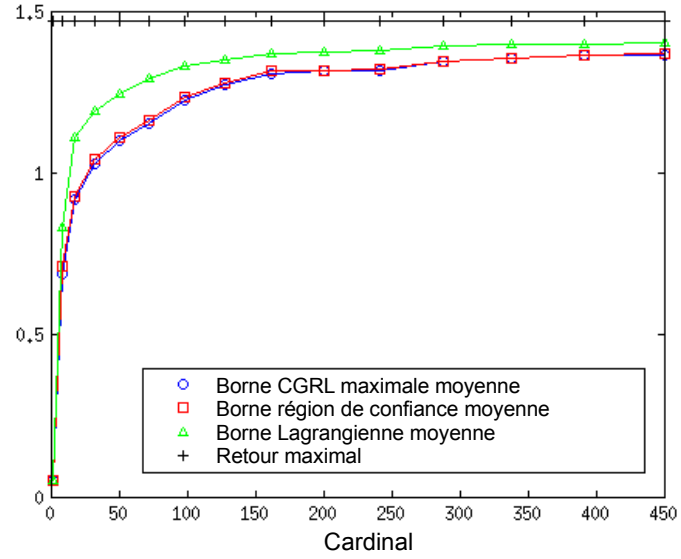

FIGURE 3 – Valeurs moyennes $A_{CGRL}(c_i)$, $A_{TR}(c_i)$ et $A_{LD}(c_i)$ des bornes calculées à partir des échantillons de trajectoires $\mathcal{F}_{c_i,k}$ $k \in \{1, \ldots, 100\}$ de cardinal $c_i = 2i^2$.

Dans le but d'observer l'influence de la dispersion de l'échantillon de transitions sur la préci-

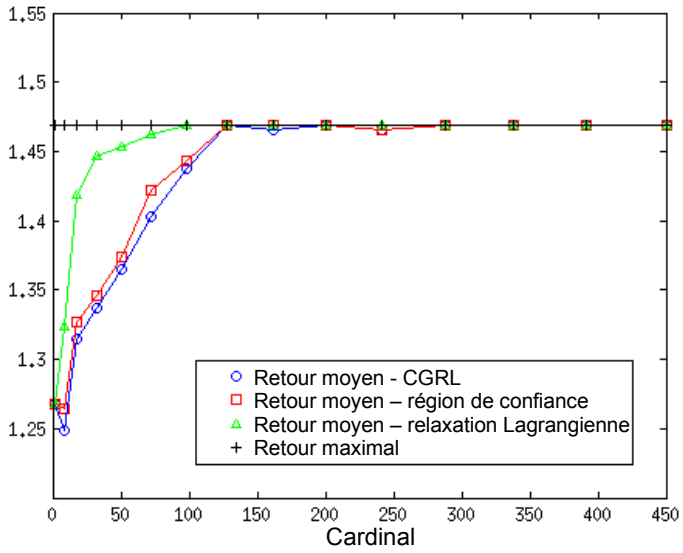

FIGURE 4 – Valeurs moyennes J_{CGRL}, J_{TR} and J_{LD} des retours des sequences de décisions maximisant les bornes calculées à partir des échantillons de trajectoires $\mathcal{F}_{c_i,k} \quad k \in \{1,\ldots,100\}$ de cardinal $c_i = 2i^2$.

sion des bornes, on propose le protocole suivant. Pour chaque valeur de cardinal $c_i = 2i^2, i = 1,\ldots,15$, on génère 100 échantillons de transitions $\mathcal{F}_{c_i,1},\ldots,\mathcal{F}_{c_i,100}$ en utilisant une distribution de probabilités uniforme sur l'espace $[0,1]^d \times \mathcal{U}$. Pour chaque échantillon de transitions $\mathcal{F}_{c_i,k} \quad i \in \{1,\ldots,15\}, k \in \{1,\ldots,100\}$, on calcule la borne CGRL maximale $B^{(*)}_{CGRL}(\mathcal{F}_{c_i,k})$, la borne région de confiance maximale $B^{(*)}_{TR}(\mathcal{F}_{c_i,k})$ et la borne Lagrangienne maximale $B^{(*)}_{LD}(\mathcal{F}_{c_i,k})$. On calcule ensuite les valeurs moyennes des bornes maximales : $\forall i \in \{1,\ldots,15\}, A_{CGRL}(c_i) = \frac{1}{100}\sum_{k=1}^{100} B^{(*)}_{CGRL}(\mathcal{F}_{c_i,k})\,, \quad A_{TR}(c_i) = \frac{1}{100}\sum_{k=1}^{100} B^{(*)}_{TR}(\mathcal{F}_{c_i,k})\,, \quad A_{LD}(c_i) = \frac{1}{100}\sum_{k=1}^{100} B^{(*)}_{LD}(\mathcal{F}_{c_i,k})$ et on donne à la figure 3 les valeurs de $A_{CGRL}(c_i)$ (resp. $A_{TR}(c_i)$ et $A_{LD}(c_i)$) en fonction du cardinal c_i de l'échantillon de transitions. On donne également à la figure 4 les valeurs moyennes des retours des séquences de décisions maximisant les bornes $(u_0,u_1)^{CGRL}_{\mathcal{F}_{c_i,k}}$, $(u_0,u_1)^{TR}_{\mathcal{F}_{c_i,k}}$ et $(u_0,u_1)^{LD}_{\mathcal{F}_{c_i,k}}$: $\forall i \in \{1,\ldots,15\}, J_{CGRL}(c_i) = \frac{1}{100}\sum_{k=1}^{100} J_2^{(u_0,u_1)^{CGRL}_{\mathcal{F}_{c_i,k}}}, J_{TR}(c_i) = \frac{1}{100}\sum_{k=1}^{100} J_2^{(u_0,u_1)^{TR}_{\mathcal{F}_{c_i,k}}}, J_{LD}(c_i) = \frac{1}{100}\sum_{k=1}^{100} J_2^{(u_0,u_1)^{LD}_{\mathcal{F}_{c_i,k}}}$ en fonction du cardinal c_i de l'échantillon de transitions.

On observe ainsi que, en moyenne, la borne Lagrangienne est plus précise que les bornes région de confiance et CGRL, alors que les bornes CGRL et région de confiance demeurent très proches. De plus, on observe que les séquences de décisions $(u_0,u_1)^{LD}_{\mathcal{F}_{c_i,k}}$ maximisant la borne Lagrangienne sont également plus performantes que les séquences de décisions maximisant les autres bornes.

7 Conclusions et perspectives d'améliorations

On s'est intéressé dans ce papier au problème de généralisation min max pour l'apprentissage par renforcement en mode batch dans un contexte déterministe et sous hypothèses de continuité Lipschitzienne. Dans un premier temps, on a montré que ce problème est NP-dur. On a ensuite proposé des schémas de relaxation pour le cas deux-étapes qui se sont montrés plus précis que l'approche initialement proposée par Fonteneau *et al.* (2011).

Une première perspective d'amélioration de ce travail consiste à étendre les schémas de relaxation proposés dans le cas où l'horizon d'optimisation est plus grand que 2 ($T \geq 3$). Les hypothèses de continuité Lipschitzienne sont désormais courantes en apprentissage par renforcement, mais on

pourrait imaginer des approches min max en généralisation sous d'autres types d'hypothèses. Enfin, il serait tout aussi intéressant d'étendre ces schémas de relaxation à des contextes où l'espace de décision est grand ou continu.

Remerciements

Raphael Fonteneau est Chargé de Recherches du F.R.S.-FNRS (Fonds de la Recherche Scientifique). Ce papier présente des résultats obtenus grâce au pôle d'attraction Inter-universitaire (PAI) belge DYSCO (Dynamical Systems, Control and Optimization) ainsi qu'au réseau européen d'excellence PASCAL2. Les auteurs remercient également Yurii Nesterov pour ses suggestions. La responsabilité scientifique demeure celle des auteurs.

Références

Banerjee A. & Tsiatis A. (2006). Adaptive two-stage designs in phase ii clinical trials. *Statistics in medicine*, **25**(19), 3382–3395.

Başar T. & Bernhard P. (1995). *H_∞-optimal control and related minimax design problems : a dynamic game approach*, volume 5. Birkhauser.

Bemporad A. & Morari M. (1999). Robust model predictive control : A survey. *Robustness in Identification and Control*, **245**, 207–226.

Bertsekas D. & Tsitsiklis J. (1996). *Neuro-Dynamic Programming*. Athena Scientific.

Birge J. & Louveaux F. (1997). *Introduction to Stochastic Programming*. Springer Verlag.

Bradtke S. & Barto A. (1996). Linear least-squares algorithms for temporal difference learning. *Machine Learning*, **22**, 33–57.

Busoniu L., Babuska R., De Schutter B. & Ernst D. (2010). *Reinforcement Learning and Dynamic Programming using Function Approximators*. Taylor & Francis CRC Press.

Camacho E. & Bordons C. (2004). *Model Predictive Control*. Springer.

Conn A., Gould N. & Toint P. (2000). *Trust-region Methods*, volume 1. Society for Industrial Mathematics.

Darby-Dowman K., Barker S., Audsley E. & Parsons D. (2000). A two-stage stochastic programming with recourse model for determining robust planting plans in horticulture. *Journal of the Operational Research Society*, p. 83–89.

Defourny B., Ernst D. & Wehenkel L. (2008). Risk-aware decision making and dynamic programming. *Selected for oral presentation at the NIPS-08 Workshop on Model Uncertainty and Risk in Reinforcement Learning, Whistler, Canada*.

Delage E. & Mannor S. (2010). Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, **58**(1), 203–213.

Ernst D., Geurts P. & Wehenkel L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, **6**, 503–556.

Ernst D., Glavic M., Capitanescu F. & Wehenkel L. (2009). Reinforcement learning versus model predictive control : a comparison on a power system problem. *IEEE Transactions on Systems, Man, and Cybernetics - Part B : Cybernetics*, **39**, 517–529.

Fonteneau R. (2011). *Contributions to Batch Mode Reinforcement Learning*. PhD thesis, University of Liège.

Fonteneau R., Ernst D., Boigelot B. & Louveaux Q. (2012). Min max generalization for deterministic batch mode reinforcement learning : relaxation schemes. *Arxiv preprint arXiv :1202.5298*.

Fonteneau R., Murphy S., Wehenkel L. & Ernst D. (2009). Inferring bounds on the performance of a control policy from a sample of trajectories. In *Proceedings of the 2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (IEEE ADPRL 09)*, Nashville, TN, USA.

Fonteneau R., Murphy S., Wehenkel L. & Ernst D. (2010a). A cautious approach to generalization in reinforcement learning. In *Proceedings of the Second International Conference on Agents and Artificial Intelligence (ICAART 2010)*, Valencia, Spain.

Fonteneau R., Murphy S. A., Wehenkel L. & Ernst D. (2010b). *Computing bounds for kernel-based policy evaluation in reinforcement learning*. Rapport interne, University of Liège.

Fonteneau R., Murphy S. A., Wehenkel L. & Ernst D. (2011). Towards min max generalization in reinforcement learning. In *Agents and Artificial Intelligence : International Conference, ICAART 2010, Valencia, Spain, January 2010, Revised Selected Papers. Series : Communications in Computer and Information Science (CCIS)*, volume 129, p. 61–77 : Springer, Heidelberg.

Frauendorfer K. (1992). *Stochastic Two-stage Programming*. Springer.

Hansen L. & Sargent T. (2001). Robust control and model uncertainty. *American Economic Review*, p. 60–66.

Hiriart-Urruty J. & Lemaréchal C. (1996). *Convex Analysis and Minimization Algorithms : Fundamentals*, volume 305. Springer-Verlag.

Ingersoll J. (1987). *Theory of Financial Decision Making*. Rowman and Littlefield Publishers, Inc.

Koenig S. (2001). Minimax real-time heuristic search. *Artificial Intelligence*, **129**(1-2), 165–197.

Lagoudakis M. & Parr R. (2003). Least-squares policy iteration. *Jounal of Machine Learning Research*, **4**, 1107–1149.

Littman M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on Machine Learning (ICML 1994)*, New Brunswick, NJ, USA.

Littman M. L. (2009). A tutorial on partially observable markov decision processes. *Journal of Mathematical Psychology*, **53**(3), 119 – 125. Special Issue : Dynamic Decision Making.

Lokhnygina Y. & Tsiatis A. (2008). Optimal two-stage group-sequential designs. *Journal of Statistical Planning and Inference*, **138**(2), 489–499.

Lunceford J., Davidian M. & Tsiatis A. (2002). Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, p. 48–57.

Mannor S., Simester D., Sun P. & Tsitsiklis J. (2004). Bias and variance in value function estimation. In *Proceedings of the Twenty-first International Conference on Machine Learning (ICML 2004)*, Banff, Alberta, Canada.

Murphy S. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B*, **65(2)**, 331–366.

Murphy S. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, **24**, 1455–1481.

Nemirovski A., Juditsky A., Lan G. & Shapiro A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, **19**(4), 1574–1609.

Ormoneit D. & Sen S. (2002). Kernel-based reinforcement learning. *Machine Learning*, **49**(2-3), 161–178.

Paduraru C., Precup D. & Pineau J. (2011). A framework for computing bounds for the return of a policy. In *Ninth European Workshop on Reinforcement Learning (EWRL9)*.

Qian M. & Murphy S. (2009). *Performance Guarantees for Individualized Treatment Rules*. Rapport interne 498, Department of Statistics, University of Michigan.

Riedmiller M. (2005). Neural fitted Q iteration - first experiences with a data efficient neural reinforcement learning method. In *Proceedings of the Sixteenth European Conference on Machine Learning (ECML 2005)*, p. 317–328, Porto, Portugal.

Rovatous M. & Lagoudakis M. (2010). Minimax search and reinforcement learning for adversarial tetris. In *Proceedings of the 6th Hellenic Conference on Artificial Intelligence (SETN'10)*, Athens, Greece.

Scokaert P. & Mayne D. (1998). Min-max feedback model predictive control for constrained linear systems. *IEEE Transactions on Automatic Control*, **43**(8), 1136–1142.

Shapiro A. (2011a). A dynamic programming approach to adjustable robust optimization. *Operations Research Letters*, **39**(2), 83–87.

Shapiro A. (2011b). *Minimax and Risk Averse Multistage Stochastic Programming*. Rapport interne, School of Industrial & Systems Engineering, Georgia Institute of Technology.

Sturm J. (1999). Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization methods and software*, **11**(1), 625–653.

Wahed A. & Tsiatis A. (2004). Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, **60**(1), 124–133.