Généralisation Min Max pour l'Apprentissage par Renforcement Batch et Déterministe : Relaxations pour le Cas Général T Etapes

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 : Cet article aborde le 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), et il a déjà été montré qu'il est NP-dur (Fonteneau *et al.* (2012b)). Deux schémas de relaxation pour le cas deux étapes ont été présentés aux JFPDA'12 (Fonteneau *et al.* (2012a)), et ce papier présente une généralisation de ces schémas au cas T étapes. Le premier schéma fonctionne en éliminant des contraintes afin d'obtenir un problème soluble en temps polynomial. Le deuxième schéma est une relaxation lagrangienne conduisant également à un problème soluble en temps polynomial. On montre théoriquement 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), que l'on sait déjà NP-dur (voir Fonteneau *et al.* (2012a,b))). On propose dans cet article 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. Ces schémas étendent au cas T étapes ceux qui étaient proposés aux JFPDA'12 et qui étaient valables dans le cas deux étapes (voir Fonteneau *et al.* (2012a,b). Le premier schéma de relaxation fonctionne par élimination de certaines contraintes afin de déboucher sur un problème soluble en temps polynomial. On obtient alors une configuration où à chaque pas de temps est associé 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 également soluble en temps polynomial. On montre que ces deux schémas permettent d'obtenir des bornes plus précises que CGRL.

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. La section 4 propose une simplification du problème, et rappelle son caractère NP-dur. La section 5 présente deux schémas de relaxation, ainsi que leurs propriétés théoriques. Quelques perspectives d'amélioration sont proposées en section 6.

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)).

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 d'une séquence de décisions)

$$\forall (u_0, \ldots, u_{T-1}) \in \mathcal{U}^T, \qquad J(u_0, \ldots, u_{T-1}) \triangleq \sum_{t=0}^{T-1} \rho(x_t, u_t) \ ,$$

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)

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

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 : $y^{(u),k} = f\left(x^{(u),k}, u\right)$ et $r^{(u),k} = \rho\left(x^{(u),k}, u\right)$.
3. Chaque ensemble $\mathcal{F}^{(u)}$ contient au moins un élément : $\forall u \in \mathcal{U}, \qquad n^{(u)} > 0$.

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 $J(\tilde{u}_0^*, \ldots, \tilde{u}_{T-1}^*)$ 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 :

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

$$\min_{\substack{\hat{\mathbf{r}}_0 \ \ldots \ \hat{\mathbf{r}}_{T-1} \in \mathbb{R} \\ \hat{\mathbf{x}}_0 \ \ldots \ \hat{\mathbf{x}}_{T-1} \in \mathcal{X}}} \sum_{t=0}^{T-1} \hat{\mathbf{r}}_t,$$

tel que

$$\left|\hat{\mathbf{r}}_t - r^{(u_t),k_t}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2 \ , \forall (t,k_t) \in \{0,\ldots,T-1\} \times \left\{1,\ldots,n^{(u_t)}\right\}, \quad (1)$$

$$\left\|\hat{\mathbf{x}}_{t+1} - y^{(u_t),k_t}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2, \forall (t,k_t) \in \{0,\ldots,T-1\} \times \left\{1,\ldots,n^{(u_t)}\right\}, \quad (2)$$

$$\left|\hat{\mathbf{r}}_t - \hat{\mathbf{r}}_{t'}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t'}\right\|^2 \ , \forall t,t' \in \{0,\ldots,T-1 | u_t = u_{t'}\}, \quad (3)$$

$$\left\|\hat{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}_{t'+1}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t'}\right\|^2 \ , \forall t,t' \in \{0,\ldots,T-2 | u_t = u_{t'}\}, \quad (4)$$

$$\hat{\mathbf{x}}_0 = x_0 \ . \quad (5)$$

Tout au long du papier, les variables d'optimisation seront écrites en caractères gras. On note $B^*(\mathcal{F}, u_0, \ldots, u_{T-1})$ la valeur optimale de la fonction objectif :

Définition 3 (Borne optimale $B^*(\mathcal{F}, u_0, \ldots, u_{T-1})$)
Soit $\hat{\mathbf{x}}_0^*, \ldots, \hat{\mathbf{x}}_{T-1}^*$ et $\hat{\mathbf{r}}_0^*, \ldots, \hat{\mathbf{r}}_{T-1}^*$ une solution optimale de $(\mathcal{P}(\mathcal{F}, u_0, \ldots, u_{T-1}))$. On définit :

$$B^*(\mathcal{F}, u_0, \ldots, u_{T-1}) = \sum_{t=0}^{T-1} \hat{\mathbf{r}}_t^*.$$

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

4 Simplification du problème et analyse de complexité

Dans cette section, on montre que les contraintes de type (3) ne sont pas nécessaires. Il en découle que dans le cas particulier à 2 étapes, c'est-à-dire le cas $T = 2$, les problèmes du temps 0 et du temps 1 sont découplés, et le problème du temps 1 est NP-dur.

4.1 Redondance de la contrainte (3)

On montre tout d'abord que les contraintes (3) ne sont pas nécessaires. En effet, on montre que toute solution optimale satisfaisant les autres contraintes, satisfait également les contraintes (3). Soit $\bar{\mathcal{P}}(\mathcal{F}, u_0, \ldots, u_{T-1})$ la relaxation de $\mathcal{P}(\mathcal{F}, u_0, \ldots, u_{T-1})$ où toutes les contraintes de type (3) sont relâchées.

Lemme 1
Soit $(\hat{\mathbf{r}}^*, \hat{\mathbf{x}}^*) \in \mathbb{R}^T \times \mathcal{X}^T$ *une solution optimale de* $\bar{\mathcal{P}}(\mathcal{F}, u_0, \ldots, u_{T-1})$*. Alors, pour tout* t, t' *tels que* $u_t = u_{t'}$*,*

$$|\hat{\mathbf{r}}_t^* - \hat{\mathbf{r}}_{t'}^*|^2 \leq L_\rho^2 \|\hat{\mathbf{x}}_t^* - \hat{\mathbf{x}}_{t'}^*\|^2 .$$

Preuve. Considérons une solution optimale de $\bar{\mathcal{P}}(\mathcal{F}, u_0, \ldots, u_{T-1})$. Observons que chaque variable $\hat{\mathbf{r}}_\mathbf{t}$ n'apparaît que dans les contraintes de type (1) :

$$\left|\hat{\mathbf{r}}_t - r^{(u_t),k_t}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2 \ , \forall (t,k_t) \in \{0,\ldots,T-1\} \times \left\{1,\ldots,n^{(u_t)}\right\} \tag{6}$$

Puisque la fonction objectif est $\min \sum_{t=0}^{T-1} \hat{\mathbf{r}}_\mathbf{t}$, on montre que, pour tout t, il existe au moins une contrainte (6) serrée. En supposant par contradiction que ce n'est pas le cas, on obtient une solution trivialement meilleure avec $\hat{\mathbf{r}}_\mathbf{t} - \epsilon$ pour un $\epsilon > 0$ bien choisi, ce qui est absurde. Il existe donc, pour tout t, un $\bar{k}_t$ tel que :

$$\hat{\mathbf{r}}_t^* = r^{(u_t),\bar{k}_t} - L_\rho \left\|\hat{\mathbf{x}}_t^* - x^{(u_t),\bar{k}_t}\right\| \tag{7}$$

Considérons maintenant une paire (t, t') telle que $u_t = u_{t'} = u$. On distingue deux cas en fonction du signe de $\hat{\mathbf{r}}_t^* - \hat{\mathbf{r}}_{t'}^*$.

– **Si** $\hat{\mathbf{r}}_t^* - \hat{\mathbf{r}}_{t'}^* \geq 0$
En utilisant (7) avec l'indice $\bar{k}_t^*$, on a :

$$\hat{\mathbf{r}}_t^* - \hat{\mathbf{r}}_{t'}^* \leq L_\rho \left(\left\|\hat{\mathbf{x}}_{t'}^* - x^{(u),\bar{k}_t^*}\right\| - \left\|\hat{\mathbf{x}}_t^* - x^{(u),\bar{k}_t^*}\right\|\right) \tag{8}$$

Etant donné que $\hat{\mathbf{r}}_t^* - \hat{\mathbf{r}}_{t'}^* \geq 0$, on a :

$$|\hat{\mathbf{r}}_t^* - \hat{\mathbf{r}}_{t'}^*| \leq L_\rho \left(\left\|\hat{\mathbf{x}}_{t'}^* - x^{(u),\bar{k}_t^*}\right\| - \left\|\hat{\mathbf{x}}_t^* - x^{(u),\bar{k}_t^*}\right\|\right) \ . \tag{9}$$

En utilisant l'inégalité triangulaire, on peut écrire :

$$\left\|\hat{\mathbf{x}}_{t'}^* - x^{(u),\bar{k}_t^*}\right\| \leq \|\hat{\mathbf{x}}_{t'}^* - \hat{\mathbf{x}}_t^*\| + \left\|\hat{\mathbf{x}}_t^* - x^{(u),\bar{k}_t^*}\right\| \ . \tag{10}$$

Remplaçant (10) dans (9), on obtient :

$$|\hat{\mathbf{r}}_{t'}^* - \hat{\mathbf{r}}_t^*| \leq L_\rho \|\hat{\mathbf{x}}_{t'}^* - \hat{\mathbf{x}}_t^*\|$$

ce qui prouve que $\hat{\mathbf{r}}_t^*$ et $\hat{\mathbf{r}}_{t'}^*$ satisfont la contrainte (3).

– **Si** $\hat{\mathbf{r}}_t^* - \hat{\mathbf{r}}_{t'}^* < 0$
En utilisant (6) avec l'indice $\bar{k}_{t'}^*$, on a :

$$\hat{\mathbf{r}}_{t'}^* - \hat{\mathbf{r}}_t^* \leq L_\rho \left(\left\|\hat{\mathbf{x}}_t^* - x^{(u),\bar{k}_{t'}^*}\right\| - \left\|\hat{\mathbf{x}}_{t'}^* - x^{(u),\bar{k}_{t'}^*}\right\|\right)$$

et, puisque $\hat{\mathbf{r}}_t^* - \hat{\mathbf{r}}_{t'}^* < 0$,

$$|\hat{\mathbf{r}}_{t'}^* - \hat{\mathbf{r}}_t^*| \leq L_\rho \left(\left\|\hat{\mathbf{x}}_t^* - x^{(u),\bar{k}_{t'}^*}\right\| - \left\|\hat{\mathbf{x}}_{t'}^* - x^{(u),\bar{k}_{t'}^*}\right\|\right). \tag{11}$$

En utilisant l'inégalité triangulaire, on peut écrire :

$$\left\|\hat{\mathbf{x}}_t^* - x^{(u),\bar{k}_{t'}^*}\right\| \leq \|\hat{\mathbf{x}}_t^* - \hat{\mathbf{x}}_{t'}^*\| + \left\|\hat{\mathbf{x}}_{t'}^* - x^{(u),\bar{k}_{t'}^*}\right\|. \tag{12}$$

En remplaçant (12) dans (11), on obtient :

$$|\hat{\mathbf{r}}_{t'}^* - \hat{\mathbf{r}}_t^*| \leq L_\rho \|\hat{\mathbf{x}}_t^* - \hat{\mathbf{x}}_{t'}^*\| \ ,$$

ce qui montre à nouveau que $\hat{\mathbf{r}}_t^*$ et $\hat{\mathbf{r}}_{t'}^*$ satisfont la contrainte (3).
Dans les deux cas $\hat{\mathbf{r}}_t^* - \hat{\mathbf{r}}_{t'}^* \geq 0$ et $\hat{\mathbf{r}}_t^* - \hat{\mathbf{r}}_{t'}^* < 0$, on a montré que la contrainte (3) est satisfaite. ■

On note que, d'après le Lemme 1, $\hat{\mathbf{r}}_0^*$ est découplée du reste du problème. Par conséquent, $\hat{\mathbf{r}}_0^*$ est solution du problème :

$(\mathcal{P}'(\mathcal{F}, u_0))$:

$$\begin{array}{l} \min \quad \hat{\mathbf{r}}_0 \\ \hat{\mathbf{r}}_0 \in \mathbb{R} \\ \hat{\mathbf{x}}_0 \in \mathcal{X} \end{array}$$

tel que

$$\left|\hat{\mathbf{r}}_0 - r^{(u_0),k_0}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_0 - x^{(u_0),k_0}\right\|^2 \ , \forall k_0 \in \left\{1, \ldots, n^{(u_0)}\right\} \ ,$$
$$\hat{\mathbf{x}}_0 = x_0 \ .$$

Lemme 2
La solution du problème $(\mathcal{P}'(\mathcal{F}, u_0))$ *est* $\hat{\mathbf{r}}_0^* = \max\limits_{k_0 \in \left\{1,\ldots,n^{(u_0)}\right\}} r^{(u_0),k_0} - L_\rho \left\|x_0 - x^{(u_0),k_0}\right\|$.

Preuve. Le résultat provient du fait que l'on minimise $\hat{\mathbf{r}}_0 \in \mathbb{R}$ sous des contraintes d'intervals. ■

Dans le cas particulier $T = 2$, le Lemme 1 implique que les deux étapes sont découplées. En particulier, le problème $\mathcal{P}(\mathcal{F}, u_0, u_1)$ peut être décomposé en deux sous-problèmes $(\mathcal{P}'(\mathcal{F}, u_0))$ et $(\mathcal{P}''(\mathcal{F}, u_0, u_1))$:

$(\mathcal{P}''(\mathcal{F}, u_0, u_1))$:

$$\begin{array}{l} \min \quad \hat{\mathbf{r}}_1 \\ \hat{\mathbf{r}}_1 \in \mathbb{R} \\ \hat{\mathbf{x}}_1 \in \mathcal{X} \end{array} \tag{13}$$

tel que

$$\left|\hat{\mathbf{r}}_1 - r^{(u_1),k_1}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_1 - x^{(u_1),k_1}\right\|^2 \ , \forall k_1 \in \left\{1, \ldots, n^{(u_1)}\right\} \ , \tag{14}$$
$$\left\|\hat{\mathbf{x}}_1 - y^{(u_0),k_0}\right\|^2 \leq L_f^2 \left\|x_0 - x^{(u_0),k_0}\right\|^2 \ , \forall k_0 \in \left\{1, \ldots, n^{(u_0)}\right\} \ . \tag{15}$$

4.2 Complexité de $(\mathcal{P}''(\mathcal{F}, u_0, u_1))$

Théorème 1
$(\mathcal{P}''(\mathcal{F}, u_0, u_1))$ *et* $(\mathcal{P}(\mathcal{F}, u_0, \ldots, u_{T-1}))$ *sont NP-durs.*

Ce théorème, déjà énoncé dans Fonteneau *et al.* (2012a), est démontré dans Fonteneau *et al.* (2012b). On note aussi que le caractère NP-dur de $(\mathcal{P}(\mathcal{F}, u_0, \ldots, u_{T-1}))$ n'implique pas nécessairement que déterminer une séquence d'actions maximisant la borne $B^*(\mathcal{F}, u_0, \ldots, u_{T-1})$ le soit également. En revanche, même dans les cas où trouver une telle séquence est simple, nous cherchons à connaître la valeur de la borne optimale associée à cette séquence, et c'est précisément ce problème qui est NP-dur.

5 Schémas de relaxation

Le problème $(\mathcal{P}(\mathcal{F}, u_0, \ldots, u_{T-1}))$ étant NP-dur (Fonteneau *et al.* (2012a,b)), il est improbable de trouver un algorithme polynomial permettant de le résoudre exactement (à moins que P = NP). Dès lors, on propose des schémas de relaxation solubles en temps polynomial. Etant donné que la philposophie du problème min max est d'obtenir une séquence de décisions avec des garanties en termes de performances, on ne propose que des schémas de relaxation menant au calcul de bornes inférieures sur le retour.

Le premier schéma de relaxation fonctionne en éliminant un certain nombre de contraintes afin d'obtenir un problème soluble en temps polynomial. On montre par ailleurs que ce schéma mène à des bornes inférieures supérieures ou égales à la borne CGRL proposée par Fonteneau *et al.* (2011). Le deuxième schéma de relaxation est fondé sur une relaxation lagrangienne pour laquelle toutes les contraintes sont dualisées. Le problème résultant est également soluble en temps polynomial en utilisant un méthode de point intérieur. On montre que ce deuxième schéma de relaxation donne de meilleures bornes que le premier schéma, et par conséquent, de meilleurs résultats que ceux obtenus dans Fonteneau *et al.* (2011).

On sait de la section précédente que le problème du premier pas de temps peut être résolu de façon immédiate (cf. Lemma 2). On s'intéresse donc plus précisément au problème correspondant aux pas de temps restants $(\mathcal{P}''(\mathcal{F}, u_0, \ldots, u_{T-1}))$:

$(\mathcal{P}''(\mathcal{F}, u_0, \ldots, u_{T-1}))$:

$$\begin{array}{c} \min \\ \hat{\mathbf{r}}_1 \quad \ldots \quad \hat{\mathbf{r}}_{T-1} \in \mathbb{R} \\ \hat{\mathbf{x}}_0 \quad \ldots \quad \hat{\mathbf{x}}_{T-1} \in \mathcal{X} \end{array} \quad \sum_{t=1}^{T-1} \hat{\mathbf{r}}_t,$$

tel que

$$\left|\hat{\mathbf{r}}_t - r^{(u_t),k_t}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2 \ ,\forall (t,k_t) \in \{1,\ldots,T-1\} \times \left\{1,\ldots,n^{(u_t)}\right\}, \tag{16}$$

$$\left\|\hat{\mathbf{x}}_{t+1} - y^{(u_t),k_t}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2, \forall (t,k_t) \in \{0,\ldots,T-1\} \times \left\{1,\ldots,n^{(u_t)}\right\}, \tag{17}$$

$$\left\|\hat{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}_{t'+1}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t'}\right\|^2 \ ,\forall t,t' \in \{0,\ldots,T-2 | u_t = u_{t'}\}, \tag{18}$$

$$\hat{\mathbf{x}}_0 = x_0 \ . \tag{19}$$

5.1 Le schéma par régions de confiance entrelacées (ITR)

Un stratégie naturelle pour obtenir une relaxation d'un problème d'optimisation est d'ignorer certaines contraintes. Un cas particulier de problèmes quadratiques quadratiquement contraints (QCQP) non-convex abordables est lorsqu'il n'y a qu'une seule contrainte quadratique. Ici, on propose de relâcher suffisamment de contraintes de telle sorte qu'on obtienne un problème abordable pour chaque pas de temps. On obtient un problème de type région de confiance pour chaque pas de temps, d'où provient l'origine du nom du schéma ITR (de l'anglais "Intertwined Trust-Region" pour "régions de confiance entrelacées").

Pour chaque $t \in \{0,\ldots,T-1\}$, on choisit $\bar{k}_t \in \{1,\ldots,n^{(u_t)}\}$. La relaxation est obtenue en supprimant toutes les contraintes du type (4) et en ne gardant qu'une seule contrainte par type et pas de temps (correspondant à l'indice $t \in \{0,\ldots,T-1\}$) . On obtient alors un problème relâché de la forme :

$\left(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1})\right)$:

$$\begin{array}{c} \min \\ \hat{\mathbf{r}}_1, \ldots, \hat{\mathbf{r}}_{T-1} \in \mathbb{R} \\ \hat{\mathbf{x}}_0, \ldots, \hat{\mathbf{x}}_{T-1} \in \mathcal{X} \end{array} \quad \sum_{t=1}^{T-1} \hat{\mathbf{r}}_t$$

tel que

$$\left|\hat{\mathbf{r}}_t - r^{(u_t),\bar{k}_t}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),\bar{k}_t}\right\|^2 \qquad t \in \{1,\ldots,T-1\} \tag{20}$$

$$\left\|\hat{\mathbf{x}}_t - y^{(u_{t-1}),\bar{k}_{t-1}}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_{t-1} - x^{(u_{t-1}),\bar{k}_{t-1}}\right\|^2 \qquad t \in \{1,\ldots,T-1\} \ . \tag{21}$$

$$\hat{\mathbf{x}}_0 = x_0 \tag{22}$$

Dans la suite, on donne la solution optimale de $\left(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1})\right)$ sous forme analytique. Cette solution est calculée par récurrence. On définit la famille de T problèmes d'op-

timisation $\left\{\left(\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, \ldots, u_j, \bar{k}_0, \ldots, \bar{k}_j)\right)\right\}_{j=0}^{j=T-1}$:

Définition 4

$\left(\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, \ldots, u_j, \bar{k}_0, \ldots, \bar{k}_j)\right)$:

$$\max_{\substack{\hat{\mathbf{r}}_1, \ldots, \hat{\mathbf{r}}_j \in \mathbb{R} \\ \hat{\mathbf{x}}_0, \ldots, \hat{\mathbf{x}}_j \in \mathcal{X}}} \left\| \hat{\mathbf{x}}_j - x^{(u_j), \bar{k}_j} \right\|$$

tel que

$$\left| \hat{\mathbf{r}}_t - r^{(u_t), \bar{k}_t} \right|^2 \leq L_\rho^2 \left\| \hat{\mathbf{x}}_t - x^{(u_i), \bar{k}_t} \right\|^2 \qquad t \in \{1, \ldots, j\} \tag{23}$$

$$\left\| \hat{\mathbf{x}}_t - y^{(u_{t-1}), \bar{k}_{t-1}} \right\|^2 \leq L_f^2 \left\| \hat{\mathbf{x}}_{t-1} - x^{(u_{t-1}), \bar{k}_{t-1}} \right\|^2 \qquad t \in \{1, \ldots, j\} \ . \tag{24}$$

$$\hat{\mathbf{x}}_0 = x_0 \tag{25}$$

L'initialisation de la récurrence est donnée par le lemme suivant :

Lemme 3
La solution optimale $D''_{ITR}(u_0, u_1, \bar{k}_0, \bar{k}_1)$ *du problème* $\left(\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, u_1, \bar{k}_0, \bar{k}_1)\right)$ *est donnée par*

$$D''_{ITR}(u_0, u_1, \bar{k}_0, \bar{k}_1) = \left\| \hat{\mathbf{x}}_1^*(\bar{k}_0, \bar{k}_1) - x^{(u_1), \bar{k}_1} \right\|,$$

où

$$\hat{\mathbf{x}}_1^*(\bar{k}_0, \bar{k}_1) \doteq y^{(u_0), \bar{k}_0} + L_f \frac{\left\| x_0 - x^{(u_0), \bar{k}_0} \right\|}{\left\| y^{(u_0), \bar{k}_0} - x^{(u_1), \bar{k}_1} \right\|} \left(y^{(u_0), \bar{k}_0} - x^{(u_1), \bar{k}_1} \right) \text{ si } y^{(u_0), \bar{k}_0} \neq x^{(u_1), \bar{k}_1}$$

et, si $y^{(u_0), \bar{k}_0} = x^{(u_1), \bar{k}_1}$, $\hat{\mathbf{x}}_1^*(\bar{k}_0, \bar{k}_1)$ *peut être tout point de la sphère centrée en* $y^{(u_0), \bar{k}_0} = x^{(u_1), \bar{k}_1}$ *et de rayon* $L_f \| x_0 - x^{(u_0), \bar{k}_0} \|$.

Preuve. Il s'agit de la maximisation d'une norme sous une contrainte de norme. Ce problème est connu et référencé dans la littérature sous le nom de problème de type région de confiance (de l'anglais *trust-region*, Conn *et al.* (2000)). Dans notre cas, la valeur optimale de $\hat{\mathbf{x}}_1$ - notée $\hat{\mathbf{x}}_1^*(\bar{k}_0, \bar{k}_1)$ - se situe sur la même droite que $x^{(u_1), \bar{k}_1}$ et $y^{(u_0), \bar{k}_0}$, avec $y^{(u_0), \bar{k}_0}$ situé entre $x^{(u_1), \bar{k}_1}$ et $\hat{\mathbf{x}}_1^*(\bar{k}_0, \bar{k}_1)$, la distance entre $y^{(u_0), \bar{k}_0}$ et $\hat{\mathbf{x}}_1^*(\bar{k}_0, \bar{k}_1)$ étant exactement égale à la distance entre x_0 et $x^{(u_0), \bar{k}_0}$. Une illustration est donnée en figure 1. ■

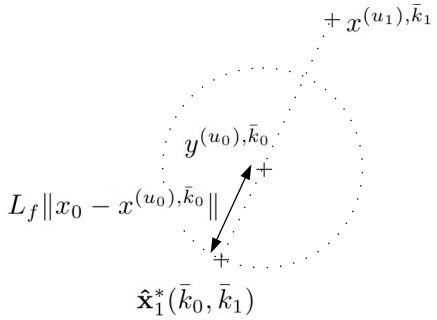

FIGURE 1 – Une approche géométrique simple pour résoudre $(\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, u_1, \bar{k}_0, \bar{k}_1))$.

Lemme 4
La solution optimale de $\left(\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, \ldots, u_j, \bar{k}_0, \ldots, \bar{k}_j)\right)$ *est donnée par :*

$$\begin{aligned}\forall t \in \{1, \ldots, j\}, \qquad & \hat{\mathbf{x}}^*_t(\bar{k}_0, \ldots, \bar{k}_t) \doteq y^{(u_{t-1}),\bar{k}_{t-1}} \\ & + L_f \frac{\left\| \hat{\mathbf{x}}^*_{t-1}(\bar{k}_0, \ldots, \bar{k}_{t-1}) - x^{(u_{t-1}),\bar{k}_{t-1}} \right\|}{\left\| y^{(u_{t-1}),\bar{k}_{t-1}} - x^{(u_t),\bar{k}_t} \right\|} \left(y^{(u_{t-1}),\bar{k}_{t-1}} - x^{(u_t),\bar{k}_t} \right) \\ & \text{si } y^{(u_{t-1}),\bar{k}_{t-1}} \neq x^{(u_t),\bar{k}_t}\end{aligned}$$

et, lorsque $y^{(u_{t-1}),\bar{k}_{t-1}} = x^{(u_t),\bar{k}_t}$, $\hat{\mathbf{x}}^*_t(\bar{k}_0, \ldots, \bar{k}_t)$ *peut être tout point de la sphère centrée en* $y^{(u_{t-1}),\bar{k}_{t-1}} = x^{(u_t),\bar{k}_t}$ *et de rayon* $L_f \| \hat{\mathbf{x}}^*_{t-1}(\bar{k}_0, \ldots, \bar{k}_{t-1}) - x^{(u_{t-1}),\bar{k}_{t-1}} \|$.

Preuve. La preuve est donnée par récurrence. L'initialisation de la récurrence est donnée par le Lemme 3. On suppose que l'hypothèse de récurrence est vraie pour le $(j-1)-$ème problème $\left(\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{j-1}, \bar{k}_0, \ldots, \bar{k}_{j-1})\right)$ et on montre que cela implique que l'hypothèse est vraie pour le $j-$ème problème. $\hat{\mathbf{x}}_j$ est contraint par une seule boule (24). Alors, si le membre de droite de l'équation (24) est fixé, la solution optimale de $\hat{\mathbf{x}}^*_j$ est induite par la même géométrie que dans le Lemme 3 (voir figure 1). Dès lors, il s'agit de maximiser le membre de droite de l'équation (24), ce qui revient à résoudre $\left(\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{j-1}, \bar{k}_0, \ldots, \bar{k}_{j-1})\right)$. Le résultat s'ensuit. ■

Théorème 2
La solution de $\left(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1})\right)$ *est :*

$$B''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1}) = \sum_{t=1}^{T-1} \hat{\mathbf{r}}^*_t$$

où

$$\begin{aligned}& \hat{\mathbf{r}}^*_t = r^{(u_t),\bar{k}_t} - L_\rho \left\| \hat{\mathbf{x}}^*_t(\bar{k}_0, \ldots, \bar{k}_t) - x^{(u_t),\bar{k}_t} \right\|, \\ & \hat{\mathbf{x}}^*_t(\bar{k}_0, \ldots, \bar{k}_t) \doteq y^{(u_{t-1}),\bar{k}_{t-1}} \\ & + L_f \frac{\left\| \hat{\mathbf{x}}^*_{t-1}(\bar{k}_0, \ldots, \bar{k}_{t-1}) - x^{(u_{t-1}),\bar{k}_{t-1}} \right\|}{\left\| y^{(u_{t-1}),\bar{k}_{t-1}} - x^{(u_t),\bar{k}_t} \right\|} \left(y^{(u_{t-1}),\bar{k}_{t-1}} - x^{(u_t),\bar{k}_t} \right) \\ & \text{si } y^{(u_{t-1}),\bar{k}_{t-1}} \neq x^{(u_t),\bar{k}_t}\end{aligned}$$

et, si $y^{(u_{t-1}),\bar{k}_{t-1}} = x^{(u_t),\bar{k}_t}$, $\hat{\mathbf{x}}^*_t(\bar{k}_0, \ldots, \bar{k}_t)$ *peut être n'importe quel point de la sphère centrée en* $y^{(u_{t-1}),\bar{k}_{t-1}} = x^{(u_t),\bar{k}_t}$ *et de rayon* $L_f \| \hat{\mathbf{x}}^*_{t-1}(\bar{k}_0, \ldots, \bar{k}_{t-1}) - x^{(u_{t-1}),\bar{k}_{t-1}} \|$.

Preuve. Observons d'abord que $\hat{\mathbf{r}}_t$ est contraint par un unique interval pour tout t. Par conséquent, comme on souhaite minimiser $\hat{\mathbf{r}}_t$, si le membre de droite de l'équation (20) est fixé, alors $\hat{\mathbf{r}}^*_t$ est donné par :

$$\hat{\mathbf{r}}^*_t = r^{(u_t),\bar{k}_t} - L_\rho \left\| \hat{\mathbf{x}}_t - x^{(u_t),\bar{k}_t} \right\|,$$

Afin de minimiser $\hat{\mathbf{r}}_t$, il est nécessaire de maximiser le membre de droite de (20), ce qui revient à résoudre $\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, \ldots, u_t, \bar{k}_0, \ldots, \bar{k}_t)$. Etant donné que la valeur de $\hat{\mathbf{x}}_j$ est la même pour toutes les solutions optimales des problèmes $\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, \ldots, u_i, \bar{k}_0, \ldots, \bar{k}_i)$ avec $i \geq j$, la valeur optimale de $\hat{\mathbf{x}}_t$ est donnée par la solution de $\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1})$ (voir Lemme 4), et le résultat s'ensuit. ■

La résolution de $(\mathcal{P}_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1}))$ aboutit à une famille de relaxations du problème initial en considérant toutes les combinaisons possibles $(\bar{k}_0, \ldots, \bar{k}_{T-1})$ de contraintes non relâchées. Choisir le maximum de toutes ces bornes inférieures donne une borne inférieure maximale pour ce type de schéma. En notant $B_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1})$ la borne obtenue en additionnant la solution du pas de temps 0 et la borne inférieure maximale ITR, on obtient :

Définition 5 (Borne régions de confiance entrelacées $B_{ITR}\left(\mathcal{F}, u_0, \ldots, u_{T-1}\right)$)

$$\begin{aligned} &B_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}) \triangleq \hat{\mathbf{r}}_0^* \\ &+ \max_{\substack{\bar{k}_{T-1} \in \{1,\ldots,n^{(u_{T-1})}\} \\ \ldots \\ \bar{k}_0 \in \{1,\ldots,n^{(u_0)}\}}} B''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1}). \end{aligned}$$

Notons que dans le cas où tous les $n^{(u_t)} \quad t = 0 \ldots T-1$ valent 1, la relaxation ITR donne une solution exacte du problème original $(\mathcal{P}(\mathcal{F}, u_0, \ldots, u_{T-1}))$:

Corollaire 1

$$\Big(\forall t \in \{0, \ldots, T-1\}, n^{(u_t)} = 1\Big) \implies B_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}) = B^*(\mathcal{F}, u_0, \ldots, u_{T-1}).$$

5.2 Relaxation lagrangienne

Une autre approche permettant d'obtenir une borne inférieure pour un problème de minimisation est d'effectuer une relaxation lagrangienne. Considérons de nouveau le problème $(\mathcal{P}''(\mathcal{F}, u_0, \ldots, u_{T-1}))$. En multipliant les contraintes (16) par les variables duales $\mu_{t,k_t} \geq 0$, les contraintes (17) par les variables duales $\lambda_{t,k_t} \geq 0$ et les contraintes (18) par les variables duales $\nu_{t,t'} \geq 0$, on obtient le dual lagrangien $(\mathcal{P}''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}))$:

$(\mathcal{P}''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}))$:

$$\begin{array}{cc} \max & \min \\ \nu_{t,t'} \in \mathbb{R} & \hat{\mathbf{r}}_1, \ldots, \hat{\mathbf{r}}_{T-1} \in \mathbb{R} \\ \lambda_{t,k_t} \in \mathbb{R} & \hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_{T-1} \in \mathcal{X} \\ \mu_{t,k_t} \in \mathbb{R} & \end{array}$$

$$\begin{aligned} &\hat{\mathbf{r}}_1 + \cdots + \hat{\mathbf{r}}_{T-1} + \\ &+ \sum_{(t,k_t) \in \{1,\ldots,T-1\} \times \left\{1,\ldots,n^{(u_t)}\right\}} \mu_{t,k_t} \left(\left| \hat{\mathbf{r}}_t - r^{(u_t),k_t} \right|^2 - L_\rho^2 \left\| \hat{\mathbf{x}}_t - x^{(u_t),k_t} \right\|^2 \right) \\ &+ \sum_{(t,k_t) \in \{1,\ldots,T-1\} \times \left\{1,\ldots,n^{(u_t)}\right\}} \lambda_{t,k_t} \left(\left\| \hat{\mathbf{x}}_{t+1} - y^{(u_t),k_t} \right\|^2 - L_f^2 \left\| \hat{\mathbf{x}}_t - x^{(u_t),k_t} \right\|^2 \right) \\ &+ \sum_{t,t' \in \{0,\ldots,T-2 | u_t = u_{t'}\}} \nu_{t,t'} \left(\|\hat{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}_{t'+1}\|^2 - L_f^2 \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t'}\|^2 \right) \end{aligned}$$

La valeur optimale de $(\mathcal{P}''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}))$ est connue pour être une borne inférieure sur la valeur optimale de $(\mathcal{P}''(\mathcal{F}, u_0, \ldots, u_{T-1}))$ (Hiriart-Urruty & Lemaréchal (1996)). Notons également que la relaxation lagrangienne donnée ci-dessus peut être résolue en temps polynomial, et est équivalente à une autre relaxation standard de problèmes QCQP qu'est la relaxation SDP (de l'anglais *semidefinite programming*). Il s'avère que ces deux relaxations sont duales l'une de l'autre (d'Aspremont & Boyd (2003); Nesterov *et al.* (2000)).

Définition 6 (Borne Lagrangienne $B_{LD}\left(\mathcal{F}, u_0, \ldots, u_{T-1}\right)$)
Soit $B''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1})$ l'optimum du dual lagrangien de $(\mathcal{P}''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}))$. On définit alors :

$$B_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}) = \mathbf{r}_0^* + B''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}) \ .$$

5.3 Comparaison des bornes

L'algorithme CGRL proposé par Fonteneau *et al.* (2010a, 2011) pour aborder le problème min max fait appel à la procédure décrite par Fonteneau *et al.* (2009) pour calculer une borne inférieure sur le retour de politiques de décision à partir d'un échantillon de transitions. Plus spécifiquement, étant donnée une séquence d'actions $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^2$, le problème $(\mathcal{P}(\mathcal{F}, u_0, \ldots, u_{T-1}))$ est remplacé par une borne $B_{CGRL}(\mathcal{F}, u_0, \ldots, u_{T-1})$. On cherche ici à savoir comment cette borne inférieure se classe parmi les deux nouvelles bornes proposées dans ce papier : la borne ITR et la borne lagrangienne.

5.3.1 ITR versus CGRL

On rappelle ici la définition de la borne CGRL (Fonteneau *et al.* (2010a)) :

Définition 7 (CGRL Bound $B_{CGRL}(\mathcal{F}, u_0, \ldots, u_{T-1})$)

$$
\begin{aligned}
&B_{CGRL}(\mathcal{F}, u_0, \ldots, u_{T-1}) \triangleq \\
&\max_{\substack{\bar{k}_{T-1} \in \{1,\ldots,n^{(u_{T-1})}\} \\ \ldots \\ \bar{k}_0 \in \{1,\ldots,n^{(u_0)}\}}} r^{(u_0),\bar{k}_0} - L_\rho \left(1 + L_f + L_f^2 + \ldots + L_f^{T-2}\right) \left\| x^{(u_0),\bar{k}_0} - x_0 \right\| \\
&\qquad + \ldots + \\
&\qquad + r^{(u_{T-2}),\bar{k}_{T-2}} - L_\rho \left(1 + L_f\right) \left\| y^{(u_{T-3}),\bar{k}_{T-3}} - x^{(u_{T-2}),\bar{k}_{T-2}} \right\| \\
&\qquad + r^{(u_{T-1}),\bar{k}_{T-1}} - L_\rho \left\| y^{(u_{T-2}),\bar{k}_{T-2}} - x^{(u_{T-1}),\bar{k}_{T-1}} \right\| .
\end{aligned}
$$

Le théorème suivant montre que la borne ITR est toujours supérieure ou égale à la borne CGRL.

Théorème 3

$$B_{CGRL}(\mathcal{F}, u_0, \ldots, u_{T-1}) \leq B_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}) \ .$$

Pour des raisons de place, la preuve de ce résultat est donnée dans Fonteneau *et al.* (2013). Elle se base sur le calcul par récurrence de $\left\| \hat{\mathbf{x}}_t^*(k_0^*, \ldots, k_t^*) - x^{(u_t),k_t^*} \right\|$ en fonction des distances $\left\| y^{(u_{t-1}),\bar{k}_{t-1}} - x^{(u_t),\bar{k}_t} \right\|, t = 1 \ldots T-1$.

5.3.2 Borne Lagrangienne versus borne ITR

Dans cette partie, on montre que la borne inférieure obtenue avec la relaxation lagragienne est toujours supérieure ou égale à la borne ITR. Pour ce faire, on montre qu'il y a dualité forte pour le dual lagrangien du problème $\left(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1})\right)$ pour un $\left(\bar{k}_0, \ldots, \bar{k}_{T-1}\right)$ fixé. Le dual lagrangien de $(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1}))$ s'écrit :

$$
\begin{aligned}
(LD''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1}))) : \quad & \max_{\substack{\lambda_1, \ldots, \lambda_{T-1} \in \mathbb{R} \\ \mu_1, \ldots, \mu_{T-1} \in \mathbb{R}}} \quad \min_{\substack{\hat{\mathbf{r}}_1, \ldots, \hat{\mathbf{r}}_{T-1} \in \mathbb{R} \\ \hat{\mathbf{x}}_0, \ldots, \hat{\mathbf{x}}_{T-1} \in \mathcal{X}}}
\end{aligned}
$$

$$
\begin{aligned}
&\hat{\mathbf{r}}_1 + \cdots + \hat{\mathbf{r}}_{T-1} + \\
&\mu_1 \left(\left| \hat{\mathbf{r}}_1 - r^{(u_1),\bar{k}_1} \right|^2 - L_\rho^2 \left\| \hat{\mathbf{x}}_1 - x^{(u_1),\bar{k}_1} \right\|^2 \right) \\
&\qquad \vdots \\
&+ \mu_{T-1} \left(\left| \hat{\mathbf{r}}_{T-1} - r^{(u_{T-1}),\bar{k}_{T-1}} \right|^2 - L_\rho^2 \left\| \hat{\mathbf{x}}_{T-1} - x^{(u_{T-1}),\bar{k}_{T-1}} \right\|^2 \right) \\
&+ \lambda_1 \left(\left\| \hat{\mathbf{x}}_1 - y^{(u_0),\bar{k}_0} \right\|^2 - L_f^2 \left\| \hat{\mathbf{x}}_0 - x^{(u_0),\bar{k}_0} \right\|^2 \right) \\
&\qquad \vdots \\
&+ \lambda_{T-1} \left(\left\| \hat{\mathbf{x}}_{T-1} - y^{(u_{T-2}),\bar{k}_{T-2}} \right\|^2 - L_f^2 \left\| \hat{\mathbf{x}}_{T-2} - x^{(u_{T-2}),\bar{k}_{T-2}} \right\|^2 \right)
\end{aligned}
$$

On considère maintenant le problème d'optimisation à l'intérieur du dual lagrangien. En considérant les λ_t, μ_t fixés, il peut être écrit comme une somme de termes monovariés, c'est à dire :

$$
\begin{aligned}
\min_{\substack{\hat{\mathbf{r}}_1,\ldots,\hat{\mathbf{r}}_{T-1} \in \mathbb{R} \\ \hat{\mathbf{x}}_0,\ldots,\hat{\mathbf{x}}_{T-1} \in \mathcal{X}}} & \sum_{t=1}^{T-1} \left(\hat{\mathbf{r}}_t + \mu_t \left| \hat{\mathbf{r}}_t - r^{(u_t),k_t} \right|^2 \right) \\
& + \sum_{t=1}^{T-1} \left(\left\| \hat{\mathbf{x}}_t - x^{(u_t),\bar{k}_t} \right\|^2 (-\mu_t L_\rho^2 - \lambda_{t+1} L_f^2) + \left\| \hat{\mathbf{x}}_t - y^{(u_{t-1}),\bar{k}_{t-1}} \right\|^2 \lambda_t \right) \qquad (26) \\
& + \lambda_1 L_f^2 \left\| \hat{x}_0 - x^{(u_0),\bar{k}_0} \right\|^2,
\end{aligned}
$$

avec $\lambda_T \triangleq 0$ par définition. On observe ainsi que la fonction objectif du problème d'optimisation (26) diverge vers $-\infty$ sauf si

$$
\forall t, \qquad \mu_t > 0 \text{ et } \begin{cases} \lambda_t - \mu_t L_\rho^2 - \lambda_{t+1} L_f^2 > 0 \\ \text{ou} \\ \lambda_t - \mu_t L_\rho^2 - \lambda_{t+1} L_f^2 = 0 \text{ avec } y^{(u_{t-1}),\bar{k}_{t-1}} = x^{(u_t),\bar{k}_t} \end{cases} \qquad (27)
$$

Le cas $\lambda_t - \mu_t L_\rho^2 - \lambda_{t+1} L_f^2 = 0$ avec $y^{(u_{t-1}),\bar{k}_{t-1}} = x^{(u_t),\bar{k}_t}$ provient du fait que si $y^{(u_{t-1}),\bar{k}_{t-1}} = x^{(u_t),\bar{k}_t}$, alors

$$
\begin{aligned}
&\left(\left\| \hat{\mathbf{x}}_t - x^{(u_t),\bar{k}_t} \right\|^2 (-\mu_t L_\rho^2 - \lambda_{t+1} L_f^2) + \left\| \hat{\mathbf{x}}_t - y^{(u_{t-1}),\bar{k}_{t-1}} \right\|^2 \lambda_t \right) \\
&\qquad = \left\| \hat{\mathbf{x}}_t - x^{(u_t),\bar{k}_t} \right\|^2 \left(\lambda_t - \mu_t L_\rho^2 - \lambda_{t+1} L_f^2 \right) \\
&\qquad = 0.
\end{aligned}
$$

Puisque le problème extérieur est une maximisation, on doit s'assurer que la condition (27) est satisfaite. A noter que la fonction objectif de (26) est une somme de fonctions monovariées, ce qui implique qu'on peut résoudre un problème d'optimisation pour chaque variables. On commence par les variables $\hat{\mathbf{r}}_t$.

Lemme 5
Soit $\mu_t > 0$. La solution du problème $\min_{\hat{\mathbf{r}}_t \in \mathbb{R}} \quad \hat{\mathbf{r}}_t + \mu_t \left| \hat{\mathbf{r}}_t - r^{(u_t),\bar{k}_t} \right|^2$ *est donnée par* $\hat{\mathbf{r}}_t^* = r^{(u_t),\bar{k}_t} - \frac{1}{2\mu_t}$.

Preuve. Cela provient directement du fait que l'on minimise une fonction quadratique monovariée :

$$
\mu_t \hat{\mathbf{r}}_t^2 + \hat{\mathbf{r}}_t \left(1 - 2\mu_t r^{(u_t),\bar{k}_t} \right) + \mu_t \left(r^{(u_t),\bar{k}_t} \right)^2.
$$

■

On s'intéresse maintenant aux problèmes impliquant les variables $\hat{\mathbf{x}}_t$, formellement définis comme suit :

$$(\mathcal{R}_t) : \min_{\hat{\mathbf{x}}_i \in \mathbb{R}^n} \left(\left\| \hat{\mathbf{x}}_t - x^{(u_t),\bar{k}_t} \right\|^2 \left(-\mu_t L_\rho^2 - \lambda_{t+1} L_f^2\right) + \left\| \hat{\mathbf{x}}_t - y^{(u_{t-1}),\bar{k}_{t-1}} \right\|^2 \lambda_t \right)$$

Lemme 6
Supposons que $x^{(u_t),\bar{k}_t} \neq y^{(u_{t-1}),\bar{k}_{t-1}}$. *La solution optimale* $\hat{\mathbf{x}}_t^*$ *de* $(\mathcal{R}_t)$ *se situe sur la même droite que* $x^{(u_t),\bar{k}_t}$ *et* $y^{(u_{t-1}),\bar{k}_{t-1}}$.

Preuve. On considère la projection orthogonale de $\hat{\mathbf{x}}_t^*$ sur $\text{aff}(x^{(u_t),\bar{k}_t}, y^{(u_{t-1}),\bar{k}_{t-1}})$ que l'on désigne par $\bar{x}_t$. On procède par l'absurde, et on suppose que $\hat{\mathbf{x}}_t^* \neq \bar{x}_t$. Par orthogonalité, on a :

$$\left\| \hat{\mathbf{x}}_t^* - x^{(u_t),\bar{k}_t} \right\|^2 = \|\hat{\mathbf{x}}_t^* - \bar{x}_t\|^2 + \left\| \bar{x}_t - x^{(u_t),\bar{k}_t} \right\|^2 \tag{28}$$

$$\left\| \hat{\mathbf{x}}_t^* - y^{(u_{t-1}),\bar{k}_{t-1}} \right\|^2 = \|\hat{\mathbf{x}}_t^* - \bar{x}_t\|^2 + \left\| \bar{x}_t - y^{(u_{t-1}),\bar{k}_{t-1}} \right\|^2. \tag{29}$$

Par conséquent, en substituant $\bar{x}_t$ dans la fonction objectif de $(\mathcal{R}_t)$, on obtient, à partir de (28) et (29),

$$\begin{aligned}
&\left(\left\| \hat{\mathbf{x}}_t^* - x^{(u_t),\bar{k}_t} \right\|^2 - \|\hat{\mathbf{x}}_t^* - \bar{x}_t\|^2 \right) \left(-\mu_t L_\rho^2 - \lambda_{t-1} L_f^2\right) \\
&\quad + \left(\left\| \hat{\mathbf{x}}_t^* - y^{(u_{t-1}),\bar{k}_{t-1}} \right\|^2 - \|\hat{\mathbf{x}}_t^* - \bar{x}_t\|^2 \right) \lambda_t \\
&\qquad = \left\| \hat{\mathbf{x}}_t^* - x^{(u_t),\bar{k}_t} \right\|^2 \left(-\mu_t L_\rho^2 - \lambda_{t+1} L_f^2\right) \\
&\qquad\quad + \left\| \hat{\mathbf{x}}_t^* - y^{(u_{t-1}),\bar{k}_{t-1}} \right\|^2 \lambda_i \\
&\qquad\quad - \|\hat{\mathbf{x}}_t^* - \bar{x}_t\|^2 \left(\lambda_t - \mu_t L_\rho^2 - \lambda_{t+1} L_f^2\right)
\end{aligned}$$

Les deux premiers termes et le terme de droite de la dernière équation correspondent à la valeur de la fonction objectif avec $\hat{\mathbf{x}}_t^*$ comme solution réalisable, et le dernier terme est toujours strictement négatif puisque $\lambda_t - \mu_t L_\rho^2 - \lambda_{t+1} L_f^2 > 0$. Par conséquent, $\bar{x}_t$ mène à un objectif inférieur à $\hat{\mathbf{x}}_t^*$, ce qui est absurde. ■

Lemme 7
Supposons que $x^{(u_t),\bar{k}_t} \neq y^{(u_{t-1}),\bar{k}_{t-1}}$. *Une solution optimale* $\hat{\mathbf{x}}_t^*$ *de* $(\mathcal{R}_t)$ *est telle que*

$$\left\| \hat{\mathbf{x}}_t^* - y^{(u_{t-1}),\bar{k}_{t-1}} \right\| = \frac{\left\| x^{(u_t),\bar{k}_t} - y^{(u_{t-1}),\bar{k}_{t-1}} \right\| \left(\mu_t L_\rho^2 + \lambda_{t+1} L_f^2\right)}{\lambda_t - \mu_t L_\rho^2 - \lambda_{t+1} L_f^2}$$

avec $\lambda_T = 0$ *par convention.*

Preuve. On connaît d'après le Lemme 6 la droite à laquelle $\hat{\mathbf{x}}_t^*$ appartient. Déterminer la solution optimale revient à calculer le minimum d'une fonction quadratique monovariée. Le calcul est laissé en exercice. ■

On a également le résultat :

Lemme 8
Supposons que $x^{(u_t),\bar{k}_t} = y^{(u_{t-1}),\bar{k}_{t-1}}$ *et* $\lambda_t - \mu_t L_\rho^2 - \lambda_{t+1} L_f^2 = 0$. *Alors la fonction objectif de* $(\mathcal{R}_t)$ *est identiquement nulle et* $\hat{\mathbf{x}}_t^*$ *peut être tout point de* $\mathcal{X}$.

On peut désormais prouver le résultat principal de cette section.

Théorème 4
Il y a dualité forte pour la relaxation lagrangienne du problème de type régions de confiance entrelacées ITR ($LD''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1})$).

Par soucis de concision, la preuve de ce résultat est donnée dans Fonteneau *et al.* (2013). Le principe de la preuve est de montrer, par récurrence, qu'il existe des $\lambda_1, \ldots, \lambda_{T-1}, \mu_1, \ldots, \mu_{T-1}$ satisfaisant les conditions données par l'Equation (27) et tels que la solution optimale correspondante pour le problème d'optimisation interne caractérisé par les Lemmes 5, 6, 7 et 8 est aussi une solution optimale de $(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1}))$. Comme toutes les contraintes y sont serrées, cela implique que la fonction objectif de la relaxation lagrangienne est égale à la fonction objectif du probème initial, ce qui prouve le résultat.

Théorème 5
$B_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}) \leq B_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1})$.

Preuve. Soit $\left(k_0^*, \ldots, k_{T-1}^*\right) \in \left\{1, \ldots, n^{(u_0)}\right\} \times \ldots \times \left\{1, \ldots, n^{(u_{T-1})}\right\}$ tel que :

$$B_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}) = \hat{\mathbf{r}}_0^* + B''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, k_0^*, \ldots, k_{T-1}^*).$$

Considérant $\left(\bar{k}_0, \ldots, \bar{k}_{T-1}\right) = \left(k_0^*, \ldots, k_{T-1}^*\right)$ dans le Théorème 4, on a :

$$B_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}) = \hat{\mathbf{r}}_0^* + B''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}, k_0^*, \ldots, k_{T-1}^*) \tag{30}$$

On observe alors que la relaxation lagrangienne $(LD''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, k_0^*, \ldots, k_{T-1}^*))$ - à partir de laquelle $B''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}, k_0^*, \ldots, k_{T-1}^*)$ est calculée - est aussi une relaxation du problème $(\mathcal{P}''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1})$ dans laquelle toutes les variables duales correspondant aux contraintes autres que celles en rapport avec la séquence de transitions

$$\left(x^{(u_0),k_0^*}, r^{(u_0),k_0^*}, y^{(u_0),k_0^*}\right) \ldots \left(x^{(u_{T-1}),k_{T-1}^*}, r^{(u_{T-1}),k_{T-1}^*}, y^{(u_{T-1}),k_{T-1}^*}\right)$$

seraient forcées à zero. Dès lors,

$$B''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}, k_0^*, \ldots, k_{T-1}^*) \leq B''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}) \ . \tag{31}$$

Par définition de la borne lagrangienne $B_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1})$, on a :

$$B_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}) = \hat{\mathbf{r}}_0^* + B''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}) \tag{32}$$

Les équations (30), (31) et (32) donnent finalement $B_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}) = B_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1})$. ■

5.3.3 Comparaison des bornes : résumé

On résume dans le théorème suivant les différents résultats obtenus dans les sections précédentes.

Théorème 6

$$\begin{aligned} \forall\, (u_0, \ldots, u_{T-1}) \in \mathcal{U}^T, B_{CGRL}(\mathcal{F}, u_0, \ldots, u_{T-1}) &\leq B_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}) \\ &\leq B_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}) \\ &\leq B^*(\mathcal{F}, u_0, \ldots, u_{T-1}) \\ &\leq J(u_0, \ldots, u_{T-1}) \ . \end{aligned}$$

Preuve. L'inégalité

$$B_{CGRL}(\mathcal{F}, u_0, \ldots, u_{T-1}) \leq B_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}) \leq B_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1})$$

est une conséquence directe des Théorèmes 3 et 5. L'inégalité

$$B_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}) \leq B^*(\mathcal{F}, u_0, \ldots, u_{T-1})$$

est une propriété de la relaxation lagrangienne, et l'inégalité $B^*(\mathcal{F}, u_0, \ldots, u_{T-1}) \leq J(u_0, \ldots, u_{T-1})$ provient de la définition de $B^*(\mathcal{F}, u_0, \ldots, u_{T-1})$. ■

5.3.4 Convergence

Le Théorème 6 implique que les propriétés de convergence de la borne CGRL lorsque la dispersion de l'échantillon de transitions tend vers 0 (voir Fonteneau *et al.* (2010a, 2011)) s'étendent aux nouvelles bornes présentées dans cet article.

6 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. Ce problème étant NP-dur, on a proposé des schémas de relaxation qui s'avèrent plus précis que l'approche initialement proposée par Fonteneau *et al.* (2011).

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. Ce papier présente des résultats obtenus grâce au pôle d'attraction Inter-universitaire (PAI) belge DYSCO. Les auteurs remercient également Yurii Nesterov, Benoit Daene et Adrien Hoarau pour leurs suggestions.

Références

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.

d'Aspremont A. & Boyd S. (2003). Relaxations and randomized methods for nonconvex qcqps. *EE392o Class Notes, Stanford University*.

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. (2012a). Généralisation min max pour l'apprentissage par renforcement batch et déterministe : schémas de relaxation. In *Journées Francophones de Planification, Décision et Apprentissage pour la conduite de systèmes (JFPDA)*.

Fonteneau R., Ernst D., Boigelot B. & Louveaux Q. (2012b). *Min max generalization for two-stage deterministic batch mode reinforcement learning : relaxation schemes*. Rapport interne, University of Liège.

Fonteneau R., Ernst D., Boigelot B. & Louveaux Q. (2013). Min max generalization for deterministic batch mode reinforcement learning : relaxation schemes. *Submitted*.

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

Fonteneau R., Murphy S., 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.

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.

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.

Nesterov Y., Wolkowicz H. & Ye Y. (2000). Semidefinite programming relaxations of nonconvex quadratic optimization. *Handbook of semidefinite programming*, p. 361–419.

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.