Relaxation Schemes for Min Max Generalization in Deterministic Batch Mode Reinforcement Learning

Raphael Fonteneau
University of Liège, Belgium
Raphael.Fonteneau@ulg.ac.be

Damien Ernst
University of Liège, Belgium
Dernst@ulg.ac.be

Bernard Boigelot
University of Liège, Belgium
Bernard.Boigelot@ulg.ac.be

Quentin Louveaux
University of Liège, Belgium
Q.Louveaux@ulg.ac.be

Abstract

We study the $\min\max$ optimization problem introduced in [6] for computing policies for batch mode reinforcement learning in a deterministic setting. This problem is NP-hard. We focus on the two-stage case for which we provide two relaxation schemes. The first relaxation scheme works by dropping some constraints in order to obtain a problem that is solvable in polynomial time. The second relaxation scheme, based on a Lagrangian relaxation where all constraints are dualized, leads to a conic quadratic programming problem. Both relaxation schemes are shown to provide better results than those given in [6].

1 Introduction

Research in Reinforcement Learning (RL) ([12]) aims at designing computational agents able to learn by themselves how to interact with their environment to maximize a numerical reward signal. The techniques developed in this field have appealed researchers trying to solve sequential decision making problems in many fields such as finance, medicine or engineering. Since the end of the nineties, several researchers have focused on the resolution of a subproblem of RL: computing a high-performance policy when the only information available on the environment is contained in a batch collection of trajectories of the agent ([2, 9, 10, 11]). This subfield of RL is known as "batch mode RL" [3].

Batch mode RL (BMRL) algorithms are challenged when dealing with large or continuous state spaces. Indeed, in such cases they have to generalize the information contained in a generally sparse sample of trajectories. The dominant approach for generalizing this information is to combine BMRL algorithms with function approximators ([1]). Usually, these approximators generalize the information contained in the sample to areas poorly covered by the sample by implicitly assuming that the properties of the system in those areas are similar to the properties of the system in the nearby areas well covered by the sample. This in turn often leads to low performance guarantees on the inferred policy. To overcome this problem, reference [6] proposes a $\min\max$-type strategy for generalizing in deterministic, Lipschitz continuous environments with continuous state spaces, finite action spaces, and finite time-horizon. The $\min\max$ approach works by determining a sequence of actions that maximizes the worst return that could possibly be obtained considering any system compatible with the sample of trajectories, and a weak prior knowledge given in the form of upper bounds on the Lipschitz constants related to the environment (dynamics, reward function). This problem is NP-hard, and reference [6] proposes an algorithm (called the CGRL algorithm - the acronym stands for "Cautious approach to Generalization in Reinforcement Learning") for computing an approximate solution in polynomial time. In this paper, we mainly focus on the 2-stage case for which we provide two relaxation schemes that are solvable in polynomial time and that provide better results than the CGRL algorithm.

2 Problem Formalization

Elements of Batch Mode Reinforcement Learning. We consider a deterministic discrete-time system whose dynamics over T stages is described by a time-invariant equation

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

where for all t, the state x_t is an element of the state space $\mathcal{X} \subset \mathbb{R}^d$ and u_t is an element of the finite (discrete) action space $\mathcal{U} = \left\{u^{(1)}, \ldots, u^{(m)}\right\}$ that we abusively identify with $\{1, \ldots, m\}$. $T \in \mathbb{N} \setminus \{0\}$ is referred to as the (finite) optimization horizon. An instantaneous reward

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

is associated with the action u_t taken while being in state x_t. For a given initial state $x_0 \in \mathcal{X}$ and for every sequence of actions $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$, the $T-$stage return is defined as follows:

$$J_T^{(u_0,\ldots,u_{T-1})} = \sum_{t=0}^{T-1} \rho\left(x_t, u_t\right) \ ,$$

where $x_{t+1} = f\left(x_t, u_t\right) \ , \forall t \in \{0, \ldots, T-1\}$.

We further make the following assumptions: (i) the system dynamics f and the reward function ρ are *unknown*, (ii) for each action $u \in \mathcal{U}$, a set of $n^{(u)} \in \mathbb{N}$ one-step system transitions

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

is known where each one-step transition is such that $y^{(u),k} = f\left(x^{(u),k}, u\right)$ and $r^{(u),k} = \rho\left(x^{(u),k}, u\right)$, and (iii) we assume that every set $\mathcal{F}^{(u)}$ contains at least one element. In the following, we denote by $\mathcal{F}$ the collection of all system transitions: $\mathcal{F} = \mathcal{F}^{(1)} \cup \ldots \cup \mathcal{F}^{(m)}$.

Min Max Generalization under Lipschitz Continuity Assumptions. The system dynamics f and the reward function ρ are assumed to be Lipschitz continuous:
$\exists L_f, L_\rho \in \mathbb{R}^+ : \forall (x, x') \in \mathcal{X}^2, \forall u \in \mathcal{U},$

$$\begin{aligned} \left\|f\left(x, u\right) - f\left(x', u\right)\right\| &\leq L_f \left\|x - x'\right\| \\ \left|\rho\left(x, u\right) - \rho\left(x', u\right)\right| &\leq L_\rho \left\|x - x'\right\| \end{aligned}$$

where $\|.\|$ denotes the Euclidean norm over the space $\mathcal{X}$. We also assume that two such constants L_f and L_ρ are known.

For a given sequence of actions, one can define the worst possible return that can be obtained by any system whose dynamics f' and ρ' would satisfy the Lipschitz inequalities and that would coincide with the values of the functions f and ρ given by the sample of system transitions $\mathcal{F}$. As shown in [6], this worst possible return can be computed by solving the following optimization problem:

$$(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1})) : \quad \min_{\substack{\mathbf{\hat{r}_0} \ \ldots \ \mathbf{\hat{r}_{T-1}} \in \mathbb{R} \\ \mathbf{\hat{x}_0} \ \ldots \ \mathbf{\hat{x}_{T-1}} \in \mathcal{X}}} \quad \sum_{t=0}^{T-1} \mathbf{\hat{r}_t},$$

subject to

$$\left|\mathbf{\hat{r}_t} - r^{(u_t),k_t}\right|^2 \leq L_\rho^2 \left\|\mathbf{\hat{x}_t} - x^{(u_t),k_t}\right\|^2 \ , \forall (t, k_t) \in \{0, \ldots, T-1\} \times \left\{1, \ldots, n^{(u_t)}\right\},$$
$$\left\|\mathbf{\hat{x}_{t+1}} - y^{(u_t),k_t}\right\|^2 \leq L_f^2 \left\|\mathbf{\hat{x}_t} - x^{(u_t),k_t}\right\|^2, \forall (t, k_t) \in \{0, \ldots, T-1\} \times \left\{1, \ldots, n^{(u_t)}\right\},$$
$$\left|\mathbf{\hat{r}_t} - \mathbf{\hat{r}_{t'}}\right|^2 \leq L_\rho^2 \left\|\mathbf{\hat{x}_t} - \mathbf{\hat{x}_{t'}}\right\|^2 \ , \forall t, t' \in \{0, \ldots, T-1 | u_t = u_{t'}\},$$
$$\left\|\mathbf{\hat{x}_{t+1}} - \mathbf{\hat{x}_{t'+1}}\right\|^2 \leq L_f^2 \left\|\mathbf{\hat{x}_t} - \mathbf{\hat{x}_{t'}}\right\|^2 \ , \forall t, t' \in \{0, \ldots, T-2 | u_t = u_{t'}\},$$
$$\mathbf{\hat{x}_0} = x_0 \ .$$

Note that optimization variables are written in bold. The min max approach to generalization aims at identifying which sequence of actions maximizes its worst possible return, that is which sequence of actions leads to the highest value of $(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1}))$. Since $\mathcal{U}$ is finite, we focus in this paper on a resolution scheme for solving this $\min\max$ problem that computes for each $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$ the value of its worst possible return.

3 The Two-stage Case

We now restrict ourselves to the case where $T = 2$, which is an important particular case of $\big(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1})\big)$. Given a two-stage sequence of actions $(u_0, u_1) \in \mathcal{U}^2$, the two-stage version of the problem $\big(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1})\big)$ reads as follows:

$$\big(\mathcal{P}_2(\mathcal{F}, L_f, L_\rho, x_0, u_0, u_1)\big): \quad \min_{\substack{\hat{\mathbf{r}}_{\mathbf{0}}, \hat{\mathbf{r}}_{\mathbf{1}} \in \mathbb{R} \\ \hat{\mathbf{x}}_{\mathbf{0}}, \hat{\mathbf{x}}_{\mathbf{1}} \in \mathcal{X}}} \quad \hat{\mathbf{r}}_{\mathbf{0}} + \hat{\mathbf{r}}_{\mathbf{1}},$$

$$\begin{aligned}
\text{subject to} \quad & \left|\hat{\mathbf{r}}_{\mathbf{0}} - r^{(u_0),k_0}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_{\mathbf{0}} - x^{(u_0),k_0}\right\|^2, \forall k_0 \in \left\{1, \ldots, n^{(u_0)}\right\}, \\
& \left|\hat{\mathbf{r}}_{\mathbf{1}} - r^{(u_1),k_1}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_{\mathbf{1}} - x^{(u_1),k_1}\right\|^2, \forall k_1 \in \left\{1, \ldots, n^{(u_1)}\right\}, \\
& \left\|\hat{\mathbf{x}}_{\mathbf{1}} - y^{(u_0),k_0}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_{\mathbf{0}} - x^{(u_0),k_0}\right\|^2, \forall k_0 \in \left\{1, \ldots, n^{(u_0)}\right\}, \\
& \left|\hat{\mathbf{r}}_{\mathbf{0}} - \hat{\mathbf{r}}_{\mathbf{1}}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_{\mathbf{0}} - \hat{\mathbf{x}}_{\mathbf{1}}\right\|^2 \text{ if } u_0 = u_1, \qquad (1) \\
& \hat{\mathbf{x}}_{\mathbf{0}} = x_0 .
\end{aligned}$$

For a matter of simplicity, we will refer $\big(\mathcal{P}_2(\mathcal{F}, L_f, L_\rho, x_0, u_0, u_1)\big)$ as $(\mathcal{P}_2^{(u_0,u_1)})$. We denote by $B_2^{(u_0,u_1)}(\mathcal{F})$ the lower bound associated with an optimal solution of $(\mathcal{P}_2^{(u_0,u_1)})$.
Let $(\mathcal{P}_2'^{(u_0,u_1)})$ and $(\mathcal{P}_2''^{(u_0,u_1)})$ be the two following subproblems:

$$(\mathcal{P}_2'^{(u_0,u_1)}): \quad \min_{\substack{\hat{\mathbf{r}}_{\mathbf{0}} \in \mathbb{R} \\ \hat{\mathbf{x}}_{\mathbf{0}} \in \mathcal{X}}} \quad \hat{\mathbf{r}}_{\mathbf{0}}$$

$$\begin{aligned}
\text{subject to} \quad & \left|\hat{\mathbf{r}}_{\mathbf{0}} - r^{(u_0),k_0}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_{\mathbf{0}} - x^{(u_0),k_0}\right\|^2, \forall k_0 \in \left\{1, \ldots, n^{(u_0)}\right\}, \\
& \hat{\mathbf{x}}_{\mathbf{0}} = x_0 .
\end{aligned}$$

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

$$\begin{aligned}
\text{subject to} \quad & \left|\hat{\mathbf{r}}_{\mathbf{1}} - r^{(u_1),k_1}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_{\mathbf{1}} - x^{(u_1),k_1}\right\|^2, \forall k_1 \in \left\{1, \ldots, n^{(u_1)}\right\}, \qquad (2) \\
& \left\|\hat{\mathbf{x}}_{\mathbf{1}} - y^{(u_0),k_0}\right\|^2 \leq L_f^2 \left\|x_0 - x^{(u_0),k_0}\right\|^2, \forall k_0 \in \left\{1, \ldots, n^{(u_0)}\right\}. \qquad (3)
\end{aligned}$$

We give hereafter a theorem that shows that an optimal solution to $(\mathcal{P}_2^{(u_0,u_1)})$ can be obtained by solving the two subproblems $(\mathcal{P}_2'^{(u_0,u_1)})$ and $(\mathcal{P}_2''^{(u_0,u_1)})$. Indeed, one can see that the stages $t = 0$ and $t = 1$ are theoretically coupled by constraint (1), except in the case where the two actions u_0 and u_1 are different for which $(\mathcal{P}_2^{(u_0,u_1)})$ is trivially decoupled. The following theorem shows that, even in the case $u_0 = u_1$, optimal solutions to the two decoupled problems $(\mathcal{P}_2'^{(u_0,u_1)})$ and $(\mathcal{P}_2''^{(u_0,u_1)})$ also satisfy constraint (1).

Theorem 1 *Let $(u_0, u_1) \in \mathcal{U}^2$. If $(\hat{\mathbf{r}}_{\mathbf{0}}^*, \hat{\mathbf{x}}_{\mathbf{0}}^*)$ is an optimal solution to $(\mathcal{P}_2'^{(u_0,u_1)})$ and $(\hat{\mathbf{r}}_{\mathbf{1}}^*, \hat{\mathbf{x}}_{\mathbf{1}}^*)$ is an optimal solution to $(\mathcal{P}_2''^{(u_0,u_1)})$, then $(\hat{\mathbf{r}}_{\mathbf{0}}^*, \hat{\mathbf{r}}_{\mathbf{1}}^*, \hat{\mathbf{x}}_{\mathbf{0}}^*, \hat{\mathbf{x}}_{\mathbf{1}}^*)$ is an optimal solution to $(\mathcal{P}_2^{(u_0,u_1)})$.*

The proof of this result is given in [4]. We now focus on $(\mathcal{P}_2'^{(u_0,u_1)})$ and $(\mathcal{P}_2''^{(u_0,u_1)})$, for which we have the two following propositions (proofs in [4]):

Proposition 2 *The solution of $(\mathcal{P}_2'^{(u_0,u_1)})$ is*

$$\hat{\mathbf{r}}_{\mathbf{0}}^* = \max_{k_0 \in \left\{1,\ldots,n^{(u_0)}\right\}} r^{(u_0),k_0} - L_\rho \left\| x_0 - x^{(u_0),k_0} \right\| .$$

Proposition 3 *In the general case, $(\mathcal{P}_2''^{(u_0,u_1)})$ is NP-hard.*

4 Relaxation Schemes for the Two-stage Case

We propose two relaxation schemes for $(\mathcal{P}_2''^{(u_0,u_1)})$ that are solvable in polynomial time and that are still leading to lower bounds on the actual return of the sequences of actions. The first relaxation scheme works by dropping some constraints. The second relaxation scheme is based on a Lagrangian relaxation where all constraints are dualized. Solving the Lagrangian dual is shown to be a conic quadratic problem that can be solved using interior-point methods.

4.1 The Trust-region Subproblem Relaxation Scheme

An easy way to obtain a relaxation from an optimization problem is to drop some constraints. We therefore suggest to drop all constraints (2) but one, indexed by k_1. Similarly we drop all constraints (3) but one, indexed by k_0. The following problem is therefore a relaxation of $(\mathcal{P}_2''^{(u_0,u_1)})$:

$$(\mathcal{P}_{TR}''^{(u_0,u_1)}(k_0,k_1)) : \quad \min_{\substack{\hat{\mathbf{r}}_{\mathbf{1}} \in \mathbb{R} \\ \hat{\mathbf{x}}_{\mathbf{1}} \in \mathcal{X}}} \quad \hat{\mathbf{r}}_{\mathbf{1}}$$

$$\text{subject to} \quad \left| \hat{\mathbf{r}}_{\mathbf{1}} - r^{(u_1),k_1} \right|^2 \leq L_\rho^2 \left\| \hat{\mathbf{x}}_{\mathbf{1}} - x^{(u_1),k_1} \right\|^2 ,$$
$$\left\| \hat{\mathbf{x}}_{\mathbf{1}} - y^{(u_0),k_0} \right\|^2 \leq L_f^2 \left\| x_0 - x^{(u_0),k_0} \right\|^2 .$$

We then have the following theorem:

Theorem 4 *The bound $B_{TR}''^{(u_0,u_1),k_0,k_1}(\mathcal{F})$ given by the resolution of $(\mathcal{P}_{TR}''^{(u_0,u_1)}(k_0,k_1))$ is*

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

where

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

and, if $y^{(u_0),k_0} = x^{(u_1),k_1}$, $\hat{\mathbf{x}}_{\mathbf{1}}^*(k_0,k_1)$ can be any point of the sphere centered in $y^{(u_0),k_0} = x^{(u_1),k_1}$ with radius $L_f \| x_0 - x^{(u_0),k_0} \|$.

The proof of this result is given in [4] and relies on the fact that $(\mathcal{P}_{TR}''^{(u_0,u_1)}(k_0,k_1))$ is equivalent to the $\max$ of a distance with a ball constraint. Solving $(\mathcal{P}_{TR}''^{(u_0,u_1)}(k_0,k_1))$ provides us with a family of relaxations for our initial problem by considering any combination (k_0,k_1) of two non-relaxed constraints. Taking the maximum out of these lower bounds yields the best possible bound out of this family of relaxations. The sum of the maximal Trust-region relaxation and the solution of $(\mathcal{P}_2'^{(u_0,u_1)})$ leads to the Trust-region bound:

Definition 5 (Trust-region Bound)

$$\forall (u_0, u_1) \in \mathcal{U}^2, \qquad B_{TR}^{(u_0,u_1)}(\mathcal{F}) = \hat{\mathbf{r}}_{\mathbf{0}}^* + \max_{\substack{k_1 \in \{1,\ldots,n^{(u_1)}\} \\ k_0 \in \{1,\ldots,n^{(u_0)}\}}} B_{TR}^{\prime\prime(u_0,u_1),k_0,k_1}(\mathcal{F}).$$

Notice that in the case where $n^{(u_0)}$ and $n^{(u_1)}$ are both equal to 1, then the trust-region relaxation scheme provides an exact solution of the original optimization problem $(\mathcal{P}_2^{(u_0,u_1)})$.

4.2 The Lagrangian Relaxation

Another way to obtain a lower bound on the value of a minimization problem is to consider a Lagrangian relaxation. If we multiply the constraints (2) by dual variables $\mu_1, \ldots, \mu_{k_1}, \ldots, \mu_{n^{(u_1)}} \geq 0$ and the constraints (3) by dual variables $\lambda_1, \ldots, \lambda_{k_0}, \ldots, \lambda_{n^{(u_0)}} \geq 0$, we obtain the Lagrangian dual:

$$\boxed{\begin{aligned}
(\mathcal{P}_{LD}^{\prime\prime(u_0,u_1)}) : \quad & \max_{\substack{\lambda_1,\ldots,\lambda_{n^{(u_0)}} \in \mathbb{R}_+ \\ \mu_1,\ldots,\mu_{n^{(u_1)}} \in \mathbb{R}_+}} \ \min_{\substack{\hat{\mathbf{r}}_{\mathbf{1}} \in \mathbb{R} \\ \hat{\mathbf{x}}_{\mathbf{1}} \in \mathcal{X}}} \quad \hat{\mathbf{r}}_{\mathbf{1}} \\
& + \sum_{k_1=1}^{n^{(u_1)}} \mu_{k_1} \left(\left(\hat{\mathbf{r}}_{\mathbf{1}} - r^{(u_1),k_1} \right)^2 - L_\rho^2 \left\| \hat{\mathbf{x}}_{\mathbf{1}} - x^{(u_1),k_1} \right\|^2 \right) \\
& + \sum_{k_0=1}^{n^{(u_0)}} \lambda_{k_0} \left(\left\| \hat{\mathbf{x}}_{\mathbf{1}} - y^{(u_0),k_0} \right\|^2 - L_f^2 \left\| x_0 - x^{(u_0),k_0} \right\|^2 \right) .
\end{aligned}}$$

Observe that the optimal value of $(\mathcal{P}_{LD}^{\prime\prime(u_0,u_1)})$ is known to provide a lower bound on the optimal value of $(\mathcal{P}_2^{\prime\prime(u_0,u_1)})$ ([8]). We have the following result (proof in [4]):

Theorem 6 *$(\mathcal{P}_{LD}^{\prime\prime(u_0,u_1)})$ is a conic quadratic program.*

The sum of the bound given by solution of $(\mathcal{P}_2^{\prime(u_0,u_1)})$ and the bound $B_{LD}^{\prime\prime(u_0,u_1)}(\mathcal{F})$ given by the resolution of the Lagrangian relaxation $(\mathcal{P}_{LD}^{\prime\prime(u_0,u_1)})$ leads to the Lagrangian relaxation bound:

Definition 7 (Lagrangian Relaxation Bound)

$$\forall (u_0, u_1) \in \mathcal{U}^2, \qquad B_{LD}^{(u_0,u_1)}(\mathcal{F}) = \hat{\mathbf{r}}_{\mathbf{0}}^* + B_{LD}^{\prime\prime(u_0,u_1)}(\mathcal{F})$$

4.3 Comparing the bounds

The CGRL algorithm proposed in [6] (initially introduced in [7]) for addressing the $\min\max$ problem uses the procedure described in [5] for computing a lower bound on the return of a policy given a sample of trajectories. More specifically, for a given sequence $(u_0, u_1) \in \mathcal{U}^2$, the program $(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1}))$ is replaced by a lower bound $B_{CGRL}^{(u_0,u_1)}(\mathcal{F})$. The following theorem (proof in [4]) shows how this bound compares in the two-stage case with the trust-region bound and the Lagrangian relaxation bound:

Theorem 8

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

Note that thanks to Theorem 8, the convergence properties of the CGRL bound (detailed in [7]) when the sparsity of $\mathcal{F}$ decreases towards zero also hold for the Trust-region and Lagrangian relaxation bounds.

5 Future Works

A natural extension of this work would be to investigate how the proposed relaxation schemes could be extended to the T-stage ($T \geq 3$) framework. Lipschitz continuity assumptions are common in a batch mode reinforcement learning setting, but one could imagine developing $\min\max$ strategies in other types of environments that are not necessarily Lipschitzian, or even not continuous. Additionally, it would also be interesting to extend the resolution schemes proposed in this paper to problems with very large/continuous action spaces.

Acknowledgments

Raphael Fonteneau is a postdoctoral fellow of the FRS-FNRS. This paper presents research results of the European Network of Excellence PASCAL2 and the Belgian Network DYSCO funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The authors thank Yurii Nesterov for pointing out the idea of using Lagrangian relaxation.

References

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

[2] D. Ernst, P. Geurts, and L. Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

[3] R. Fonteneau. *Contributions to Batch Mode Reinforcement Learning*. PhD thesis, University of Liège, 2011.

[4] R. Fonteneau, D. Ernst, B. Boigelot, and Q. Louveaux. Min max generalization for deterministic batch mode reinforcement learning: relaxation schemes. *Submitted*.

[5] R. Fonteneau, S. Murphy, L. Wehenkel, and D. Ernst. 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, 2009.

[6] R. Fonteneau, S. A. Murphy, L. Wehenkel, and D. Ernst. 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, pages 61–77. Springer, Heidelberg, 2011.

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

[8] J.B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms: Fundamentals*, volume 305. Springer-Verlag, 1996.

[9] M.G. Lagoudakis and R. Parr. Least-squares policy iteration. *Jounal of Machine Learning Research*, 4:1107–1149, 2003.

[10] D. Ormoneit and S. Sen. Kernel-based reinforcement learning. *Machine Learning*, 49(2-3):161–178, 2002.

[11] M. Riedmiller. 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)*, pages 317–328, Porto, Portugal, 2005.

[12] R.S. Sutton and A.G. Barto. *Reinforcement Learning*. MIT Press, 1998.