Lipschitz Robust Control from Off-Policy Trajectories

Raphael Fonteneau[†], Damien Ernst[†], Bernard Boigelot[†], Quentin Louveaux[†]

Abstract— **We study the $\min\max$ optimization problem introduced in [Fonteneau et al. (2011), "Towards min max reinforcement learning", Springer CCIS, vol. 129, pp. 61-77] for computing control policies for batch mode reinforcement learning in a deterministic setting with fixed, finite optimization horizon. First, we state that the $\min$ part of this problem is NP-hard. We then 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, can also be solved in polynomial time. We theoretically show that both relaxation schemes provide better results than those given in [Fonteneau et al. (2011)].**

I. Introduction

Research in Reinforcement Learning (RL) [1] 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 [2], Medicine [3], [4] or Engineering [5]. 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 [6], [7], [8], [9], [5], [10]. This subfield of RL is known as "batch mode RL".

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 [11], [8], [7], [12]. 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 when large state space areas are poorly covered by the sample. This can be explained by the fact that when computing the performance guarantees of these policies, one needs to take into account that they may actually drive the system into the poorly visited areas to which the generalization strategy associates a favorable environment behavior, while the environment may actually be particularly adversarial in those areas. This is corroborated by theoretical results which show that the performance guarantees of the policies inferred by these algorithms degrade with the sample dispersion where, loosely speaking, the dispersion can be seen as the radius of the largest non-visited state space area [13].

To overcome this problem, reference [14] 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). However, they show that finding an exact solution of the $\min\max$ problem is far from trivial, even after reformulating the problem so as to avoid the search in the space of all compatible functions. To circumvent these difficulties, they propose to replace, inside this $\min\max$ problem, the search for the worst environment given a sequence of actions by an expression that lower-bounds the worst possible return which leads to their so called CGRL algorithm (the acronym stands for "Cautious approach to Generalization in Reinforcement Learning"). This lower bound is derived from their previous work [15], [16] and has a tightness that depends on the sample dispersion. However, in some configurations where areas of the state space are not well covered by the sample of trajectories, the CGRL bound turns to be very conservative.

This paper - which is a shortened version of [18] - proposes to further investigate the $\min\max$ generalization optimization problem that was initially proposed in [14]. We first state that the $\min$ part of this optimization problem is NP-hard. Since it seems hopeless to exactly solve the problem, we propose two relaxation schemes that preserve the nature of the $\min\max$ generalization problem by targeting policies leading to high performance guarantees. The first relaxation scheme works by dropping some constraints in order to obtain a problem that is solvable in polynomial time for a given finite time horizon. This results into a configuration where each stage resorts in solving a *trust-region subproblem* [17]. The second relaxation scheme, based on a Lagrangian relaxation where all constraints are dualized, can be solved in polynomial time. We state that both relaxation schemes always provide bounds that are greater or equal to the CGRL bound. For conciseness reasons, proofs are not reported in this version (see [18]).

[†] Department of Electrical Engineering and Computer Science, University of Liège, 4000 Liège, Belgium{raphael.fonteneau, dernst, bernard.boigelot, q.louveaux}@ulg.ac.be

II. Related Work

Several works have already been built upon $\min\max$ paradigms for computing policies in a RL setting. In stochastic frameworks, $\min\max$ approaches are often successful for deriving robust solutions with respect to uncertainties in the (parametric) representation of the probability distributions associated with the environment [19]. In the context where several agents interact with each other in the same environment, $\min\max$ approaches appear to be efficient strategies for designing policies that maximize one agent's reward given the worst adversarial behavior of the other agents. [20], [21]. They have also received some attention for solving partially observable Markov decision processes [22], [23].

The $\min\max$ approach towards generalization, originally introduced in [14], implicitly relies on a methodology for computing lower bounds on the worst possible return (considering any compatible environment) in a deterministic setting with a mostly unknown actual environment. In this respect, it is related to other approaches that aim at computing performance guarantees on the returns of inferred policies [24], [25], [26].

Other fields of research have proposed $\min\max$-type strategies for computing control policies. This includes Robust Control theory [27] with H_∞ methods [28], but also Model Predictive Control (MPC) theory - where usually the environment is supposed to be fully known [29], [30] - for which $\min\max$ approaches have been used to determine an optimal sequence of actions with respect to the "worst case" disturbance sequence occurring [31], [32]. Finally, there is a broad stream of works in the field of Stochastic Programming [33] that have addressed the problem of safely planning under uncertainties, mainly known as "robust stochastic programming" or "risk-averse stochastic programming" [34], [35], [36], [37].

III. Problem Formalization

We first formalize the BMRL setting in Section III-A, and we state the $\min\max$ generalization problem in Section III-B.

A. Batch Mode Reinforcement Learning

We consider a deterministic discrete-time system whose dynamics over T stages is described by a time-invariant equation

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

where for all t, the state x_t is an element of the state space $\mathcal{X} \subset \mathbb{R}^d$ where $\mathbb{R}^d$ denotes the $d-$dimensional Euclidean space 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 $\{1, \ldots, m\}$. We assume that the (finite) optimization horizon $T \in \mathbb{N} \setminus \{0\}$ is a given (fixed) parameter of the problem. An instantaneous reward

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

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 cumulated reward over T stages (also named $T-$stage return) is defined as follows:

Definition 1 ($T-$stage Return):

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

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

An optimal sequence of actions is a sequence that leads to the maximization of the $T-$stage return:

Definition 2 (Optimal $T-$stage Return):

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

We further make the following assumptions that characterize the *batch mode setting*:

1) The system dynamics f and the reward function ρ are *unknown*;
2) 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:
$$y^{(u),k} = f\left(x^{(u),k}, u\right) \text{ and } r^{(u),k} = \rho\left(x^{(u),k}, u\right).$$
3) We assume that every set $\mathcal{F}^{(u)}$ contains at least one element: $\forall u \in \mathcal{U}, n^{(u)} > 0$.

In the following, we denote by $\mathcal{F}$ the collection of all system transitions:

$$\mathcal{F} = \mathcal{F}^{(1)} \cup \ldots \cup \mathcal{F}^{(m)}. \quad (1)$$

Under those assumptions, batch mode reinforcement learning (BMRL) techniques propose to infer from the sample of one-step system transitions $\mathcal{F}$ a high-performance sequence of actions, i.e. a sequence of actions $\left(\tilde{u}_0^*, \ldots, \tilde{u}_{T-1}^*\right) \in \mathcal{U}^T$ such that $J(\tilde{u}_0^*, \ldots, \tilde{u}_{T-1}^*)$ is as close as possible to J_T^*.

B. Min max Generalization under Lipschitz Continuity Assumptions

In this section, we state the $\min\max$ generalization problem that we study in this paper. The formalization was originally proposed in [14].

In all this paper, we assume that the system dynamics f and the reward function ρ are Lipschitz continuous, i.e. there exist finite constants $L_f, L_\rho \in \mathbb{R}$ such that $\forall (x, x') \in \mathcal{X}^2, \forall u \in \mathcal{U}$:

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

where $\|.\|$ denotes the Euclidean norm over the space $\mathcal{X}$. We also assume that two constants L_f and L_ρ satisfying the above-written inequalities are known. Such Lipschitz continuity assumptions are very standard in the field of batch mode reinforcement learning in continuous state spaces.

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 [14], this worst possible return can be computed by solving a finite-dimensional optimization problem over $\mathcal{X}^{T-1} \times \mathbb{R}^T$. Intuitively, solving such an optimization problem amounts to determining a most pessimistic trajectory of the system that is still compliant with the sample of data and the Lipschitz continuity assumptions. More specifically, for a given sequence of actions $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$, some given constants L_f and L_ρ, a given initial state $x_0 \in \mathcal{X}$ and a given sample of transitions $\mathcal{F}$, this optimization problem writes:

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

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

subject to

$$\left|\hat{\mathbf{r}}_t - r^{(u_t),k_t}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2,$$
$$\forall (t, k_t) \in \{0, \ldots, T-1\} \times \left\{1, \ldots, n^{(u_t)}\right\}, \tag{2}$$
$$\left\|\hat{\mathbf{x}}_{t+1} - y^{(u_t),k_t}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2,$$
$$\forall (t, k_t) \in \{0, \ldots, T-1\} \times \left\{1, \ldots, n^{(u_t)}\right\}, \tag{3}$$
$$\left|\hat{\mathbf{r}}_t - \hat{\mathbf{r}}_{t'}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t'}\right\|^2,$$
$$\forall t, t' \in \{0, \ldots, T-1 | u_t = u_{t'}\}, \tag{4}$$
$$\left\|\hat{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}_{t'+1}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t'}\right\|^2,$$
$$\forall t, t' \in \{0, \ldots, T-2 | u_t = u_{t'}\}, \tag{5}$$
$$\hat{\mathbf{x}}_0 = x_0. \tag{6}$$

For short, we refer to this problem as $(\mathcal{P}(\mathcal{F}, u_0, \ldots, u_{T-1}))$. Intuitively, the objective of the optimization problem modelizes the sum of rewards gathered along a trajectory $\hat{\mathbf{x}}_0, \ldots, \hat{\mathbf{x}}_{T-1}$. The idea of minimizing this objective comes from the fact that we want to find a most pessimistic trajectory. The constraints ensure that Lipschitz inequalities hold (i) between states / rewards from the pessimistic trajectory and states / rewards from the sample of data $\mathcal{F}$ and (ii) between states / rewards from different time-steps within the pessimistic trajectory. We also define the "optimal lower bound" $B^*(\mathcal{F}, u_0, \ldots, u_{T-1})$:

Definition 3 (Optimal lower bound $B^*(\mathcal{F}, u_0, \ldots, u_{T-1})$): Let $\hat{\mathbf{x}}_0^*, \ldots, \hat{\mathbf{x}}_{T-1}^*$ and $\hat{\mathbf{r}}_0^*, \ldots, \hat{\mathbf{r}}_{T-1}^*$ be an optimal solution to $(\mathcal{P}(\mathcal{F}, u_0, \ldots, u_{T-1}))$. We define the optimal lower bound $B^*(\mathcal{F}, u_0, \ldots, u_{T-1})$ as follows:

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

Note that, throughout the paper, optimization variables will be written in bold. The objective function represents the search for the most pessimistic trajectory. The constraints (2) and (4) (resp. (3) and (5)) express the fact that the reward function (resp. the system dynamics) must satisfy the Lipschitz inequalities for every pair of points from both the sample of data $\mathcal{F}$ and the pessimistic trajectory $(\hat{\mathbf{x}}_0, \hat{\mathbf{r}}_0, \ldots, \hat{\mathbf{x}}_{T-1}, \hat{\mathbf{r}}_{T-1})$. Constraint 6 ensures that the pessimistic trajectory starts in x_0.

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 $(\mathcal{P}(\mathcal{F}, u_0, \ldots, u_{T-1}))$.

We focus in this paper on the design of resolution schemes for solving the program $(\mathcal{P}(\mathcal{F}, u_0, \ldots, u_{T-1}))$. These schemes can afterwards be used for solving the $\min\max$ problem through exhaustive search over the set of all sequences of actions.

Later in this paper, we will also analyze the computational complexity of this min max generalization problem. When carrying out this analysis, we will assume that all the data of the problem (i.e., $T, \mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1}$) are given in the form of rational numbers.

IV. Analysis of the complexity

This section states that solving the $\min$ problem $(\mathcal{P}(\mathcal{F}, u_0, \ldots, u_{T-1}))$ is NP-hard. More precisely, we show that, in the case where $T = 2$, the problems of stage 0 and stage 1 are decoupled, and that the second stage problem is NP-hard.

A. Redundancy of constraint (4)

We first state that the constraints (4) are not needed. Indeed, in any optimal solution, they are always satisfied. Let $\bar{\mathcal{P}}\left(\mathcal{F}, u_0, \ldots, u_{T-1}\right)$ be the relaxation of $\mathcal{P}\left(\mathcal{F}, u_0, \ldots, u_{T-1}\right)$ where all constraints of type (4) are relaxed.

Lemma 1: Consider $(\hat{\mathbf{r}}^*, \hat{\mathbf{x}}^*) \in \mathbb{R}^T \times \mathcal{X}^T$ an optimal solution to $\bar{\mathcal{P}}\left(\mathcal{F}, u_0, \ldots, u_{T-1}\right)$. Then, for all t, t' such that $u_t = u_{t'}$,

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

Observe that Lemma 1 implies that $\hat{\mathbf{r}}_0^*$ is decoupled from the rest of the problem. Therefore, $\hat{\mathbf{r}}_0^*$ is the solution of:

$$(\mathcal{P}'(\mathcal{F}, u_0)) : \min_{\substack{\hat{\mathbf{r}}_0 \in \mathbb{R} \\ \hat{\mathbf{x}}_0 \in \mathcal{X}}} \hat{\mathbf{r}}_0$$

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

Lemma 2: The solution of the problem $(\mathcal{P}'(\mathcal{F}, u_0))$ is

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

In the particular case $T = 2$, Lemma 1 implies that the two stages are decoupled. In particular, the problem $\mathcal{P}(\mathcal{F}, u_0, u_1)$ can be decomposed in two subproblems $(\mathcal{P}'(\mathcal{F}, u_0))$ and $(\mathcal{P}''(\mathcal{F}, u_0, u_1))$:

$$(\mathcal{P}''(\mathcal{F}, u_0, u_1)) : \min_{\substack{\hat{\mathbf{r}}_1 \in \mathbb{R} \\ \hat{\mathbf{x}}_1 \in \mathcal{X}}} \quad \hat{\mathbf{r}}_1 \tag{7}$$

subject to

$$\left|\hat{\mathbf{r}}_1 - r^{(u_1),k_1}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_1 - x^{(u_1),k_1}\right\|^2$$
$$\forall k_1 \in \left\{1, \ldots, n^{(u_1)}\right\}, \tag{8}$$
$$\left\|\hat{\mathbf{x}}_1 - y^{(u_0),k_0}\right\|^2 \leq L_f^2 \left\|x_0 - x^{(u_0),k_0}\right\|^2$$
$$\forall k_0 \in \left\{1, \ldots, n^{(u_0)}\right\}. \tag{9}$$

B. Complexity of $(\mathcal{P}''(\mathcal{F}, u_0, u_1))$

The problem $(\mathcal{P}'(\mathcal{F}, u_0))$ being solved, we now focus in this section on the resolution of $(\mathcal{P}''(\mathcal{F}, u_0, u_1))$. In particular, we show that it is NP-hard, even in the particular case where there is only one element in the sample $\mathcal{F}^{(u_1)} = \left\{\left(x^{(u_1),1}, r^{(u_1),1}, y^{(u_1),1}\right)\right\}$. In this particular case, the problem $(\mathcal{P}''(\mathcal{F}, u_0, u_1))$ amounts to maximizing the distance $\left\|\hat{\mathbf{x}}_1 - x^{(u_1),1}\right\|$ under an intersection of balls as we show in the following lemma.

Lemma 3: If the cardinality of $\mathcal{F}^{(u_1)}$ is equal to 1:

$$\mathcal{F}^{(u_1)} = \left\{\left(x^{(u_1),1}, r^{(u_1),1}, y^{(u_1),1}\right)\right\},$$

then the optimal solution to $(\mathcal{P}''(\mathcal{F}, u_0, u_1))$ satisfies $\hat{\mathbf{r}}_1^* = r^{(u_1),1} - L_\rho \left\|\hat{\mathbf{x}}_1^* - x^{(u_1),1}\right\|$ where $\hat{\mathbf{x}}_1^*$ maximizes $\left\|\hat{\mathbf{x}}_1 - x^{(u_1),1}\right\|$ subject to

$$\left\|\hat{\mathbf{x}}_1 - y^{(u_0),k_0}\right\|^2 \leq L_f^2 \left\|x_0 - x^{(u_0),k_0}\right\|^2,$$

$\forall \left(x^{(u_0),k_0}, r^{(u_0),k_0}, y^{(u_0),k_0}\right) \in \mathcal{F}^{(u_0)}$.

Note that if the cardinality $n^{(u_0)}$ of $\mathcal{F}^{(u_0)}$ is also equal to 1, then $(\mathcal{P}(\mathcal{F}, u_0, u_1))$ can be solved exactly, as we will later show in Corollary 10. But, in the general case where $n^{(u_0)}$ is not fixed this problem of maximizing a distance under a set of ball-constraints is NP-hard as we state in Lemma 4. To do it, we introduce the MNBC (for "Max Norm with Ball Constraints") decision problem:

Definition 4 (MNBC Decision Problem): Given $x^{(0)} \in \mathbb{Q}^d, y^i \in \mathbb{Q}^d, \gamma_i \in \mathbb{Q}, i \in \{1, \ldots, I\}, C \in \mathbb{Q}$, the MNBC problem is to determine whether there exists $x \in \mathbb{R}^d$ such that $\left\|x - x^{(0)}\right\|^2 \geq C$ and $\left\|x - y^i\right\|^2 \leq \gamma_i$, $\forall i \in \{1, \ldots, I\}$.

Lemma 4: MNBC is NP-hard.

The MNBC problem amounts to maximizing the Euclidean norm of a vector over a finite intersection of spheres. Let us first mention that the problem of maximizing the norm of a vector over a finite intersection of concentric ellipsoids, which directly reduces to MNBC, is claimed to be NP-hard in [38] and [39], but without proof. Additionally, the complexity class of some related problems has already been investigated. In particular, it has been established that minimizing (or, equivalently, maximizing) a quadratic function under linear constraints is a NP-hard problem [40]. Furthermore, containment problems between polyhedra and spheres are known to be NP-hard as well [41]. However, those problems do not admit immediate reductions to MNBC. This motivates our development of a proof in [18] relying on a reduction from $\{0,1\}-$programming.

Note that the NP-hardness of MNBC is independent from the choice of the norm used over the state space $\mathcal{X}$. Also observe that, since $\{0,1\}-$programming is strongly NP-hard [42], it is also the case for MNBC. The two results follow:

Corollary 5: $(\mathcal{P}''(\mathcal{F}, u_0, u_1))$ is NP-hard.

Theorem 6: The two-stage problem $(\mathcal{P}(\mathcal{F}, u_0, u_1))$ and the generalized $T-$stage problem $(\mathcal{P}(\mathcal{F}, u_0, \ldots, u_{T-1}))$ are NP-hard.

Observe that the NP-hardness of $(\mathcal{P}(\mathcal{F}, u_0, \ldots, u_{T-1}))$ does not imply that finding a sequence of actions maximizing $B^*(\mathcal{F}, u_0, \ldots, u_{T-1})$ is also NP-hard. However, even for cases where finding such a sequence is easy, we are still interested in computing the value of the optimal lower bound associated with such a sequence, which is NP-hard.

V. Relaxation Schemes

The two-stage case with only one element in the set $\mathcal{F}^{(u_1)}$ was shown to be NP-hard in the previous section. It is therefore unlikely that one can design an algorithm that optimally solves the general case in polynomial time (unless P = NP). Therefore, we propose relaxation schemes that are computationally more tractable. Note that since the main motivation for solving the $\min\max$ optimization problem is to obtain a sequence of actions that has a performance guarantee, we will only propose relaxation schemes that are leading to lower bounds on the actual return of the sequences of actions. Note that all relaxation schemes are designed for the general $T-$stage case.

The first relaxation scheme works by dropping some constraints in order to obtain a problem that is solvable in polynomial time. We state that this scheme provides bounds that are greater or equal to the CGRL bound introduced in [14]. The second relaxation scheme is based on a Lagrangian relaxation where all constraints are dualized. The resulting problem can be solved in polynomial time using interior-point methods. We also state that this relaxation scheme always gives better bounds than the first relaxation scheme mentioned above, and consequently, better bounds than [14]. We also deduce from CGRL properties that the bounds computed from these relaxation schemes converge towards the actual return of the sequence $(u_0, \ldots, u_{T-1})$ when the sample dispersion converges towards zero. As a consequence, the sequences of actions that maximize those bounds also become optimal when the dispersion decreases towards zero.

From the previous section, we know that the first stage problem can be solved straightforwardly (cd. Lemma 2). We therefore only focus on relaxing the problem corresponding to the remaining stages $(\mathcal{P}''(\mathcal{F}, u_0, \ldots, u_{T-1}))$:

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

$$\min_{\substack{\hat{\mathbf{r}}_1 \quad \ldots \quad \hat{\mathbf{r}}_{T-1} \in \mathbb{R} \\ \hat{\mathbf{x}}_0 \quad \ldots \quad \hat{\mathbf{x}}_{T-1} \in \mathcal{X}}} \sum_{t=1}^{T-1} \hat{\mathbf{r}}_t,$$

subject to

$$\left|\hat{\mathbf{r}}_t - r^{(u_t),k_t}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2 ,$$
$$\forall (t, k_t) \in \{1, \ldots, T-1\} \times \left\{1, \ldots, n^{(u_t)}\right\}, \tag{10}$$
$$\left\|\hat{\mathbf{x}}_{t+1} - y^{(u_t),k_t}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2,$$
$$\forall (t, k_t) \in \{0, \ldots, T-1\} \times \left\{1, \ldots, n^{(u_t)}\right\}, \tag{11}$$
$$\|\hat{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}_{t'+1}\|^2 \leq L_f^2 \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t'}\|^2 ,$$
$$\forall t, t' \in \{0, \ldots, T-2 | u_t = u_{t'}\}, \tag{12}$$
$$\hat{\mathbf{x}}_0 = x_0 . \tag{13}$$

A. The Intertwined Trust-region (ITR) Relaxation Scheme

A natural way to obtain a relaxation from an optimization problem is to drop some constraints. A particular case of tractable non-convex quadratically constrained quadratic programs (QCQP) is where there is only one quadratic constraint. The idea here is to relax many constraints in order to obtain a tractable problem for each stage.

For all $t \in \{0, \ldots, T-1\}$, we select $\bar{k}_t$ in $\{1, \ldots, n^{(u_t)}\}$. The relaxation is obtained by dropping all constraints of type (5) and keeping one constraint by stage and by type. We therefore obtain a relaxed problem of the form:

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

$$\min_{\substack{\hat{\mathbf{r}}_1, \ldots, \hat{\mathbf{r}}_{T-1} \in \mathbb{R} \\ \hat{\mathbf{x}}_0, \ldots, \hat{\mathbf{x}}_{T-1} \in \mathcal{X}}} \sum_{t=1}^{T-1} \hat{\mathbf{r}}_t$$

subject to

$$\left|\hat{\mathbf{r}}_t - r^{(u_t),\bar{k}_t}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),\bar{k}_t}\right\|^2$$
$$t \in \{1, \ldots, T-1\} \tag{14}$$
$$\left\|\hat{\mathbf{x}}_t - y^{(u_{t-1}),\bar{k}_{t-1}}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_{t-1} - x^{(u_{t-1}),\bar{k}_{t-1}}\right\|^2$$
$$t \in \{1, \ldots, T-1\} \tag{15}$$
$$\hat{\mathbf{x}}_0 = x_0 \tag{16}$$

In the following, we provide the optimal solution of $\left(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1})\right)$ in closed-form. Such a solution is obtained by induction. It is more practical to work with the following family of T optimization problems $\left\{\left(\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, \ldots, u_j, \bar{k}_0, \ldots, \bar{k}_j)\right)\right\}_{j=0}^{j=T-1}$:

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

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

subject to

$$\left|\hat{\mathbf{r}}_t - r^{(u_t),\bar{k}_t}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_t - x^{(u_i),\bar{k}_t}\right\|^2$$
$$t \in \{1, \ldots, j\} \tag{17}$$
$$\left\|\hat{\mathbf{x}}_t - y^{(u_{t-1}),\bar{k}_{t-1}}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_{t-1} - x^{(u_{t-1}),\bar{k}_{t-1}}\right\|^2$$
$$t \in \{1, \ldots, j\} \tag{18}$$
$$\hat{\mathbf{x}}_0 = x_0 \tag{19}$$

The initialization of the induction is provided by the following Lemma:

Lemma 7: The optimal solution $D''_{ITR}(u_0, u_1, \bar{k}_0, \bar{k}_1)$ to $\left(\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, u_1, \bar{k}_0, \bar{k}_1)\right)$ is given by

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

where $\hat{\mathbf{x}}_1^*(\bar{k}_0, \bar{k}_1) \doteq y^{(u_0),\bar{k}_0} + L_f \frac{\left\|x_0 - x^{(u_0),\bar{k}_0}\right\|}{\left\|y^{(u_0),\bar{k}_0} - x^{(u_1),\bar{k}_1}\right\|}\left(y^{(u_0),\bar{k}_0} - x^{(u_1),\bar{k}_1}\right)$ if $y^{(u_0),\bar{k}_0} \neq x^{(u_1),\bar{k}_1}$ and, if $y^{(u_0),\bar{k}_0} = x^{(u_1),\bar{k}_1}$, $\hat{\mathbf{x}}_1^*(\bar{k}_0, \bar{k}_1)$ can be any point of the sphere centered in $y^{(u_0),\bar{k}_0} = x^{(u_1),\bar{k}_1}$ with radius $L_f\|x_0 - x^{(u_0),\bar{k}_0}\|$.

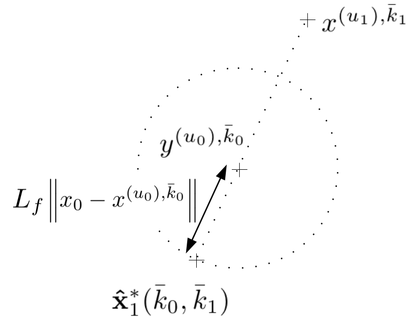
$+\ x^{(u_1),\bar{k}_1}$
$y^{(u_0),\bar{k}_0}$
$L_f \left\|x_0 - x^{(u_0),\bar{k}_0}\right\|$
$\hat{\mathbf{x}}_1^*(\bar{k}_0, \bar{k}_1)$

Fig. 1. A simple geometric algorithm to solve $(\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, u_1, \bar{k}_0, \bar{k}_1))$.

Lemma 8: The optimal solution to $\left(\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, \ldots, u_j, \bar{k}_0, \ldots, \bar{k}_j)\right)$ is given by: $\forall t \in \{1, \ldots, j\}$, $\hat{\mathbf{x}}_t^*(\bar{k}_0, \ldots, \bar{k}_t) \doteq y^{(u_{t-1}),\bar{k}_{t-1}} + L_f \frac{\left\|\hat{\mathbf{x}}_{t-1}^*(\bar{k}_0, \ldots, \bar{k}_{t-1}) - x^{(u_{t-1}),\bar{k}_{t-1}}\right\|}{\left\|y^{(u_{t-1}),\bar{k}_{t-1}} - x^{(u_t),\bar{k}_t}\right\|}\left(y^{(u_{t-1}),\bar{k}_{t-1}} - x^{(u_t),\bar{k}_t}\right)$ if $y^{(u_{t-1}),\bar{k}_{t-1}} \neq x^{(u_t),\bar{k}_t}$ and, if $y^{(u_{t-1}),\bar{k}_{t-1}} = x^{(u_t),\bar{k}_t}$, $\hat{\mathbf{x}}_t^*(\bar{k}_0, \ldots, \bar{k}_t)$ can be any point of the sphere centered in $y^{(u_{t-1}),\bar{k}_{t-1}} = x^{(u_t),\bar{k}_t}$ with radius $L_f\|\hat{\mathbf{x}}_{t-1}^*(\bar{k}_0, \ldots, \bar{k}_{t-1}) - x^{(u_{t-1}),\bar{k}_{t-1}}\|$.

Theorem 9: The solution to $\left(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1})\right)$ is given by:

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

where $\hat{\mathbf{r}}_t^* = r^{(u_t),\bar{k}_t} - L_\rho \left\| \hat{\mathbf{x}}_t^*(\bar{k}_0, \ldots, \bar{k}_t) - x^{(u_t),\bar{k}_t} \right\|$, $\hat{\mathbf{x}}_t^*(\bar{k}_0, \ldots, \bar{k}_t) \doteq y^{(u_{t-1}),\bar{k}_{t-1}} + L_f \frac{\left\| \hat{\mathbf{x}}_{t-1}^*(\bar{k}_0,\ldots,\bar{k}_{t-1}) - x^{(u_{t-1}),\bar{k}_{t-1}} \right\|}{\left\| y^{(u_{t-1}),\bar{k}_{t-1}} - x^{(u_t),\bar{k}_t} \right\|} \left(y^{(u_{t-1}),\bar{k}_{t-1}} - x^{(u_t),\bar{k}_t} \right)$ if $y^{(u_{t-1}),\bar{k}_{t-1}} \neq x^{(u_t),\bar{k}_t}$ and, if $y^{(u_{t-1}),\bar{k}_{t-1}} = x^{(u_t),\bar{k}_t}$, $\hat{\mathbf{x}}_t^*(\bar{k}_0, \ldots, \bar{k}_t)$ can be any point of the sphere centered in $y^{(u_{t-1}),\bar{k}_{t-1}} = x^{(u_t),\bar{k}_t}$ with radius $L_f \|\hat{\mathbf{x}}_{t-1}^*(\bar{k}_0, \ldots, \bar{k}_{t-1}) - x^{(u_{t-1}),\bar{k}_{t-1}}\|$.

Solving $(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1}))$ provides us with a family of relaxations for our initial problem by considering any combination $(\bar{k}_0, \ldots, \bar{k}_{T-1})$ of non-relaxed constraints. Taking the maximum out of these lower bounds yields the best possible bound out of this family of relaxations. Finally, if we denote by $B_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1})$ the bound made of the sum of the solution of the first stage problem and the maximal ITR relaxation of the problem $(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1}))$ over all possible couples of constraints, we have:

Definition 5 (Intertwined Trust-region Bound):

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

Notice that in the case where all $n^{(u_t)} \quad t = 0 \ldots T-1$ are equal to 1, then the ITR relaxation scheme provides an exact solution of the original problem $(\mathcal{P}(\mathcal{F}, u_0, \ldots, u_{T-1}))$:

Corollary 10: $\left(\forall t \in \{0, \ldots, T-1\}, n^{(u_t)} = 1\right) \implies B_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}) = B^*(\mathcal{F}, u_0, \ldots, u_{T-1})$.

B. The Lagrangian Relaxation

Another way to obtain a lower bound on the value of a minimization problem is to consider a Lagrangian relaxation. Consider again the optimization problem $(\mathcal{P}''(\mathcal{F}, u_0, \ldots, u_{T-1}))$. If we multiply the constraints (10) by dual variables $\mu_{t,k_t} \geq 0$, the constraints (11) by dual variables $\lambda_{t,k_t} \geq 0$ and the constraints (12) by dual variables $\nu_{t,t'} \geq 0$, we get the Lagrangian dual problem $(\mathcal{P}''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}))$:

$$\boxed{\begin{array}{l}
(\mathcal{P}''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1})) \quad : \\
\max\limits_{\substack{\nu_{t,t'} \in \mathbb{R} \\ \lambda_{t,k_t} \in \mathbb{R} \; \mu_{t,k_t} \in \mathbb{R}}} \quad \min\limits_{\hat{\mathbf{r}}_1, \ldots, \hat{\mathbf{r}}_{T-1} \in \mathbb{R} \; \hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_{T-1} \in \mathcal{X}} \\
\hat{\mathbf{r}}_1 + \cdots + \hat{\mathbf{r}}_{T-1} + \sum_{(t,k_t) \in \{1,\ldots,T-1\} \times \left\{1,\ldots,n^{(u_t)}\right\}} \\
\mu_{t,k_t} \left(\left| \hat{\mathbf{r}}_t - r^{(u_t),k_t} \right|^2 - L_\rho^2 \left\| \hat{\mathbf{x}}_t - x^{(u_t),k_t} \right\|^2 \right) \\
+ \sum_{(t,k_t) \in \{1,\ldots,T-1\} \times \left\{1,\ldots,n^{(u_t)}\right\}} \\
\lambda_{t,k_t} \left(\left\| \hat{\mathbf{x}}_{t+1} - y^{(u_t),k_t} \right\|^2 - L_f^2 \left\| \hat{\mathbf{x}}_t - x^{(u_t),k_t} \right\|^2 \right) \\
+ \sum_{t,t' \in \{0,\ldots,T-2 | u_t = u_{t'}\}} \\
\nu_{t,t'} \left(\left\| \hat{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}_{t'+1} \right\|^2 - L_f^2 \left\| \hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t'} \right\|^2 \right).
\end{array}}$$

Observe that the optimal value of $(\mathcal{P}''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}))$ is known to provide a lower bound on the optimal value of $(\mathcal{P}''(\mathcal{F}, u_0, \ldots, u_{T-1}))$ [43]. Note that the above Lagrangian relaxation can be solved in polynomial time and is equivalent to another standard relaxation of quadratically constrained quadratic programs known as the SDP relaxation. It turns out that one relaxation is the dual of the other [44], [45], [46].

Definition 6 (Lagrandian Bound $B_{LD}\,(\mathcal{F}, u_0, \ldots, u_{T-1})$): Let $B''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1})$ be the optimal Lagrangian dual of $(\mathcal{P}''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}))$. Then,

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

C. Comparing the Bounds

The CGRL algorithm proposed in [16], [14] for addressing the $\min \max$ problem uses the procedure described in [15] for computing a lower bound on the return of a policy given a sample of trajectories. More specifically, for a given sequence $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^2$, the program $(\mathcal{P}(\mathcal{F}, u_0, \ldots, u_{T-1}))$ is replaced by a lower bound $B_{CGRL}(\mathcal{F}, u_0, \ldots, u_{T-1})$. We may now wonder how this bound compares with the two new bounds of $(\mathcal{P}(\mathcal{F}, u_0, \ldots, u_{T-1}))$ that we have proposed: the ITR bound and the Lagrangian bound.

1) Trust-region Versus CGRL: We first recall the definition of the CGRL bound.

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

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

The following theorem states that the ITR bound is always greater than or equal to the CGRL bound.

Theorem 11:

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

2) Lagrangian Relaxation Versus Intertwined Trust-region Relaxation: In this section, we state that the lower bound obtained with the Lagrangian relaxation is always greater than or equal to the ITR bound. To do so, we show that strong duality holds for the Lagrangian dual of $\left(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1})\right)$ for a given $\left(\bar{k}_0, \ldots, \bar{k}_{T-1}\right)$. The Lagrangian dual of $(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1}))$ reads

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

$$\max_{\substack{\lambda_1, \ldots, \lambda_{T-1} \in \mathbb{R} \\ \mu_1, \ldots, \mu_{T-1} \in \mathbb{R}}} \quad \min_{\substack{\hat{\mathbf{r}}_1, \ldots, \hat{\mathbf{r}}_{T-1} \in \mathbb{R} \\ \hat{\mathbf{x}}_0, \ldots, \hat{\mathbf{x}}_{T-1} \in \mathcal{X}}}$$

$$\hat{\mathbf{r}}_1 + \cdots + \hat{\mathbf{r}}_{T-1} + \mu_1 \left(\left| \hat{\mathbf{r}}_1 - r^{(u_1),\bar{k}_1} \right|^2 - L_\rho^2 \left\| \hat{\mathbf{x}}_1 - x^{(u_1),\bar{k}_1} \right\|^2 \right)$$
$$+ \ldots +$$

$$\mu_{T-1}\left(\left|\hat{\mathbf{r}}_{T-1}-r^{(u_{T-1}),\bar{k}_{T-1}}\right|^2-L_\rho^2\left\|\hat{\mathbf{x}}_{T-1}-x^{(u_{T-1}),\bar{k}_{T-1}}\right\|^2\right)$$
$$+\lambda_1\left(\left\|\hat{\mathbf{x}}_1-y^{(u_0),\bar{k}_0}\right\|^2-L_f^2\left\|\hat{\mathbf{x}}_0-x^{(u_0),\bar{k}_0}\right\|^2\right) \quad + \quad \ldots$$
$$+\lambda_{T-1}\Bigg(\left\|\hat{\mathbf{x}}_{T-1}-y^{(u_{T-2}),\bar{k}_{T-2}}\right\|^2$$
$$-L_f^2\left\|\hat{\mathbf{x}}_{T-2}-x^{(u_{T-2}),\bar{k}_{T-2}}\right\|^2\Bigg).$$

Theorem 12: Strong duality holds for the Lagrangian relaxation of the Intertwined Trust-region problem $(LD''_{ITR}(\mathcal{F},u_0,\ldots,u_{T-1},\bar{k}_0,\ldots,\bar{k}_{T-1}))$.

Theorem 13:

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

3) Bounds Inequalities: Summary: We summarize in the following theorem all the results that were obtained in the previous sections.

Theorem 14: $\forall\,(u_0,\ldots,u_{T-1})\in\mathcal{U}^T$,

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

D. Convergence Properties

Theorem 14 implies that the convergence properties of the CGRL bound - when the dispersion of the sample of transitions goes to zero - also apply to other bounds presented in this paper (see [18] for more informations).

VI. Conclusions

This paper - which is a shortened version of [18] - addresses the problem of computing $\min\max$ policies for deterministic, Lipschitz continuous batch mode reinforcement learning. First, we have shown that this $\min\max$ problem is NP-hard. Afterwards, we have proposed two relaxation schemes. Both have been extensively studied and, in particular, they have been shown to perform better than the CGRL algorithm that has been introduced earlier to address this min-max generalization problem.

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.

Acknowledgements

Raphael Fonteneau is a Post-doctoral fellow of the F.R.S.-FNRS (Belgian Funds for Scientific Research). This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control and Optimization) funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The authors also thank Yurii Nesterov, Adrien Hoarau and Benoit Daene for fruitful discussions. The scientific responsibility rests with its authors.

References

[1] R. Sutton and A. Barto, *Reinforcement Learning*. MIT Press, 1998.

[2] J. Ingersoll, *Theory of Financial Decision Making*. Rowman and Littlefield Publishers, Inc., 1987.

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

[4] ——, "An experimental design for the development of adaptive treatment strategies," *Statistics in Medicine*, vol. 24, pp. 1455–1481, 2005.

[5] 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)*, Porto, Portugal, 2005, pp. 317–328.

[6] S. Bradtke and A. Barto, "Linear least-squares algorithms for temporal difference learning," *Machine Learning*, vol. 22, pp. 33–57, 1996.

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

[8] M. Lagoudakis and R. Parr, "Least-squares policy iteration," *Jounal of Machine Learning Research*, vol. 4, pp. 1107–1149, 2003.

[9] D. Ormoneit and S. Sen, "Kernel-based reinforcement learning," *Machine Learning*, vol. 49, no. 2-3, pp. 161–178, 2002.

[10] R. Fonteneau, "Contributions to Batch Mode Reinforcement Learning," Ph.D. dissertation, University of Liège, 2011.

[11] D. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, 1996.

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

[13] R. Fonteneau, S. Murphy, L. Wehenkel, and D. Ernst, "Computing bounds for kernel-based policy evaluation in reinforcement learning," University of Liège, Tech. Rep., 2010.

[14] ——, "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)*, vol. 129. Springer, Heidelberg, 2011, pp. 61–77.

[15] ——, "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.

[16] ——, "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.

[17] A. Conn, N. Gould, and P. Toint, *Trust-region Methods*. Society for Industrial Mathematics, 2000, vol. 1.

[18] R. Fonteneau, D. Ernst, B. Boigelot, and Q. Louveaux, "Min max generalization for deterministic batch mode reinforcement learning: Relaxation schemes," *SIAM Journal on Control and Optimization*, vol. 51, no. 5, pp. 3355–3385, 2013. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/120867263

[19] E. Delage and S. Mannor, "Percentile optimization for Markov decision processes with parameter uncertainty," *Operations Research*, vol. 58, no. 1, pp. 203–213, 2010.

[20] M. L. Littman, "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, 1994.

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

[22] M. L. Littman, "A tutorial on partially observable markov decision processes," *Journal of Mathematical Psychology*, vol. 53, no. 3, pp. 119 – 125, 2009, special Issue: Dynamic Decision Making.

[23] S. Koenig, "Minimax real-time heuristic search," *Artificial Intelligence*, vol. 129, no. 1-2, pp. 165–197, 2001.

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

[25] M. Qian and S. Murphy, "Performance guarantees for individualized treatment rules," Department of Statistics, University of Michigan, Tech. Rep. 498, 2009.

[26] C. Paduraru, D. Precup, and J. Pineau, "A framework for computing bounds for the return of a policy," in *Ninth European Workshop on Reinforcement Learning (EWRL9)*, 2011.

[27] L. Hansen and T. Sargent, "Robust Control and Model Uncertainty," *American Economic Review*, pp. 60–66, 2001.

[28] T. Başar and P. Bernhard, *H_∞-optimal control and related minimax design problems: a dynamic game approach*. Birkhauser, 1995, vol. 5.

[29] E. Camacho and C. Bordons, *Model Predictive Control*. Springer, 2004.

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

[31] P. Scokaert and D. Mayne, "Min-max feedback model predictive control for constrained linear systems," *IEEE Transactions on Automatic Control*, vol. 43, no. 8, pp. 1136–1142, 1998.

[32] A. Bemporad and M. Morari, "Robust model predictive control: A survey," *Robustness in Identification and Control*, vol. 245, pp. 207–226, 1999.

[33] J. Birge and F. Louveaux, *Introduction to Stochastic Programming*. Springer Verlag, 1997.

[34] B. Defourny, D. Ernst, and L. Wehenkel, "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*, 2008.

[35] A. Shapiro, "A dynamic programming approach to adjustable robust optimization," *Operations Research Letters*, vol. 39, no. 2, pp. 83–87, 2011.

[36] ——, "Minimax and risk averse multistage stochastic programming," School of Industrial & Systems Engineering, Georgia Institute of Technology, Tech. Rep., 2011.

[37] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[38] D. Henrion, S. Tarbouriech, and D. Arzelier, "LMI approximations for the radius of the intersection of ellipsoids: Survey," *Journal of Optimization Theory and Applications*, vol. 108, no. 1, pp. 1–28, 2001.

[39] S. Boyd, L. El-Ghaoui, E. Feron, V. Balakrishnan, and E. Yaz, "Linear matrix inequalities in system and control theory," *Proceedings of the IEEE*, vol. 85, no. 4, pp. 698–699, 1997.

[40] P. Pardalos and S. Vavasis, "Quadratic programming with one negative eigenvalue is NP-hard," *Journal of Global Optimization*, vol. 1, no. 1, pp. 15–22, 1991.

[41] R. Freund and J. Orlin, "On the complexity of four polyhedral set containment problems," *Mathematical programming*, vol. 33, no. 2, pp. 139–145, 1985.

[42] C. Papadimitriou, "On the complexity of integer programming," *Journal of the ACM (JACM)*, vol. 28, no. 4, pp. 765–768, 1981.

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

[44] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Review*, vol. 38, no. 1, pp. 49–95, 1996.

[45] A. d'Aspremont and S. Boyd, "Relaxations and randomized methods for nonconvex qcqps," *EE392o Class Notes, Stanford University*, 2003.

[46] Y. Nesterov, H. Wolkowicz, and Y. Ye, "Semidefinite programming relaxations of nonconvex quadratic optimization," *Handbook of semidefinite programming*, pp. 361–419, 2000.