SIAM J. CONTROL OPTIM.
Vol. 51, No. 5, pp. 3355–3385

MIN MAX GENERALIZATION FOR DETERMINISTIC BATCH MODE REINFORCEMENT LEARNING: RELAXATION SCHEMES[*]

R. FONTENEAU[†], D. ERNST[†], B. BOIGELOT[†], AND Q. LOUVEAUX[†]

Abstract. We study the min max optimization problem introduced in Fonteneau et al. [*Towards min max reinforcement learning*, ICAART 2010, Springer, Heidelberg, 2011, pp. 61–77] for computing policies for batch mode reinforcement learning in a deterministic setting with fixed, finite time horizon. First, we show 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 also theoretically prove and empirically illustrate that both relaxation schemes provide better results than those given in [Fonteneau et al., 2011, as cited above].

Key words. reinforcement learning, min max generalization, nonconvex optimization, computational complexity

AMS subject classification. 60J05

DOI. 10.1137/120867263

1. Introduction.

Research in reinforcement learning (RL) [47] 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 to researchers trying to solve sequential decision making problems in many fields such as finance [25], medicine [31, 32], or engineering [41]. Since the end of the 1990s, 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, 13, 27, 35, 41, 20]. This subfield of RL is known as "batch mode RL (BMRL)."

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 [3, 27, 13, 7]. 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

[*] Received by the editors February 24, 2012; accepted for publication (in revised form) June 28, 2013; published electronically September 4, 2013. 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.

http://www.siam.org/journals/sicon/51-5/86726.html

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

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 nonvisited state space area [18].

To overcome this problem, reference [19] 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 [16, 17] 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.

In this paper, we propose to further investigate the min max generalization optimization problem that was initially proposed in [19]. We first show 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 in a configuration where each stage resorts to solving a *trust-region subproblem* [9]. The second relaxation scheme, based on a Lagrangian relaxation where all constraints are dualized, can be solved in polynomial time. We prove that both relaxation schemes always provide bounds that are greater than or equal to the CGRL bound. We also deduce from CGRL properties that these bounds are tight in a sense that they converge towards the actual return when the sample dispersion converges towards zero, and that the sequences of actions that maximize these bounds converge towards optimal ones.

The paper is organized as follows:

- in section 2, we give a short summary of the literature related to this work;
- section 3 formalizes the min max generalization problem in a Lipschitz continuous, deterministic BMRL context;
- in section 4, we analyze the complexity of the min max generalization problem. To this end, we focus on the particular two-stage case, for which we prove that it can be decoupled into two independent problems corresponding, respectively, to the first stage and the second stage (Lemma 4.1):
  - the first stage problem leads to a trivial optimization problem that can be solved in closed form (Corollary 4.2);
  - we prove in section 4.2 that the second stage problem is NP-hard (Corollary 4.6), which consequently proves the NP-hardness of the min part of the general min max generalization problem (Theorem 4.7);
- we then describe in section 5 the two relaxation schemes that we propose:

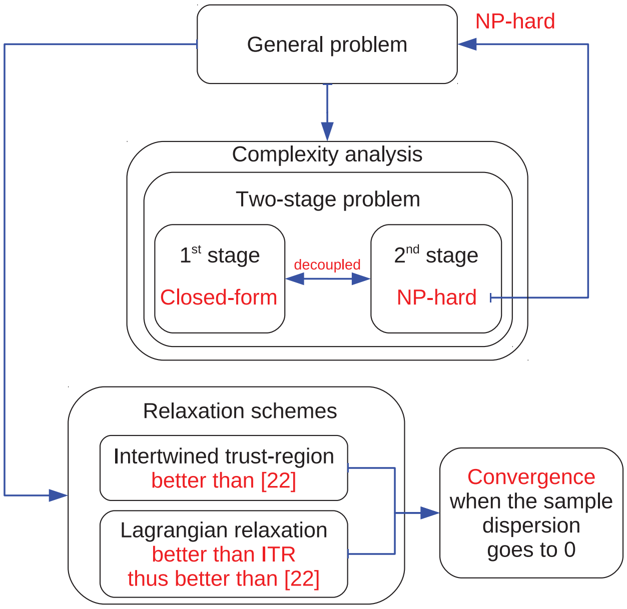

Fig. 1.1. *Main results of the paper.*

  - the intertwined trust-region (ITR) relaxation scheme (section 5.1);
  - the Lagrangian relaxation scheme (section 5.2);
- we prove in section 5.3.1 that the ITR relaxation scheme gives better results than CGRL (Theorem 5.9);
- we show in section 5.3.2 that the Lagrangian relaxation scheme povides better results than the ITR relaxation scheme (Theorem 5.17), and consequently better results than CGRL (Theorem 5.18);
- we provide in section 5.4 results about the asymptotic behavior of the relaxation schemes as a function of the sample dispersion:
  - the bounds provided by the relaxation schemes converge towards the actual return when the sample dispersion decreases towards zero (Theorem 5.21);
  - the sequences of actions maximizing such bounds converge towards optimal sequences of actions when the sample dispersion decreases towards zero (Theorem 5.24);
- section 6 illustrates the relaxation schemes on an academic benchmark;
- section 7 concludes the paper.

We provide in Figure 1.1 an illustration of the road map of the main results of this paper.

2. Related work. Several works have already been built upon min max paradigms for computing policies in an 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 [12]. 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 [28, 42]. They have also received some attention for solving partially observable Markov decision processes [29, 26].

The min max approach towards generalization, originally introduced in [19], 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 [30, 40, 36].

Other fields of research have proposed min-max–type strategies for computing control policies. This includes not only robust control theory [22] with H_∞ methods [1], but also model predictive control (MPC) theory—where usually the environment is supposed to be fully known [8, 14]—for which min max approaches have been used to determine an optimal sequence of actions with respect to the "worst case" disturbance sequence occurring [43, 2]. Finally, there is a broad stream of works in the field of stochastic programming [4] that have addressed the problem of safely planning under uncertainties, mainly known as "robust stochastic programming" or "risk-averse stochastic programming" [11, 44, 45, 33].

3. Problem formalization.

We first formalize the BMRL setting in section 3.1, and we state the min max generalization problem in section 3.2.

3.1. Batch mode reinforcement learning.

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

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

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 $\mathcal{U} = \{u^{(1)}, \dots, u^{(m)}\}$ that we abusively identify with $\{1, \dots, m\}$. We assume that the (finite) optimization horizon $T \in \mathbb{N} \setminus \{0\}$ is a given (fixed) parameter of the problem. An instantaneous reward

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

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 $(u_0, \dots, u_{T-1}) \in \mathcal{U}^T$, the cumulated reward over T stages (also named T-stage return) is defined as follows.

Definition 3.1 (T-stage return).

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

where

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

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

Definition 3.2 (optimal T-stage return).

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

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:

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

Under those assumptions, 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 $(\tilde{u}_0^*, \ldots, \tilde{u}_{T-1}^*) \in \mathcal{U}^T$ such that $J(\tilde{u}_0^*, \ldots, \tilde{u}_{T-1}^*)$ is as close as possible to J_T^*.

3.2. 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 [19].

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

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

where $\|.\|$ denotes the Euclidean norm over the space $\mathcal{X}$. We also assume that two constants L_f and L_ρ satisfying the above-written inequalities are known. Such Lipschitz continuity assumptions are very standard in the field of BMRL 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 [19], 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 is written as follows:

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

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

subject to

(3.1)
$$\left| \hat{\mathbf{r}}_t - r^{(u_t),k_t} \right|^2 \leq L_\rho^2 \left\| \hat{\mathbf{x}}_t - x^{(u_t),k_t} \right\|^2 \ \forall (t, k_t) \in \{0, \ldots, T-1\} \times \left\{1, \ldots, n^{(u_t)}\right\},$$

(3.2)
$$\left\| \hat{\mathbf{x}}_{t+1} - y^{(u_t),k_t} \right\|^2 \leq L_f^2 \left\| \hat{\mathbf{x}}_t - x^{(u_t),k_t} \right\|^2 \ \forall (t, k_t) \in \{0, \ldots, T-1\} \times \left\{1, \ldots, n^{(u_t)}\right\},$$

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

(3.4)
$$\|\hat{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}_{t'+1}\|^2 \leq L_f^2 \|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t'}\|^2 \ \forall t, t' \in \{0, \ldots, T-2 | u_t = u_{t'}\},$$

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

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 3.3 (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 (3.1) and (3.3) (resp., (3.2) and (3.4)) 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 3.5 ensures that the pessimistic trajectory starts at 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., $T, \mathcal{F}, L_f, L_\rho, x_0, u_0, \dots, u_{T-1}$) are given in the form of rational numbers.

4. Analysis of the complexity.

In this section, we prove that solving the min problem ($\mathcal{P}(\mathcal{F}, u_0, \dots, u_{T-1})$) is NP-hard. More precisely, we will prove 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.

4.1. Redundancy of constraint (3.3).

We first want to show that the constraints (3.3) are not needed. Indeed, in any optimal solution, they are always satisfied. Let $\bar{\mathcal{P}}\left(\mathcal{F}, u_0, \dots, u_{T-1}\right)$ be the relaxation of $\mathcal{P}\left(\mathcal{F}, u_0, \dots, u_{T-1}\right)$, where all constraints of type (3.3) are relaxed.

Lemma 4.1. *Consider* $(\hat{\mathbf{r}}^*, \hat{\mathbf{x}}^*) \in \mathbb{R}^T \times \mathcal{X}^T$ *an optimal solution to* $\bar{\mathcal{P}}(\mathcal{F}, u_0, \dots, u_{T-1})$. *Then, for all* t, t' *such that* $u_t = u_{t'}$,

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

Proof. Consider an optimal solution to $\bar{\mathcal{P}}(\mathcal{F}, u_0, \dots, u_{T-1})$. Observe that any variable $\hat{\mathbf{r}}_\mathbf{t}$ only appears in constraints (3.1) in a series of interval constraints of the type

$$\left|\hat{\mathbf{r}}_t - r^{(u_t),k_t}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2 \forall (t, k_t) \in \{0, \dots, T-1\} \times \left\{1, \dots, n^{(u_t)}\right\}. \tag{4.1}$$

Since the objective function is $\min \sum_{t=0}^{T-1} \hat{\mathbf{r}}_\mathbf{t}$, we claim that, for each t, there exists at least one constraint (4.1) that is tight. Indeed, assume by contradiction that it is not the case; by considering $\hat{\mathbf{r}}_\mathbf{t} - \epsilon$, $\epsilon > 0$, we obtain a trivially better feasible solution, a contradiction. Therefore, for each t, there exists $\bar{k}_t$ such that

$$\hat{\mathbf{r}}_t^* = r^{(u_t),\bar{k}_t} - L_\rho \left\|\hat{\mathbf{x}}_t^* - x^{(u_t),\bar{k}_t}\right\|. \tag{4.2}$$

Consider now a pair (t, t') such that $u_t = u_{t'} = u$. We now discuss two cases depending on the sign of $\hat{\mathbf{r}}_t^* - \hat{\mathbf{r}}_{t'}^*$.

• **If** $\hat{\mathbf{r}}_t^* - \hat{\mathbf{r}}_{t'}^* \geq 0$.
Using (4.2) with index $\bar{k}_t^*$, we have

$$\hat{\mathbf{r}}_t^* - \hat{\mathbf{r}}_{t'}^* \leq L_\rho \left(\left\|\hat{\mathbf{x}}_{t'}^* - x^{(u),\bar{k}_t^*}\right\| - \left\|\hat{\mathbf{x}}_t^* - x^{(u),\bar{k}_t^*}\right\|\right). \tag{4.3}$$

Since $\hat{\mathbf{r}}_t^* - \hat{\mathbf{r}}_{t'}^* \geq 0$, we therefore have

$$\left|\hat{\mathbf{r}}_t^* - \hat{\mathbf{r}}_{t'}^*\right| \leq L_\rho \left(\left\|\hat{\mathbf{x}}_{t'}^* - x^{(u),\bar{k}_t^*}\right\| - \left\|\hat{\mathbf{x}}_t^* - x^{(u),\bar{k}_t^*}\right\|\right). \tag{4.4}$$

Using the triangle inequality we can write

$$\left\|\hat{\mathbf{x}}_{t'}^* - x^{(u),\bar{k}_t^*}\right\| \leq \left\|\hat{\mathbf{x}}_{t'}^* - \hat{\mathbf{x}}_t^*\right\| + \left\|\hat{\mathbf{x}}_t^* - x^{(u),\bar{k}_t^*}\right\|. \tag{4.5}$$

Replacing (4.5) in (4.4) we obtain

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

which shows that $\hat{\mathbf{r}}_t^*$ and $\hat{\mathbf{r}}_{t'}^*$ satisfy constraint (3.3).

• **If** $\hat{\mathbf{r}}^*_t - \hat{\mathbf{r}}^*_{t'} < 0$.
Using (4.1) with index $\bar{k}^*_{t'}$, we have

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

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

$$|\hat{\mathbf{r}}^*_{t'} - \hat{\mathbf{r}}^*_t| \leq L_\rho \left(\left\| \hat{\mathbf{x}}^*_t - x^{(u),\bar{k}^*_{t'}} \right\| - \left\| \hat{\mathbf{x}}^*_{t'} - x^{(u),\bar{k}^*_{t'}} \right\| \right). \tag{4.6}$$

Using the triangle inequality we can write

$$\left\| \hat{\mathbf{x}}^*_t - x^{(u),\bar{k}^*_{t'}} \right\| \leq \| \hat{\mathbf{x}}^*_t - \hat{\mathbf{x}}^*_{t'} \| + \left\| \hat{\mathbf{x}}^*_{t'} - x^{(u),\bar{k}^*_{t'}} \right\|. \tag{4.7}$$

Replacing (4.7) in (4.6) yields

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

which again shows that $\hat{\mathbf{r}}^*_t$ and $\hat{\mathbf{r}}^*_{t'}$ satisfy constraint (3.3).

In both cases $\hat{\mathbf{r}}^*_t - \hat{\mathbf{r}}^*_{t'} \geq 0$ and $\hat{\mathbf{r}}^*_t - \hat{\mathbf{r}}^*_{t'} < 0$, we have shown that constraint (3.3) is satisfied. □

Observe that Lemma 4.1 implies that $\hat{\mathbf{r}}^*_0$ is decoupled from the rest of the problem. Therefore, $\hat{\mathbf{r}}^*_0$ is the solution of

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

$$\min_{\substack{\hat{\mathbf{r}}_0 \in \mathbb{R} \\ \hat{\mathbf{x}}_0 \in \mathcal{X}}} \quad \hat{\mathbf{r}}_0$$

subject to

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

Lemma 4.2. *The solution of the problem* $(\mathcal{P}'(\mathcal{F}, u_0))$ *is*

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

Proof. This follows directly from the fact that we minimize $\hat{\mathbf{r}}_0 \in \mathbb{R}$ under interval constraints. □

In the particular case $T = 2$, Lemma 4.1 implies that the two stages are decoupled. In particular, the problem $\mathcal{P}(\mathcal{F}, u_0, u_1)$ can be decomposed into two subproblems $(\mathcal{P}'(\mathcal{F}, u_0))$ and $(\mathcal{P}''(\mathcal{F}, u_0, u_1))$:

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

$$\min_{\substack{\hat{\mathbf{r}}_1 \in \mathbb{R} \\ \hat{\mathbf{x}}_1 \in \mathcal{X}}} \quad \hat{\mathbf{r}}_1 \tag{4.8}$$

subject to

$$\left| \hat{\mathbf{r}}_1 - r^{(u_1),k_1} \right|^2 \leq L_\rho^2 \left\| \hat{\mathbf{x}}_1 - x^{(u_1),k_1} \right\|^2 \ \forall k_1 \in \left\{ 1, \dots, n^{(u_1)} \right\}, \tag{4.9}$$

$$\left\| \hat{\mathbf{x}}_1 - y^{(u_0),k_0} \right\|^2 \leq L_f^2 \left\| x_0 - x^{(u_0),k_0} \right\|^2 \ \forall k_0 \in \left\{ 1, \dots, n^{(u_0)} \right\}. \tag{4.10}$$

4.2. 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 $\mathcal{F}^{(u_1)} = \{(x^{(u_1),1}, r^{(u_1),1}, y^{(u_1),1})\}$. In this particular case, the problem $(\mathcal{P}''(\mathcal{F}, u_0, u_1))$ amounts to maximizing the distance $\|\hat{\mathbf{x}}_1 - x^{(u_1),1}\|$ under an intersection of balls as we show in the following lemma.

LEMMA 4.3. *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 $\|\hat{\mathbf{x}}_1 - x^{(u_1),1}\|$ subject to

$$\left\| \hat{\mathbf{x}}_1 - y^{(u_0),k_0} \right\|^2 \leq L_f^2 \left\| x_0 - x^{(u_0),k_0} \right\|^2 \qquad \forall \left(x^{(u_0),k_0}, r^{(u_0),k_0}, y^{(u_0),k_0} \right) \in \mathcal{F}^{(u_0)}.$$

Proof. The unique constraint concerning $\hat{\mathbf{r}}_1$ is an interval. Therefore $\hat{\mathbf{r}}_1^*$ takes the value of the lower bound of the interval. In order to obtain the lowest such value, the right-hand side of (4.9) must be maximized under the other constraints. □

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 5.6. 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 now prove. To do it, we introduce the MNBC (for "max norm with ball constraints") decision problem.

DEFINITION 4.4 (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

$$\left\| x - y^i \right\|^2 \leq \gamma_i \qquad \forall i \in \{1, \ldots, I\}$$

LEMMA 4.5. *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 [23] and [5], 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 an NP-hard problem [39]. Furthermore, containment problems between polyhedra and spheres are known to be NP-hard as well [21]. However, those problems do not admit immediate reductions to MNBC. This motivates our development of a proof relying on a reduction from $\{0, 1\}$-programming.

Proof. To prove it, we will do a reduction from the $\{0, 1\}$-programming feasibility problem [38]. More precisely, we consider in this proof the $\{0, 2\}$-programming feasibility problem, which is equivalent. The problem is, given $p \in \mathbb{N}, A \in \mathbb{Z}^{p \times d}, b \in \mathbb{Z}^p$ to find whether there exists $x \in \{0,2\}^d$ that satisfies $Ax \leq b$. This problem is known to be NP-hard and we now provide a polynomial reduction to MNBC.

The dimension d is kept the same in both problems. The first step is to define a set of constraints for MNBC such that the only potential feasible solutions are exactly $x \in \{0,2\}^d$. We define

$$x^{(0)} \triangleq (1, \ldots, 1)$$

and

$$C \triangleq d.$$

For $i = 1, \ldots, d$, we define

$$y^{2i} \triangleq \left(y_1^{2i}, \ldots, y_d^{2i}\right)$$

with $y_i^{2i} \triangleq 0$ and $y_j^{2i} \triangleq 1$ for all $j \neq i$ and $\gamma_i \triangleq d + 3$.

Similarly for $i = 1, \ldots, d$, we define

$$y^{2i+1} \triangleq \left(y_1^{2i+1}, \ldots, y_d^{2i+1}\right)$$

with $y_i^{2i+1} \triangleq 2$ and $y_j^{2i+1} \triangleq 1$ for all $j \neq i$ and $\gamma_i \triangleq d + 3$.

Claim.

$$\left\{x \in \mathbb{R}^d \mid \|x - x^{(0)}\|^2 \geq d\right\} \cap \left(\bigcap_{i=2}^{2d+1} \left\{x \in \mathbb{R}^d \mid \|x - y^i\|^2 \leq \gamma_i\right\}\right) = \{0,2\}^d.$$

It is readily verified that any $x \in \{0,2\}^d$ belongs to the $2d+1$ above sets.

Consider $x \in \mathbb{R}^d$ that belongs to the $2d+1$ above sets. Consider an index $k \in \{1, \ldots, d\}$. Using the constraints defining the sets, we can in particular write

$$\begin{gathered}
\|(x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_d) - (1, \ldots, 1)\|^2 \geq d, \\
\|(x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_d) - (1, \ldots, 1, 0, 1, \ldots, 1)\|^2 \leq d + 3, \\
\|(x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_d) - (1, \ldots, 1, 2, 1, \ldots, 1)\|^2 \leq d + 3,
\end{gathered}$$

that we can write algebraically

$$\sum_{j \neq k} (x_j - 1)^2 + (x_k - 1)^2 \geq d, \tag{4.11}$$

$$\sum_{j \neq k} (x_j - 1)^2 + x_k^2 \leq d + 3, \tag{4.12}$$

$$\sum_{j \neq k} (x_j - 1)^2 + (x_k - 2)^2 \leq d + 3. \tag{4.13}$$

By computing (4.12)–(4.11) and (4.13)–(4.11), we obtain $x_k \leq 2$ and $x_k \geq 0$, respectively. This implies that

$$\sum_{k=1}^{d} (x_k - 1)^2 \leq d$$

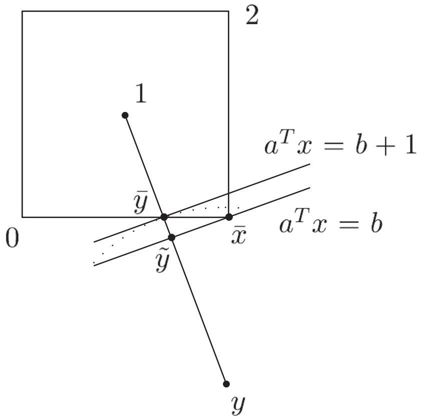

FIG. 4.1. *The case when* $a^T\bar{x} \le b$.

and the equality is obtained if and only if we have that $x_k \in \{0, 2\}$ for all k which proves the claim.

It remains to prove that we can encode any linear inequality through a ball constraint. Consider an inequality of the type $\sum_{j=1}^{d} a_j x_j \le b$. We assume that $a \neq 0$ and that b is even and therefore that there exists no $x \in \{0,2\}^d$ such that $a^T x = b + 1$. We want to show that there exist $y \in \mathbb{Q}^d$ and $\gamma \in \mathbb{Q}$ such that

$$\left\{x \in \{0,2\}^d \mid a^T x \le b\right\} = \left\{x \in \{0,2\}^d \mid \|x - y\|^2 \le \gamma\right\}. \tag{4.14}$$

Let $\bar{y} \in \mathbb{R}^d$ be the intersection point of the hyperplane $a^T x = b + 1$ and the line $(1\ \cdots\ 1)^T + \lambda(a_1\ \cdots\ a_d)^T, \lambda \in \mathbb{R}$. Note that λ is a rational number that can be expressed in closed form with both numerator and denominator of polynomial encoding length. Let r be defined as follows:

$$r = \left\lceil \frac{d}{2}\sqrt{\sum_{j=1}^{d} a_j^2} + 1 \right\rceil .$$

Observe that since r is an integer, the square root in its formula can be approximated with polynomial precision. We claim that choosing $\gamma \triangleq r^2$ and $y \triangleq \bar{y} - ra$ allows us to obtain (4.14). To prove it, we need to show that $x \in \{0,2\}^d$ belongs to the ball if and only if it satisfies the constraint $a^T x \le b$. Let $\bar{x} \in \{0,2\}^d$. There are two cases to consider.

- Suppose first that $a^T\bar{x} \ge b + 2$. Since $\bar{y}$ is the closest point to y that satisfies $a^T y = b + 1$, it also implies that any point x such that $a^T x > b + 1$ is such that $\|x - y\|^2 > r^2$ proving that

$$\bar{x} \notin \left\{x \in \mathbb{R}^d \mid \|x - y\|^2 \le r^2\right\}.$$

- Suppose now that $a^T\bar{x} \le b$ and in particular that $a^T\bar{x} = b - k$ with $k \in \mathbb{N}$ (see Figure 4.1). Let $\tilde{y} \in \mathbb{R}^d$ be the intersection point of the hyperplane $a^T x = b - k$ and the line $(1\ \cdots\ 1)^T + \lambda(a_1\ \cdots\ a_d)^T, \lambda \in \mathbb{R}$. Since $\big((1\cdots1)^T, \tilde{y}, \bar{x}\big)$ form a right triangle with the right angle in $\tilde{y}$ and since $\left\|(1\cdots1)^T - \bar{x}\right\|^2 \le d$, we have

$$\|\tilde{y} - \bar{x}\|^2 \le d. \tag{4.15}$$

By the definition of y, we have

$$\|y - \bar{y}\| = r,$$

and by the definition of $\bar{y}$ and $\tilde{y}$, we have

$$\|\bar{y} - \tilde{y}\| \geq \frac{1}{\sqrt{\sum_{j=1}^{d} a_j^2}}.$$

Since $\bar{y}, \tilde{y}$, and y belong to the same line, we have

$$\|y - \tilde{y}\| \leq r - \frac{1}{\sqrt{\sum_{j=1}^{d} a_j^2}}. \tag{4.16}$$

As $(y, \tilde{y}, \bar{x})$ form a right triangle with the right angle in $\tilde{y}$, we have that

$$\begin{aligned}\|\bar{x} - y\|^2 &= \|y - \tilde{y}\|^2 + \|\bar{x} - \tilde{y}\|^2 \\ &\leq \left(r - \frac{1}{\sqrt{\sum_{j=1}^{d} a_j^2}}\right)^2 + d \qquad \text{using (4.15), (4.16)} \\ &= r^2 - \frac{2r}{\sqrt{\sum_{j=1}^{d} a_j^2}} + \frac{1}{\sum_{j=1}^{d} a_j^2} + d.\end{aligned}$$

Since by definition, $r \geq \frac{d}{2}\sqrt{\sum_{j=1}^{d} a_j^2} + 1$, we can write

$$\begin{aligned}\|\bar{x} - y\|^2 &\leq r^2 - d - \frac{2}{\sqrt{\sum_{j=1}^{d} a_j^2}} + \frac{1}{\sum_{j=1}^{d} a_j^2} + d \\ &= r^2 - \frac{1}{\sum_{j=1}^{d} a_j^2} \\ &\leq r^2.\end{aligned}$$

This proves that the chosen ball $\{x \in \mathbb{R}^d \mid \|x - y\|^2 \leq r^2\}$ includes the same points from $\{0, 2\}^d$ as the linear inequality $a^T x \leq b$.

The encoding length of all data is furthermore polynomial in the encoding length of the initial inequalities. This completes the reduction and proves the NP-hardness of MNBC. ▯

Note that the NP-hardness of MNBC is independent of the choice of the norm used over the state space $\mathcal{X}$. Also observe that, since $\{0, 1\}$-programming is strongly NP-hard [37], it is also the case for MNBC. The two results follow.

Corollary 4.6. *$(\mathcal{P}''(\mathcal{F}, u_0, u_1))$ is NP-hard.*

Theorem 4.7. *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.

5. Relaxation schemes.

The two-stage case with only one element in the set $\mathcal{F}^{(u_1)}$ was proven 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 show that this scheme provides bounds that are greater than or equal to the CGRL bound introduced in [19]. 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 prove that this relaxation scheme always gives better bounds than the first relaxation scheme mentioned above, and consequently, better bounds than [19]. 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 (cf. Lemma 4.2). We therefore only focus on relaxing the problem corresponding to the remaining stages $(\mathcal{P}''(\mathcal{F}, u_0, \ldots, u_{T-1}))$.

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

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

subject to

$$\left|\hat{\mathbf{r}}_t - r^{(u_t),k_t}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2 \ \forall (t, k_t) \in \{1, \ldots, T-1\} \times \left\{1, \ldots, n^{(u_t)}\right\}, \tag{5.1}$$

$$\left\|\hat{\mathbf{x}}_{t+1} - y^{(u_t),k_t}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t),k_t}\right\|^2 \ \forall (t, k_t) \in \{0, \ldots, T-1\} \times \left\{1, \ldots, n^{(u_t)}\right\}, \tag{5.2}$$

$$\left\|\hat{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}_{t'+1}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t'}\right\|^2 \ \forall t, t' \in \{0, \ldots, T-2 | u_t = u_{t'}\}, \tag{5.3}$$

$$\hat{\mathbf{x}}_0 = x_0 \ . \tag{5.4}$$

5.1. The ITR relaxation scheme.

A natural way to obtain a relaxation from an optimization problem is to drop some constraints. A particular case of tractable nonconvex 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 (3.4) and keeping one constraint by stage and by type. We therefore obtain a relaxed problem of the form

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

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

subject to

$$\left|\hat{\mathbf{r}}_t - r^{(u_t), \bar{k}_t}\right|^2 \le L_\rho^2 \left\|\hat{\mathbf{x}}_t - x^{(u_t), \bar{k}_t}\right\|^2, \qquad t \in \{1, \ldots, T-1\}, \tag{5.5}$$

$$\left\|\hat{\mathbf{x}}_t - y^{(u_{t-1}), \bar{k}_{t-1}}\right\|^2 \le L_f^2 \left\|\hat{\mathbf{x}}_{t-1} - x^{(u_{t-1}), \bar{k}_{t-1}}\right\|^2, \qquad t \in \{1, \ldots, T-1\}, \tag{5.6}$$

$$\hat{\mathbf{x}}_0 = x_0. \tag{5.7}$$

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

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

Definition 5.1.

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

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

subject to

$$\left|\hat{\mathbf{r}}_t - r^{(u_t), \bar{k}_t}\right|^2 \le L_\rho^2 \left\|\hat{\mathbf{x}}_t - x^{(u_i), \bar{k}_t}\right\|^2, \qquad t \in \{1, \ldots, j\}, \tag{5.8}$$

$$\left\|\hat{\mathbf{x}}_t - y^{(u_{t-1}), \bar{k}_{t-1}}\right\|^2 \le L_f^2 \left\|\hat{\mathbf{x}}_{t-1} - x^{(u_{t-1}), \bar{k}_{t-1}}\right\|^2, \qquad t \in \{1, \ldots, j\}, \tag{5.9}$$

$$\hat{\mathbf{x}}_0 = x_0. \tag{5.10}$$

The initialization of the induction is provided by the following lemma.

Lemma 5.2. *The optimal solution $D''_{ITR}(u_0, u_1, \bar{k}_0, \bar{k}_1)$ to $(\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, u_1, \bar{k}_0, \bar{k}_1))$ is given by*

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

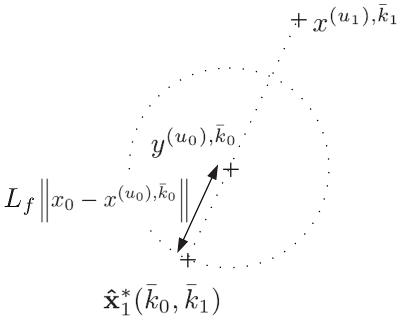
$+\, x^{(u_1),\bar{k}_1}$
$y^{(u_0),\bar{k}_0}$
$L_f \left\| x_0 - x^{(u_0),\bar{k}_0} \right\|$
$\hat{\mathbf{x}}_1^*(\bar{k}_0, \bar{k}_1)$

FIG. 5.1. *A simple geometric algorithm to solve* $(\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, u_1, \bar{k}_0, \bar{k}_1))$.

where

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

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* $L_f \| x_0 - x^{(u_0),\bar{k}_0} \|$.

Proof. This is the maximization of a norm under a norm constraint. This problem is referred to in the literature as the *trust-region subproblem* [9]. In our case, the optimal value for $\hat{\mathbf{x}}_1$—denoted by $\hat{\mathbf{x}}_1^*(\bar{k}_0, \bar{k}_1)$—lies on the same line as $x^{(u_1),\bar{k}_1}$ and $y^{(u_0),\bar{k}_0}$, with $y^{(u_0),\bar{k}_0}$ lying in between $x^{(u_1),\bar{k}_1}$ and $\hat{\mathbf{x}}_1^*(\bar{k}_0, \bar{k}_1)$, the distance between $y^{(u_0),\bar{k}_0}$ and $\hat{\mathbf{x}}_1^*(\bar{k}_0, \bar{k}_1)$ being exactly equal to the distance between x_0 and $x^{(u_0),\bar{k}_0}$. An illustration is given in Figure 5.1. □

LEMMA 5.3. *The optimal solution to* $(\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, \ldots, u_j, \bar{k}_0, \ldots, \bar{k}_j))$ *is given by*

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

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* $L_f \| \hat{\mathbf{x}}_{t-1}^*(\bar{k}_0, \ldots, \bar{k}_{t-1}) - x^{(u_{t-1}),\bar{k}_{t-1}} \|$.

Proof. We proceed by induction. The basis of the induction is provided by Lemma 5.2. We assume that the statement is correct for the $(j-1)$th optimization problem $(\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{j-1}, \bar{k}_0, \ldots, \bar{k}_{j-1}))$ and we show that it is also true for the jth problem. $\hat{\mathbf{x}}_j$ is constrained by a single ball (5.9). So, if the right-hand side of (5.9) is fixed, the optimal solution $\hat{\mathbf{x}}_j^*$ is induced by the same geometry as Lemma 5.2 (see Figure 5.1). It is therefore profitable to maximize the right-hand side of (5.9), which resorts to solving $(\mathcal{Q}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{j-1}, \bar{k}_0, \ldots, \bar{k}_{j-1}))$. The result follows by induction. □

THEOREM 5.4. *The solution to* $(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1}))$ *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

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

and, if $y^{(u_{t-1}),\bar{k}_{t-1}} = x^{(u_t),\bar{k}_t}$, $\hat{\mathbf{x}}_t^*(\bar{k}_0,\dots,\bar{k}_t)$ *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* $L_f \|\hat{\mathbf{x}}_{t-1}^*(\bar{k}_0,\dots,\bar{k}_{t-1}) - x^{(u_{t-1}),\bar{k}_{t-1}}\|$.

Proof. Observe that $\hat{\mathbf{r}}_t$ is constrained by one interval for all t. Therefore, as we want to minimize $\hat{\mathbf{r}}_t$, if the right-hand side of (5.5) is fixed, then $\hat{\mathbf{r}}_t^*$ is given by

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

In order to minimize $\hat{\mathbf{r}}_t$, it is profitable to maximize the right-hand side of (5.5), which resorts to solving $\mathcal{Q}''_{ITR}(\mathcal{F}, u_0,\dots,u_t,\bar{k}_0,\dots,\bar{k}_t)$. Since the value of $\hat{\mathbf{x}}_j$ is the same in every optimal solution of every $\mathcal{Q}''_{ITR}(\mathcal{F}, u_0,\dots,u_i,\bar{k}_0,\dots,\bar{k}_i)$ with $i \geq j$, then the optimal values of $\hat{\mathbf{x}}_t$ are provided by the solution of $\mathcal{Q}''_{ITR}(\mathcal{F}, u_0,\dots,u_{T-1},\bar{k}_0,\dots,\bar{k}_{T-1})$ (see Lemma 5.3), and the result follows. □

Solving $(\mathcal{P}_{ITR}(\mathcal{F}, u_0,\dots,u_{T-1},\bar{k}_0,\dots,\bar{k}_{T-1}))$ provides us with a family of relaxations for our initial problem by considering any combination $(\bar{k}_0,\dots,\bar{k}_{T-1})$ 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

$$
B_{ITR}(\mathcal{F}, u_0,\dots,u_{T-1})
$$

the bound made of the sum of the solution of the first-stage problem and the maximal ITR relaxation of the problem $(\mathcal{P}''_{ITR}(\mathcal{F}, u_0,\dots,u_{T-1},\bar{k}_0,\dots,\bar{k}_{T-1}))$ over all possible couples of constraints, we have the following.

Definition 5.5 (ITR bound $B_{ITR}(\mathcal{F}, u_0,\dots,u_{T-1})$).

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

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

Corollary 5.6.

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

5.2. 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 $(\mathcal{P}''(\mathcal{F}, u_0,\dots,u_{T-1}))$. If we multiply the constraints (5.1) by dual variables $\mu_{t,k_t} \geq 0$, the constraints (5.2) by dual variables $\lambda_{t,k_t} \geq 0$, and the constraints (5.3) by dual variables $\nu_{t,t'} \geq 0$, we get the Lagrangian dual problem $(\mathcal{P}''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}))$:

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

$$\max_{\substack{\nu_{t,t'} \in \mathbb{R} \\ \lambda_{t,k_t} \in \mathbb{R} \\ \mu_{t,k_t} \in \mathbb{R}}} \quad \min_{\substack{\hat{\mathbf{r}}_1, \ldots, \hat{\mathbf{r}}_{T-1} \in \mathbb{R} \\ \hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_{T-1} \in \mathcal{X}}}$$

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

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 $(\mathcal{P}''(\mathcal{F}, u_0, \ldots, u_{T-1}))$ [24]. 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 [48, 10, 34].

DEFINITION 5.7 (Lagrangian 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,*

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

5.3. Comparing the bounds. The CGRL algorithm proposed in [17, 19] for addressing the min max problem uses the procedure described in [16] 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.

5.3.1. Trust region versus CGRL. We first recall the definition of the CGRL bound.

DEFINITION 5.8 (CGRL bound $B_{CGRL}(\mathcal{F}, u_0, \ldots, u_{T-1})$).

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

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

THEOREM 5.9.

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

Proof. Let $(k_0^*, \dots, k_{T-1}^*) \in \{1, \dots, n^{(u_0)}\} \times \cdots \times \{1, \dots, n^{(u_{T-1})}\}$ be such that

$$B_{CGRL}(\mathcal{F}, u_0, \dots, u_{T-1}) = r^{(u_0),k_0^*} - L_\rho(1 + L_f + \cdots + L_f^{T-2}) \left\| x^{(u_0),k_0^*} - x_0 \right\| \\ + \cdots + r^{(u_1),k_1^*} - L_\rho \left\| y^{(u_{T-2}),k_{T-2}^*} - x^{(u_{T-1}),k_{T-1}^*} \right\| .$$

Now, let us consider the solution $B''_{ITR}(\mathcal{F}, u_0, \dots, u_{T-1}, k_0^*, \dots, k_{T-1}^*)$ of the problem $(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \dots, u_{T-1}, k_0^*, \dots, k_{T-1}^*))$, and let us denote by $\beta(u_0, \dots, u_{T-1}, k_0^*, \dots, k_{T-1}^*)$ the bound obtained if, in the definition of the value of $\hat{\mathbf{r}}_0^*$ given in Corollary 4.2, we fix the value of k'_0 to k_0^* instead of maximizing over all possible k'_0:

$$\beta\big(u_0, \dots, u_{T-1}, k_0^*, \dots, k_{T-1}^*\big) = r^{(u_0),k_0^*} - L_\rho \left\| x_0 - x^{(u_0),k_0^*} \right\| \\ + B''_{ITR}(\mathcal{F}, u_0, \dots, u_{T-1}, k_0^*, \dots, k_{T-1}^*).$$

Since $r^{(u_0),k_0^*} - L_\rho \| x_0 - x^{(u_0),k_0^*} \|$ is smaller than or equal to the solution $\hat{\mathbf{r}}_0^*$ of $(\mathcal{P}'(\mathcal{F}, u_0))$, one has

$$B_{ITR}(\mathcal{F}, u_0, \dots, u_{T-1}, k_0^*, \dots, k_{T-1}^*) \geq \beta\left(u_0, \dots, u_{T-1}, k_0^*, \dots, k_{T-1}^*\right) \ . \tag{5.11}$$

Back to the solution of the ITR relaxation (see Theorem 5.4), we have that

$$\beta\left(u_0, \dots, u_{T-1}, k_0^*, \dots, k_{T-1}^*\right) = \sum_{t=0}^{T-1} r^{(u_t),k_t} - L_\rho \left\| \hat{\mathbf{x}}_t^*(k_0^*, \dots, k_t^*) - x^{(u_t),k_t^*} \right\| \ .$$

Let $t \geq 1$. We now recursively compute the value of $\| \hat{\mathbf{x}}_t^*(k_0^*, \dots, k_t^*) - x^{(u_t),k_t^*} \|$.

- First case: $\left\| y^{(u_{t-1}),k_{t-1}^*} - x^{(u_t),k_t^*} \right\| > 0$.

From Theorem 5.4, we have

$$\left\| \hat{\mathbf{x}}_t^*(k_0^*, \dots, k_t^*) - x^{(u_t),k_t^*} \right\| \\ = \left\| y^{(u_{t-1}),k_{t-1}^*} - x^{(u_t),k_t^*} \right\| \left(1 + L_f \frac{\left\| \hat{\mathbf{x}}_{t-1}^*(k_0^*, \dots, k_{t-1}^*) - x^{(u_{t-1}),k_{t-1}^*} \right\|}{\left\| y^{(u_{t-1}),k_{t-1}^*} - x^{(u_t),k_t^*} \right\|} \right) \\ = \left\| y^{(u_{t-1}),k_{t-1}^*} - x^{(u_t),k_t^*} \right\| + L_f \left\| \hat{\mathbf{x}}_{t-1}^*(k_0^*, \dots, k_{t-1}^*) - x^{(u_{t-1}),k_{t-1}^*} \right\| \ .$$

- Second case: $\| y^{(u_{t-1}),k_{t-1}^*} - x^{(u_t),k_t^*} \| = 0$.

From Theorem 5.4, we know that $\hat{\mathbf{x}}_t^*(\bar{k}_0, \dots, \bar{k}_t)$ 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 $L_f \| \hat{\mathbf{x}}_{t-1}^*(\bar{k}_0, \dots, \bar{k}_{t-1}) - x^{(u_{t-1}),\bar{k}_{t-1}} \|$, so it means that

$$\left\| \hat{\mathbf{x}}_t^*(k_0^*, \dots, k_t^*) - x^{(u_t),k_t^*} \right\| = L_f \left\| \hat{\mathbf{x}}_{t-1}^*(k_0^*, \dots, k_{t-1}^*) - x^{(u_{t-1}),k_{t-1}^*} \right\| \ .$$

In both cases, we straightforwardly obtain the following result:

$$\begin{aligned}&\left\|\hat{\mathbf{x}}_t^*(k_0^*,\dots,k_t^*) - x^{(u_t),k_t^*}\right\| \\ &\quad= \left\|y^{(u_{t-1}),k_{t-1}^*} - x^{(u_t),k_t^*}\right\| + L_f \left\|y^{(u_{t-2}),k_{t-2}^*} - x^{(u_{t-1}),k_{t-1}^*}\right\| \\ &\qquad + \dots + L_f^{t-1}\left\|x_0 - x^{(u_0),k_0^*}\right\| \ .\end{aligned}$$

Going back to $\beta(u_0,\dots,u_{T-1},k_0^*,\dots,k_{T-1}^*)$, we have

$$\begin{aligned}&\beta\left(u_0,\dots,u_{T-1},k_0^*,\dots,k_{T-1}^*\right) \\ &\qquad= \sum_{t=0}^{T-1} r^{(u_t),k_t} - L_\rho\left(\sum_{t'=0}^{t} L_f^{t-t'}\left\|y^{(u_{t'-1}),k_{t'-1}^*} - x^{(u_{t'}),k_{t'}^*}\right\|\right).\end{aligned}$$

By reorganizing the terms of the sum, one directly obtains the value of the CGRL bound

$$\beta\big(u_0,\dots,u_{T-1},k_0^*,\dots,k_{T-1}^*\big) = B_{CGRL}(\mathcal{F},u_0,\dots,u_{T-1}) \ . \tag{5.12}$$

The final result is given by combining (5.11) and (5.12). □

From the previous proof, one can observe that the gap between the CGRL bound and the ITR bound is only due to the resolution of $(\mathcal{P}'(\mathcal{F},u_0))$. Note that in the case where k_0^* also belongs to the set $\arg\max_{k_0\in\{1,\dots,n^{(u_0)}\}} r^{(u_0),k_0} - L_\rho\|x^{(u_0),k_0} - x_0\|$, then the bounds are equal. The two corollaries follow.

COROLLARY 5.10. *Let* $k_0^* \in \{1,\dots,n^{(u_0)}\},\dots,k_{T-1}^* \in \{1,\dots,n^{(u_{T-1})}\}$ *be such that*

$$\begin{aligned}B_{CGRL}(\mathcal{F},u_0,\dots,u_{T-1}) &= r^{(u_0),k_0^*} - L_\rho(1+L_f+\dots+L_f^{T-2})\left\|x^{(u_0),k_0^*} - x_0\right\| \\ &\quad + \dots + r^{(u_1),k_1^*} - L_\rho\left\|y^{(u_{T-2}),k_{T-2}^*} - x^{(u_{T-1}),k_{T-1}^*}\right\|.\end{aligned}$$

Then,

$$\begin{aligned}&\left(k_0^* \in \underset{k_0\in\{1,\dots,n^{(u_0)}\}}{\arg\max}\ r^{(u_0),k_0} - L_\rho\left\|x^{(u_0),k_0} - x_0\right\|\right) \\ &\qquad\qquad \implies B_{CGRL}(\mathcal{F},u_0,\dots,u_{T-1}) = B_{ITR}(\mathcal{F},u_0,\dots,u_{T-1}) \ .\end{aligned}$$

COROLLARY 5.11.

$$\begin{aligned}\left(\forall t\in\{0,\dots,T-1\}, n^{(u_t)} = 1\right) \implies &B_{CGRL}(\mathcal{F},u_0,\dots,u_{T-1}) \\ &= B_{ITR}(\mathcal{F},u_0,\dots,u_{T-1}) \\ &= B^*(\mathcal{F},u_0,\dots,u_{T-1}) \ .\end{aligned}$$

5.3.2. Lagrangian relaxation versus ITR relaxation. In this section, we prove that the lower bound obtained with the Lagrangian relaxation is always greater than or equal to the ITR bound. To do so, we prove that strong duality holds for the

Lagrangian dual of $(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1}))$ for a given $(\bar{k}_0, \ldots, \bar{k}_{T-1})$. The Lagrangian dual of $(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1}))$ reads

$$
\begin{aligned}
(LD''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1}))) : \quad & \max_{\substack{\lambda_1, \ldots, \lambda_{T-1} \in \mathbb{R} \\ \mu_1, \ldots, \mu_{T-1} \in \mathbb{R}}} \quad \min_{\substack{\hat{\mathbf{r}}_1, \ldots, \hat{\mathbf{r}}_{T-1} \in \mathbb{R} \\ \hat{\mathbf{x}}_0, \ldots, \hat{\mathbf{x}}_{T-1} \in \mathcal{X}}} \\
& \hat{\mathbf{r}}_1 + \cdots + \hat{\mathbf{r}}_{T-1} \\
& + \mu_1 \left(\left| \hat{\mathbf{r}}_1 - r^{(u_1),\bar{k}_1} \right|^2 - L_\rho^2 \left\| \hat{\mathbf{x}}_1 - x^{(u_1),\bar{k}_1} \right\|^2 \right) \\
& \qquad \vdots \\
& + \mu_{T-1} \left(\left| \hat{\mathbf{r}}_{T-1} - r^{(u_{T-1}),\bar{k}_{T-1}} \right|^2 - L_\rho^2 \left\| \hat{\mathbf{x}}_{T-1} - x^{(u_{T-1}),\bar{k}_{T-1}} \right\|^2 \right) \\
& + \lambda_1 \left(\left\| \hat{\mathbf{x}}_1 - y^{(u_0),\bar{k}_0} \right\|^2 - L_f^2 \left\| \hat{\mathbf{x}}_0 - x^{(u_0),\bar{k}_0} \right\|^2 \right) \\
& \qquad \vdots \\
& + \lambda_{T-1} \left(\left\| \hat{\mathbf{x}}_{T-1} - y^{(u_{T-2}),\bar{k}_{T-2}} \right\|^2 - L_f^2 \left\| \hat{\mathbf{x}}_{T-2} - x^{(u_{T-2}),\bar{k}_{T-2}} \right\|^2 \right).
\end{aligned}
$$

We now consider the inner optimization problem in the expression of the Lagrangian dual. It can be written, by considering λ_t, μ_t fixed, as a sum of terms that each include one variable, i.e.,

$$
\begin{aligned}
& \min_{\substack{\hat{\mathbf{r}}_1, \ldots, \hat{\mathbf{r}}_{T-1} \in \mathbb{R} \\ \hat{\mathbf{x}}_0, \ldots, \hat{\mathbf{x}}_{T-1} \in \mathcal{X}}} \sum_{t=1}^{T-1} \left(\hat{\mathbf{r}}_t + \mu_t \left| \hat{\mathbf{r}}_t - r^{(u_t),k_t} \right|^2 \right) \\
& \qquad + \sum_{t=1}^{T-1} \left(\left\| \hat{\mathbf{x}}_t - x^{(u_t),\bar{k}_t} \right\|^2 (-\mu_t L_\rho^2 - \lambda_{t+1} L_f^2) + \left\| \hat{\mathbf{x}}_t - y^{(u_{t-1}),\bar{k}_{t-1}} \right\|^2 \lambda_t \right) \qquad (5.13) \\
& \qquad + \lambda_1 L_f^2 \left\| \hat{x}_0 - x^{(u_0),\bar{k}_0} \right\|^2,
\end{aligned}
$$

where we define, for ease of notation, $\lambda_T \triangleq 0$. We first observe that the objective function of the optimization problem (5.13) goes to $-\infty$ unless

$$
(5.14) \quad \forall t \in \{1, \ldots, T-1\}, \mu_t > 0 \\
\text{and} \begin{cases} \lambda_t - \mu_t L_\rho^2 - \lambda_{t+1} L_f^2 > 0 \\ \text{or} \\ \lambda_t - \mu_t L_\rho^2 - \lambda_{t+1} L_f^2 = 0 \text{ with } y^{(u_{t-1}),\bar{k}_{t-1}} = x^{(u_t),\bar{k}_t}. \end{cases}
$$

The condition (5.14) comes from the fact that, since the objective function is a polynomial of degree 2, for each variable, there are two ways to obtain a finite minimum: either (i) the coefficient of the term of degree 2 is positive or (ii) the coefficient of degree 2 and the corresponding coefficient of degree 1 are both equal to 0. These two conditions lead to the two cases of (5.14). In particular, for the case

$\lambda_t - \mu_t L_\rho^2 - \lambda_{t+1} L_f^2 = 0$ with $y^{(u_{t-1}),\bar{k}_{t-1}} = x^{(u_t),\bar{k}_t}$, we have

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

Since the outer optimization problem is a maximization problem, we henceforth assume that condition (5.14) holds. Note that the objective function of (5.13) is a sum of univariate functions, which implies that we can solve one single optimization problem for each variable. We start with variables $\hat{\mathbf{r}}_t$.

LEMMA 5.12. *Let $\mu_t > 0$. The optimal solution to the problem*

$$\min_{\hat{\mathbf{r}}_t \in \mathbb{R}} \quad \hat{\mathbf{r}}_t + \mu_t \left|\hat{\mathbf{r}}_t - r^{(u_t),\bar{k}_t}\right|^2$$

is given by

$$\hat{\mathbf{r}}_t^* = r^{(u_t),\bar{k}_t} - \frac{1}{2\mu_t}.$$

Proof. It follows directly from the fact that we minimize the quadratic univariate function

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

We now turn to the optimization problems involving one variable $\hat{\mathbf{x}}_t$. It is formally defined as

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

LEMMA 5.13. *Assume that $x^{(u_t),\bar{k}_t} \neq y^{(u_{t-1}),\bar{k}_{t-1}}$. The optimal solution $\hat{\mathbf{x}}_t^*$ to $(\mathcal{R}_t)$ lies on the same line as $x^{(u_t),\bar{k}_t}$ and $y^{(u_{t-1}),\bar{k}_{t-1}}$.*

Proof. We consider the orthogonal projection of $\hat{\mathbf{x}}_t^*$ onto $\operatorname{aff}(x^{(u_t),\bar{k}_t}, y^{(u_{t-1}),\bar{k}_{t-1}})$ that we denote by $\bar{x}_t$. We assume by contradiction that $\hat{\mathbf{x}}_t^* \neq \bar{x}_t$. From orthogonality we have

$$\left\|\hat{\mathbf{x}}_t^* - x^{(u_t),\bar{k}_t}\right\|^2 = \left\|\hat{\mathbf{x}}_t^* - \bar{x}_t\right\|^2 + \left\|\bar{x}_t - x^{(u_t),\bar{k}_t}\right\|^2, \tag{5.15}$$

$$\left\|\hat{\mathbf{x}}_t^* - y^{(u_{t-1}),\bar{k}_{t-1}}\right\|^2 = \left\|\hat{\mathbf{x}}_t^* - \bar{x}_t\right\|^2 + \left\|\bar{x}_t - y^{(u_{t-1}),\bar{k}_{t-1}}\right\|^2. \tag{5.16}$$

Therefore if we substitute $\bar{x}_t$ in the objective function of $(\mathcal{R}_t)$, we obtain, using (5.15) and (5.16),

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

The two first terms of the right-hand side of the last equation correspond to the value of the objective function with $\hat{\mathbf{x}}_t^*$ as a feasible solution, and the last term is always negative since $\lambda_t - \mu_t L_\rho^2 - \lambda_{t+1} L_f^2 > 0$. Therefore, $\bar{x}_t$ has a lower objective value than $\hat{\mathbf{x}}_t^*$, a contradiction. □

LEMMA 5.14. *Assume that $x^{(u_t),\bar{k}_t} \neq y^{(u_{t-1}),\bar{k}_{t-1}}$. An optimal solution $\hat{\mathbf{x}}_t^*$ of $(\mathcal{R}_t)$ is such that*

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

where $\lambda_T = 0$ by convention.

Proof. We know from Lemma 5.13 the line to which $\hat{\mathbf{x}}_t^*$ belongs. Finding the optimal solution resorts to finding the minimum of a univariate quadratic function. The complete calculation is left as an exercise to the reader. □

We also have the straightforward following lemma.

LEMMA 5.15. *Assume that $x^{(u_t),\bar{k}_t} = y^{(u_{t-1}),\bar{k}_{t-1}}$ and $\lambda_t - \mu_t L_\rho^2 - \lambda_{t+1} L_f^2 = 0$. Then, the objective function of $(\mathcal{R}_t)$ is identically equal to zero and $\hat{\mathbf{x}}_t^*$ can be any vector of $\mathcal{X}$.*

We are now ready to prove the main result of this section.

THEOREM 5.16. *Strong duality holds for the Lagrangian relaxation of the ITR problem $(LD''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1}))$.*

Proof. We will prove that there exist $\lambda_1, \ldots, \lambda_{T-1}, \mu_1, \ldots, \mu_{T-1}$ satisfying conditions given by (5.14) and such that the corresponding optimal solution to the inner optimization problems which is characterized by Lemmas 5.12, 5.13, 5.14, and 5.15, is also an optimal solution to $(\mathcal{P}''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, \bar{k}_0, \ldots, \bar{k}_{T-1}))$. Since the latter is tight for all constraints, this implies that the objective value for the Lagrangian relaxation is equal to the objective function of the initial problem and proves the result.

We exhibit appropriate values of the dual variables by backward induction. Let us first assume that $\|\hat{\mathbf{x}}_t^*(\bar{k}_0, \ldots, \bar{k}_t) - x^{(u_t),\bar{k}_t}\| > 0 \; \forall t \in \{0, \ldots, T-1\}$.

- Basis. By identifying the values of $\hat{\mathbf{r}}_{T-1}^*$ obtained with the ITR relaxation (Theorem 5.4) and with the Lagrangian relaxation (Lemma 5.12), we get

$$\mu_{T-1}^* = \frac{1}{2L_\rho \left\|\hat{\mathbf{x}}_{T-1}^*(\bar{k}_0, \ldots, \bar{k}_{T-1}) - x^{(u_{T-1}),\bar{k}_{T-1}}\right\|}$$

which is positive by assumption. Similarly, by identifying the values of $\hat{\mathbf{x}}_{T-1}^*$ obtained with the ITR relaxation (Theorem 5.4) and with the Lagrangian relaxation (Lemma 5.14), we get

$$\lambda_{T-1}^* = \frac{\left\|x^{(u_{T-1}),\bar{k}_{T-1}} - y^{(u_{T-2}),\bar{k}_{T-2}}\right\| \mu_{T-1}^* L_\rho^2}{L_f \left\|\hat{\mathbf{x}}_{T-2}^* - x^{(u_{T-2}),\bar{k}_{T-2}}\right\|} + \mu_{T-1}^* L_\rho^2.$$

Observe that λ_{T-1}^* and μ_{T-1}^* satisfy the conditions given by (5.14). Indeed, if $x^{(u_{T-1}),\bar{k}_{T-1}} \neq y^{(u_{T-2}),\bar{k}_{T-2}}$, then the first case of (5.14) holds, whereas if $x^{(u_{T-1}),\bar{k}_{T-1}} = y^{(u_{T-2}),\bar{k}_{T-2}}$, then the second case of (5.14) holds.

• Inductive step. By identifying the values of $\hat{\mathbf{r}}_t^*$ obtained with the ITR relaxation (Theorem 5.4) and with the Lagrangian relaxation (Lemma 5.12), we get

$$\mu_t^* = \frac{1}{2L_\rho \left\| \hat{\mathbf{x}}_t^*(\bar{k}_0, \ldots, \bar{k}_t) - x^{(u_t), \bar{k}_t} \right\|}$$

which is positive by assumption. Similarly, by identifying the values of $\hat{\mathbf{x}}_t^*$ obtained with the ITR relaxation (Theorem 5.4) and with the Lagrangian relaxation (Lemma 5.14), we get

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

Observe again that λ_t^* and μ_t^* satisfy the conditions given by (5.14).

We now discuss the case where

$$\exists t_0 \in \{0, \ldots, T-1\}, \qquad \left\| \hat{\mathbf{x}}_{t_0}^*(\bar{k}_0, \ldots, \bar{k}_{t_0}) - x^{(u_{t_0}), \bar{k}_{t_0}} \right\| = 0.$$

According to Theorem 5.4, if

$$\left(\left\| \hat{\mathbf{x}}_{t_0}^*(\bar{k}_0, \ldots, \bar{k}_{t_0}) - x^{(u_{t_0}), \bar{k}_{t_0}} \right\| = 0 \right) \Longrightarrow \begin{cases} \left\| \hat{\mathbf{x}}_{t_0-1}^*(\bar{k}_0, \ldots, \bar{k}_{t_0-1}) - x^{(u_{t_0-1}), \bar{k}_{t_0-1}} \right\| = 0, \\ \left\| y^{(u_{t_0-1}), \bar{k}_{t_0-1}} - x^{(u_{t_0}), \bar{k}_{t_0}} \right\| = 0, \end{cases}$$

then this implies, by backward induction, that

$$\forall t \in \{0, \ldots, t_0\}, \quad \begin{aligned} x^{(u_t), \bar{k}_t} &= y^{(u_{t-1}), \bar{k}_{t-1}} \\ &= \hat{\mathbf{x}}_t^*(\bar{k}_0, \ldots, \bar{k}_t). \end{aligned}$$

The results follows by choosing $\mu_t^* \to \infty$ $\forall t \in \{0, \ldots, t_0\}$ and the λ_t^* are chosen according to conditions (5.14). This case corresponds to a specific configuration where the sample of data $\mathcal{F}$ contains a sequence of transitions which forms a trajectory that starts from x_0 and exactly follows the sequence of actions $(u_0, \ldots, u_{T-1})$ until t_0. In such a specific case, the optimization problem is trivial for all time steps preceding t_0. □

THEOREM 5.17.

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

Proof. Let $(k_0^*, \ldots, k_{T-1}^*) \in \{1, \ldots, n^{(u_0)}\} \times \cdots \times \{1, \ldots, n^{(u_{T-1})}\}$ be such that

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

Considering $(\bar{k}_0, \ldots, \bar{k}_{T-1}) = (k_0^*, \ldots, k_{T-1}^*)$ in Theorem 5.16, we have

$$B_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}) = \hat{\mathbf{r}}_0^* + B''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}, k_0^*, \ldots, k_{T-1}^*). \tag{5.17}$$

Then, one can observe that the Lagrangian relaxation $(LD''_{ITR}(\mathcal{F}, u_0, \ldots, u_{T-1}, k_0^*, \ldots, k_{T-1}^*))$—from which $B''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}, k_0^*, \ldots, k_{T-1}^*)$ is computed—is also a relaxation of the problem $(\mathcal{P}''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1})$ for which all the dual variables corresponding to constraints that are not related to the sequence of transitions $(x^{(u_0), k_0^*},$

$r^{(u_0),k_0^*}, y^{(u_0),k_0^*}), \ldots, (x^{(u_{T-1}),k_{T-1}^*}, r^{(u_{T-1}),k_{T-1}^*}, y^{(u_{T-1}),k_{T-1}^*})$ would be forced to zero. We therefore have

$$B''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}, k_0^*, \ldots, k_{T-1}^*) \leq B''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}). \tag{5.18}$$

By definition of the Lagrangian relaxation bound $B_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1})$, we have

$$B_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}) = \hat{\mathbf{r}}_0^* + B''_{LD}(\mathcal{F}, u_0, \ldots, u_{T-1}). \tag{5.19}$$

Equations (5.17), (5.18), and (5.19) finally give

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

5.3.3. Bounds inequalities: Summary. We summarize in the following theorem all the results that were obtained in the previous sections.

THEOREM 5.18. $\forall (u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$,

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

Proof. The inequality

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

is a straightforward consequence of Theorems 5.9 and 5.17. The inequality

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

is a property of the Lagrangian relaxation, and the inequality

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

comes from the definition of $B^*(\mathcal{F}, u_0, \ldots, u_{T-1})$. $\square$

5.4. Convergence properties. We finally propose to analyze the convergence of the bounds, as well as the sequences of actions that lead to the maximization of the bounds, when the sample dispersion decreases towards zero. We assume in this section that the state space $\mathcal{X}$ is bounded:

$$\exists C_{\mathcal{X}} > 0 : \forall (x, x') \in \mathcal{X}^2, \qquad \|x - x'\| \leq C_{\mathcal{X}}.$$

Let us now introduce the sample dispersion.

DEFINITION 5.19 (sample dispersion). *Since $\mathcal{X}$ is bounded, one has*

$$\exists\ \alpha > 0 : \forall u \in \mathcal{U}\ , \qquad \sup_{x \in \mathcal{X}} \quad \min_{k \in \left\{1, \ldots, n^{(u)}\right\}} \left\| x^{(u),k} - x \right\| \leq \alpha. \tag{5.20}$$

The smallest α which satisfies (5.20) *is named the sample dispersion and is denoted by $\alpha^*(\mathcal{F})$.*

Intuitively, the sample dispersion $\alpha^*(\mathcal{F})$ can be seen as the radius of the largest nonvisited state space area.

5.4.1. Bounds. We analyze in this subsection the tightness of the ITR and the Lagrangian relaxation lower bounds as a function of the sample dispersion.

LEMMA 5.20. $\exists\ C > 0 : \forall (u_0, u_1) \in \mathcal{U}^2, \forall \beta \in \{B_{CGRL}(\mathcal{F}, u_0, \dots, u_{T-1}), B_{ITR}(\mathcal{F}, u_0, \dots, u_{T-1}), B_{LD}(\mathcal{F}, u_0, \dots, u_{T-1})\}$,

$$J(u_0, \dots, u_{T-1}) - \beta \le C\alpha^*(\mathcal{F}).$$

Proof. The proof for the case where $\beta = B_{CGRL}(\mathcal{F}, u_0, \dots, u_{T-1})$ is given in [17], and the remainder of the proof directly follows from Theorem 5.18. □

We therefore have the following theorem.

THEOREM 5.21. $\forall (u_0, \dots, u_{T-1}) \in \mathcal{U}^T, \forall \beta \in \{B_{CGRL}(\mathcal{F}, u_0, \dots, u_{T-1}), B_{ITR}(\mathcal{F}, u_0, \dots, u_{T-1}), B_{LD}(\mathcal{F}, u_0, \dots, u_{T-1})\}$,

$$\lim_{\alpha^*(\mathcal{F}) \to 0} J(u_0, \dots, u_{T-1}) - \beta = 0\ .$$

5.4.2. Bound-optimal sequences of actions. In the following, we denote by $B^{(*)}_{CGRL}(\mathcal{F})$ (resp., $B^{(*)}_{ITR}(\mathcal{F})$ and $B^{(*)}_{LD}(\mathcal{F})$) the maximal CGRL bound (resp., the maximal ITR bound and maximal Lagrangian bound) over the set of all possible sequences of actions, i.e., which is shown by the following.

DEFINITION 5.22 (maximal bounds).

$$B^{(*)}_{T,CGRL}(\mathcal{F}) \triangleq \max_{(u_0,\dots,u_{T-1}) \in \mathcal{U}^T} B_{CGRL}(\mathcal{F}, u_0, \dots, u_{T-1}),$$
$$B^{(*)}_{T,ITR}(\mathcal{F}) \triangleq \max_{(u_0,\dots,u_{T-1}) \in \mathcal{U}^T} B_{ITR}(\mathcal{F}, u_0, \dots, u_{T-1}),$$
$$B^{(*)}_{T,LD}(\mathcal{F}) \triangleq \max_{(u_0,\dots,u_{T-1}) \in \mathcal{U}^T} B_{LD}(\mathcal{F}, u_0, \dots, u_{T-1}).$$

We also denote by $(u_0, \dots, u_{T-1})^{CGRL}_{\mathcal{F}}$ (resp., $(u_0, \dots, u_{T-1})^{ITR}_{\mathcal{F}}$ and $(u_0, \dots, u_{T-1})^{LD}_{\mathcal{F}}$) three sequences of actions that maximize the bounds.

DEFINITION 5.23 (bound-optimal sequences of actions).

$$(u_0, \dots, u_{T-1})^{CGRL}_{\mathcal{F}} \in \left\{(u_0, \dots, u_{T-1}) \in \mathcal{U}^T | B_{CGRL}(\mathcal{F}, u_0, \dots, u_{T-1}) = B^{(*)}_{T,CGRL}(\mathcal{F})\right\},$$
$$(u_0, \dots, u_{T-1})^{ITR}_{\mathcal{F}} \in \left\{(u_0, \dots, u_{T-1}) \in \mathcal{U}^T | B_{ITR}(\mathcal{F}, u_0, \dots, u_{T-1}) = B^{(*)}_{T,ITR}(\mathcal{F})\right\},$$
$$(u_0, \dots, u_{T-1})^{LD}_{\mathcal{F}} \in \left\{(u_0, \dots, u_{T-1}) \in \mathcal{U}^T | B_{LD}(\mathcal{F}, u_0, \dots, u_{T-1}) = B^{(*)}_{T,LD}(\mathcal{F})\right\}.$$

We finally give in this section a last theorem that shows the convergence of the sequences of actions $(u_0, \dots, u_{T-1})^{CGRL}_{\mathcal{F}}$, $(u_0, \dots, u_{T-1})^{ITR}_{\mathcal{F}}$, and $(u_0, \dots, u_{T-1})^{LD}_{\mathcal{F}}$ towards optimal sequences of actions—i.e., sequences of actions that lead to an optimal return J^*_T—when the sample dispersion $\alpha^*(\mathcal{F})$ decreases towards zero.

THEOREM 5.24. *Let $\mathfrak{J}^*_T$ be the set of optimal sequences of actions*

$$\mathfrak{J}^*_T \triangleq \left\{(u_0, \dots, u_{T-1}) \in \mathcal{U}^T | J(u_0, \dots, u_{T-1}) = J^*_T\right\},$$

and let us suppose that $\mathfrak{J}^*_T \neq \mathcal{U}^T$ (if $\mathfrak{J}^*_T = \mathcal{U}^T$, the search for an optimal sequence of actions is indeed trivial). We define

$$\epsilon \triangleq \min_{(u_0,\dots,u_{T-1}) \in \mathcal{U}^T \setminus \mathfrak{J}^*_T} \left\{J^*_T - J(u_0, \dots, u_{T-1})\right\}.$$

Then, $\forall (\tilde{u}_0, \dots, \tilde{u}_{T-1})_{\mathcal{F}} \in \{(u_0, \dots, u_{T-1})_{\mathcal{F}}^{CGRL}, (u_0, \dots, u_{T-1})_{\mathcal{F}}^{ITR}, (u_0, \dots, u_{T-1})_{\mathcal{F}}^{LD}\}$,

$$\left(C\alpha^*(\mathcal{F}) < \epsilon \right) \implies (\tilde{u}_0, \dots, \tilde{u}_{T-1})_{\mathcal{F}} \in \mathfrak{J}_T^*.$$

The proof of this theorem is given in [17] in the specific case of the CGRL bound, and straightforwardly follows for other bounds. The extension of this result to the ITR and Lagrangian bounds is a direct consequence of Theorem 5.18.

5.4.3. Remark. It is important to notice that the tightness of the bounds resulting from the relaxation schemes proposed in this paper does not depend *explicitly* on the sample dispersion (which suffers from the curse of dimensionality, i.e., that exponentially depends on the dimension of the state space), but depends rather on the initial state for which the sequence of actions is computed and on the local concentration of samples around the actual (unknown) trajectories of the system. Therefore, this may lead to cases where the bounds are tight for some specific initial states, even if the sample does not cover every area of the state space well enough.

6. Experimental results.

We provide some experimental results to illustrate the theoretical properties of the CGRL, ITR, and Lagrangian bounds given below. We compare the tightness of the bounds, as well as the performances of the bound-optimal sequences of actions, on an academic benchmark.

6.1. Benchmark.

The optimization horizon T is chosen equal to 2. We consider a linear benchmark whose dynamics is defined as follows:

$$\forall (x, u) \in \mathcal{X} \times \mathcal{U}, \qquad f(x, u) = x + 3.1416 \times u \times 1_d,$$

where $1_d \in \mathbb{R}^d$ denotes a d-dimensional vector for which each component is equal to 1. The reward function is defined as follows:

$$\forall (x, u) \in \mathcal{X} \times \mathcal{U}, \qquad \rho(x, u) = \sum_{i=1}^{d} x(i),$$

where $x(i)$ denotes the ith component of x. The state space $\mathcal{X}$ is included in $\mathbb{R}^d$ and the finite action space is equal to $\mathcal{U} = \{0, 0.1\}$. The system dynamics f is 1-Lipschitz continuous and the reward function is $\sqrt{d}$-Lipschitz continuous. The initial state of the system is set to

$$x_0 = 0.5772 \times 1_d \ .$$

The dimension d of the state space is set to $d = 2$. In all our experiments, the computation of the Lagrangian relaxations, which requires us to solve a conic-quadratic program (see [15] for a detailed description of the two-stage case), is done using SeDuMi [46].

6.2. Protocol and results.

6.2.1. Typical run. For different cardinalities $c_i = 2i^2, i = 1, \dots, 15$, we generate a sample of transitions $\mathcal{F}_{c_i}$ using a grid over $[0,1]^d \times \mathcal{U}$, as follows: $\forall u \in \mathcal{U}$,

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

and

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

We report in Figure 6.1 the values of the maximal CGRL bound $B_{CGRL}^{(*)}(\mathcal{F}_{c_i})$, the maximal ITR bound $B_{ITR}^{(*)}(\mathcal{F}_{c_i})$, and the maximal Lagrangian bound $B_{LD}^{(*)}(\mathcal{F}_{c_i})$ as

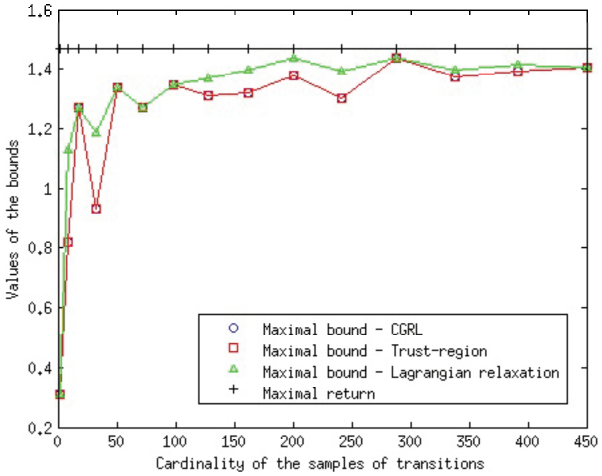

FIG. 6.1. *Bounds* $B^{(*)}_{CGRL}(\mathcal{F}_{c_i})$, $B^{(*)}_{ITR}(\mathcal{F}_{c_i})$, *and* $B^{(*)}_{LD}(\mathcal{F}_{c_i})$ *computed from all samples of transitions* $\mathcal{F}_{c_i}, i \in \{1, \ldots, 15\}$ *of cardinality* $c_i = 2i^2$.

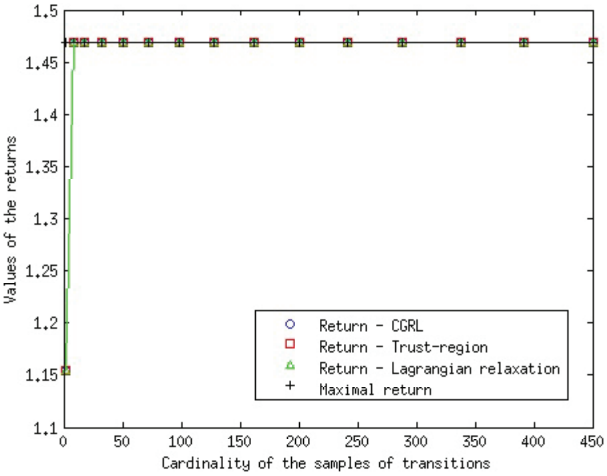

FIG. 6.2. *Returns of the sequences* $(u_0, u_1)^{CGRL}_{\mathcal{F}_{c_i}}$, $(u_0, u_1)^{ITR}_{\mathcal{F}_{c_i}}$, *and* $(u_0, u_1)^{LD}_{\mathcal{F}_{c_i}}$ *computed from all samples of transitions* $\mathcal{F}_{c_i}, i \in \{1, \ldots, 15\}$ *of cardinality* $c_i = 2i^2$.

a function of the cardinality c_i of the samples of transitions $\mathcal{F}_{c_i}$. We also report in Figure 6.2 the returns $J((u_0, u_1)^{CGRL}_{\mathcal{F}_{c_i}})$, $J((u_0, u_1)^{ITR}_{\mathcal{F}_{c_i}})$, and $J((u_0, u_1)^{LD}_{\mathcal{F}_{c_i}})$ of the bound-optimal sequences of actions $(u_0, u_1)^{CGRL}_{\mathcal{F}_{c_i}}$, $(u_0, u_1)^{ITR}_{\mathcal{F}_{c_i}}$, and $(u_0, u_1)^{LD}_{\mathcal{F}_{c_i}}$.

As expected, we observe that the bound computed with the Lagrangian relaxation is always greater than or equal to the ITR bound, which is also greater than or equal to the CGRL bound as predicted by Theorem 5.18. On the other hand, no differences were observed in terms of return of the bound-optimal sequences of actions.

6.2.2. Uniformly drawn samples of transitions. In order to observe the influence of the dispersion of the state-action points of the transitions on the quality of the bounds, we propose the following protocol. For each cardinality $c_i =$

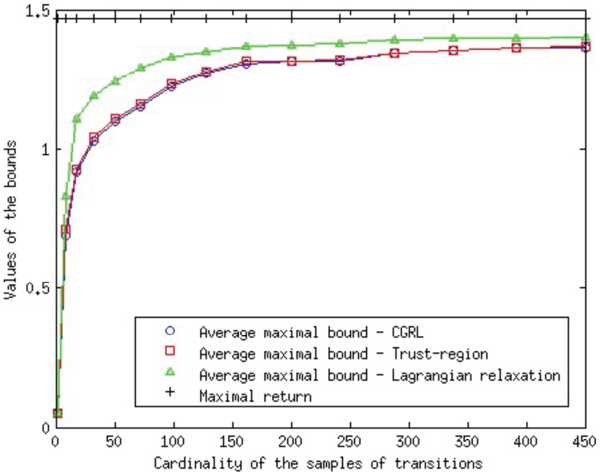

FIG. 6.3. *Average values $A_{CGRL}(c_i)$, $A_{ITR}(c_i)$, and $A_{LD}(c_i)$ of the bounds computed from all samples of transitions $\mathcal{F}_{c_i,k}, k \in \{1,\ldots,100\}$ of cardinality $c_i = 2i^2$.*

$2i^2, i = 1,\ldots,15$, we generate 100 samples of transitions $\mathcal{F}_{c_i,1},\ldots,\mathcal{F}_{c_i,100}$ using a uniform probability distribution over the space $[0,1]^d \times \mathcal{U}$. For each sample of transition $\mathcal{F}_{c_i,k}, i \in \{1,\ldots,15\}, k \in \{1,\ldots,100\}$, we compute the maximal CGRL bound $B^{(*)}_{CGRL}(\mathcal{F}_{c_i,k})$, the maximal ITR bound $B^{(*)}_{ITR}(\mathcal{F}_{c_i,k})$, and the maximal Lagrangian relaxation bound $B^{(*)}_{LD}(\mathcal{F}_{c_i,k})$. We then compute the average values of the maximal CGRL, ITR, and Lagrangian bounds:

$$\forall i \in \{1,\ldots,15\}, \qquad A_{CGRL}(c_i) = \frac{1}{100}\sum_{k=1}^{100} B^{(*)}_{CGRL}\left(\mathcal{F}_{c_i,k}\right),$$
$$A_{ITR}(c_i) = \frac{1}{100}\sum_{k=1}^{100} B^{(*)}_{ITR}\left(\mathcal{F}_{c_i,k}\right),$$
$$A_{LD}(c_i) = \frac{1}{100}\sum_{k=1}^{100} B^{(*)}_{LD}\left(\mathcal{F}_{c_i,k}\right),$$

and we report in Figure 6.3 the values $A_{CGRL}(c_i)$ (resp., $A_{ITR}(c_i)$ and $A_{LD}(c_i)$) as a function of the cardinality c_i of the samples of transitions. We also report in Figure 6.4 the average returns of the bound-optimal sequences of actions $(u_0,u_1)^{CGRL}_{\mathcal{F}_{c_i,k}}$, $(u_0,u_1)^{ITR}_{\mathcal{F}_{c_i,k}}$, and $(u_0,u_1)^{LD}_{\mathcal{F}_{c_i,k}}$:

$$\forall i \in \{1,\ldots,15\}, \qquad J_{CGRL}(c_i) = \frac{1}{100}\sum_{k=1}^{100} J\left((u_0,u_1)^{CGRL}_{\mathcal{F}_{c_i,k}}\right),$$
$$J_{ITR}(c_i) = \frac{1}{100}\sum_{k=1}^{100} J\left((u_0,u_1)^{ITR}_{\mathcal{F}_{c_i,k}}\right),$$
$$J_{LD}(c_i) = \frac{1}{100}\sum_{k=1}^{100} J\left((u_0,u_1)^{LD}_{\mathcal{F}_{c_i,k}}\right)$$

as a function of the cardinality c_i of the samples of transitions.

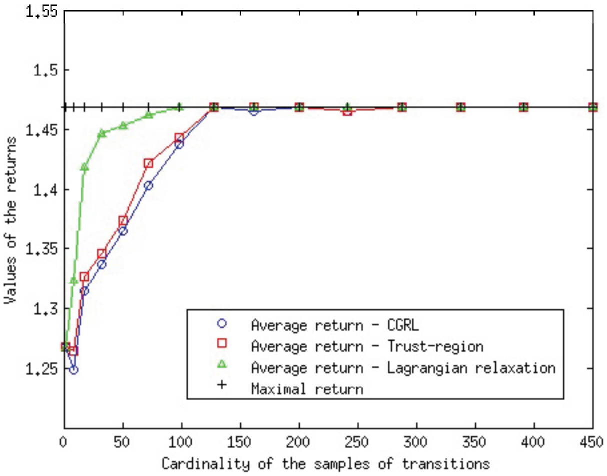

FIG. 6.4. *Average values J_{CGRL}, J_{ITR}, and J_{LD} of the return of the bound-optimal sequences of actions computed from all samples of transitions $\mathcal{F}_{c_i,k}, k \in \{1,\ldots,100\}$ of cardinality $c_i = 2i^2$.*

We observe that, on average, the Lagrangian relaxation bound is much tighter than the ITR and the CGRL bounds. The CGRL bound and the ITR bound remain very close on average, which illustrates, in a sense, Corollary 5.10. Moreover, we also observe that the bound-optimal sequences of actions $(u_0, u_1)^{LD}_{\mathcal{F}_{c_i,k}}$ perform better on average.

7. Conclusions. We have considered in this paper the problem of computing min max policies for deterministic, Lipschitz continuous BMRL. 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 BMRL 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. The authors thank Yurii Nesterov, Adrien Hoarau, and Benoit Daene for fruitful discussions. The scientific responsibility rests with its authors.

REFERENCES

[1] T. Başar and P. Bernhard, *H_∞-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, Vol. 5, Birkhäuser, Boston, 1995.

[2] A. Bemporad and M. Morari, *Robust model predictive control: A survey*, in Robustness in Identification and Control, Lecture Notes in Control and Inform. Sci. 245, 1999, pp. 207–226.

[3] D.P. Bertsekas and J.N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.

[4] J.R. Birge and F. Louveaux, *Introduction to Stochastic Programming*, Springer Verlag, New York, 1997.

[5] S. Boyd, L. El-Ghaoui, E. Feron, V. Balakrishnan, and E.E. Yaz, *Linear matrix inequalities in system and control theory*, Proc. IEEE, 85 (1997), pp. 698–699.

[6] S.J. Bradtke and A.G. Barto, *Linear least-squares algorithms for temporal difference learning*, Mach. Learn., 22 (1996), pp. 33–57.

[7] L. Busoniu, R. Babuska, B. De Schutter, and D. Ernst, *Reinforcement Learning and Dynamic Programming using Function Approximators*, CRC Press, Boca Raton, FL, 2010.

[8] E.F. Camacho and C. Bordons, *Model Predictive Control*, Springer, London, 2004.

[9] A.R. Conn, N.I.M. Gould, and P.L. Toint, *Trust-Region Methods*, MPS-SIAM Ser. Optim. 1, SIAM, Philadelphia, 2000.

[10] A. d'Aspremont and S. Boyd, *Relaxations and Randomized Methods for Nonconvex QCAPs*, EE392o Class Notes, Stanford University, Stanford, CA, 2003.

[11] B. Defourny, D. Ernst, and L. Wehenkel, *Risk-aware decision making and dynamic programming*, NIPS-08 Workshop on Model Uncertainty and Risk in Reinforcement Learning, Whistler, Canada, 2008.

[12] E. Delage and S. Mannor, *Percentile optimization for Markov decision processes with parameter uncertainty*, Oper. Res., 58 (2010), pp. 203–213.

[13] D. Ernst, P. Geurts, and L. Wehenkel, *Tree-based batch mode reinforcement learning*, J. Mach. Learn. Res., 6 (2005), pp. 503–556.

[14] D. Ernst, M. Glavic, F. Capitanescu, and L. Wehenkel, *Reinforcement learning versus model predictive control: A comparison on a power system problem*, IEEE Trans. Syst., Man, Cybernet. Part B, 39 (2009), pp. 517–529.

[15] R. Fonteneau, D. Ernst, B. Boigelot, and Q. Louveaux, *Min Max Generalization for Two-Stage Deterministic Batch Mode Reinforcement Learning: Relaxation Schemes*, Technical report, University of Liège, Liège, Belgium, 2012.

[16] R. Fonteneau, S.A. 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 (ADPRL 09), Nashville, TN, IEEE, Piscataway, NJ, 2009, pp. 117–123.

[17] R. Fonteneau, S.A. Murphy, L. Wehenkel, and D. Ernst, *A cautious approach to generalization in reinforcement learning*, in Agents and Artificial Intelligence: International Conference, ICAART 2010, Valencia, Spain, 2010, Commun. Comp. Inform. Sci. 129, Springer, Heidelberg, 2011, pp. 61–77.

[18] R. Fonteneau, S.A. Murphy, L. Wehenkel, and D. Ernst, *Computing Bounds for Kernel-Based Policy Evaluation in Reinforcement Learning*, Technical Report, University of Liège, Liège, Belgium, 2010.

[19] 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, 2010, Commun. Comp. Inform. Sci. 129, Springer, Heidelberg, 2011, pp. 61–77.

[20] R. Fonteneau, *Contributions to Batch Mode Reinforcement Learning*, Ph.D. thesis, University of Liège, Liège, Belgium, 2011.

[21] R.M. Freund and J.B. Orlin, *On the complexity of four polyhedral set containment problems*, Math. Program., 33 (1985), pp. 139–145.

[22] L.P. Hansen and T.J. Sargent, *Robust control and model uncertainty*, Amer. Econom. Rev., 91 (2001), pp. 60–66.

[23] D. Henrion, S. Tarbouriech, and D. Arzelier, *LMI approximations for the radius of the intersection of ellipsoids: Survey*, J. Optim. Theory Appl., 108 (2001), pp. 1–28.

[24] J.B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms: Fundamentals*, Grundlehren Math. Wiss. 305, Springer-Verlag, Berlin, 1996.

[25] J.E. Ingersoll, *Theory of Financial Decision Making*, Rowman and Littlefield, Totowa, NJ, 1987.

[26] S. Koenig, *Minimax real-time heuristic search*, Artif. Intell., 129 (2001), pp. 165–197.

[27] M.G. Lagoudakis and R. Parr, *Least-squares policy iteration*, J. Mach. Learn. Res., 4 (2003), pp. 1107–1149.

[28] M.L. Littman, *Markov games as a framework for multi-agent reinforcement learning*, in Proceedings of the 11th International Conference on Machine Learning (ICML 1994), New Brunswick, NJ, 1994, Morgan Kaufman, San Francisco, pp. 157–163.

[29] M.L. Littman, *A tutorial on partially observable Markov decision processes*, J. Math. Psychol., 53 (2009), pp. 119–125.

[30] S. Mannor, D. Simester, P. Sun, and J.N. Tsitsiklis, *Bias and variance in value function estimation*, in Proceedings of the 21st International Conference on Machine Learning (ICML 2004), Banff, Alberta, Canada, 2004, AAAI Press, Menlo Park, CA, 2004, 72.

[31] S.A. Murphy, *Optimal dynamic treatment regimes*, J. Roy. Statist. Soc. Ser. B, 65 (2003), pp. 331–366.

[32] S.A. Murphy, *An experimental design for the development of adaptive treatment strategies*, Statist. Med., 24 (2005), pp. 1455–1481.

[33] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim., 19 (2009), pp. 1574–1609.

[34] Y. Nesterov, H. Wolkowicz, and Y. Ye, *Semidefinite programming relaxations of nonconvex quadratic optimization*, in Handbook of Semidefinite Programming, Kluwer, Boston, 2000, pp. 361–419.

[35] D. Ormoneit and S. Sen, *Kernel-based reinforcement learning*, Mach. Learn., 49 (2002), pp. 161–178.

[36] 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), Dagstuhl, Germany, 2011.

[37] C.H. Papadimitriou, *On the complexity of integer programming*, J. ACM, 28 (1981), pp. 765–768.

[38] C.H. Papadimitriou, *Computational Complexity*, Addison-Wesley, Reading, MA, 2003.

[39] P.M. Pardalos and S.A. Vavasis, *Quadratic programming with one negative eigenvalue is NP-hard*, J. Global Optim., 1 (1991), pp. 15–22.

[40] M. Qian and S.A. Murphy, *Performance Guarantees for Individualized Treatment Rules*, Technical report 498, Department of Statistics, University of Michigan, Ann Arbor, MI, 2009.

[41] M. Riedmiller, *Neural fitted Q iteration - first experiences with a data efficient neural reinforcement learning method*, in Proceedings of the 16th European Conference on Machine Learning (ECML 2005), Porto, Portugal, Springer, Berlin, 2005, pp. 317–328.

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

[43] P. Scokaert and D. Mayne, *Min-max feedback model predictive control for constrained linear systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 1136–1142.

[44] A. Shapiro, *A dynamic programming approach to adjustable robust optimization*, Oper. Res. Lett., 39 (2011), pp. 83–87.

[45] A. Shapiro, *Minimax and Risk Averse Multistage Stochastic Programming*, Technical report, School of Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 2011.

[46] J.F. Sturm, *Using SeDuMi* 1.02*, a MATLAB toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11 (1999), pp. 625–653.

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

[48] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.