MIN MAX GENERALIZATION FOR TWO-STAGE 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 [22] for computing policies for batch mode reinforcement learning in a deterministic setting. First, we show that this problem is NP-hard. In the two-stage case, we provide two relaxation schemes. The first relaxation scheme works by dropping some constraints in order to obtain a problem that is solvable in polynomial time. The second relaxation scheme, based on a Lagrangian relaxation where all constraints are dualized, leads to a conic quadratic programming problem. We also theoretically prove and empirically illustrate that both relaxation schemes provide better results than those given in [22].

Key words. Reinforcement Learning, Min Max Generalization, Non-convex Optimization, Computational Complexity

AMS subject classifications. 60J05 Discrete-time Markov processes on general state spaces

1. Introduction. Research in Reinforcement Learning (RL) [48] aims at designing computational agents able to learn by themselves how to interact with their environment to maximize a numerical reward signal. The techniques developed in this field have appealed researchers trying to solve sequential decision making problems in many fields such as Finance [26], Medicine [34, 35] or Engineering [42]. Since the end of the nineties, several researchers have focused on the resolution of a subproblem of RL: computing a high-performance policy when the only information available on the environment is contained in a batch collection of trajectories of the agent [10, 17, 28, 38, 42, 19]. This subfield of RL is known as "batch mode RL".

Batch mode RL (BMRL) algorithms are challenged when dealing with large or continuous state spaces. Indeed, in such cases they have to generalize the information contained in a generally sparse sample of trajectories. The dominant approach for generalizing this information is to combine BMRL algorithms with function approximators [6, 28, 17, 11]. Usually, these approximators generalize the information contained in the sample to areas poorly covered by the sample by implicitly assuming that the properties of the system in those areas are similar to the properties of the system in the nearby areas well covered by the sample. This in turn often leads to low performance guarantees on the inferred policy when large state space areas are poorly covered by the sample. This can be explained by the fact that when computing the performance guarantees of these policies, one needs to take into account that they may actually drive the system into the poorly visited areas to which the generalization strategy associates a favorable environment behavior, while the environment may actually be particularly adversarial in those areas. This is corroborated by theoretical results which show that the performance guarantees of the policies inferred by these algorithms degrade with the sample dispersion where, loosely speaking, the dispersion can be seen as the radius of the largest non-visited state space area.

To overcome this problem, [22] propose 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

[†]Department of Electrical Engineering and Computer Science, University of Liège, Belgium

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 [20, 21] and has a tightness that depends on the sample dispersion. However, in some configurations where areas of the 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 [22]. We first show that this optimization problem is NP-hard. We then focus on the two-stage case, which is still 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 targetting 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. This results into a well known configuration called the *trust-region subproblem* [13]. The second relaxation scheme, based on a Lagrangian relaxation where all constraints are dualized, can be solved using conic quadratic programming in polynomial time. We prove that both relaxation schemes always provide bounds that are greater or equal to the CGRL bound. We also show 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 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 (Theorem 4.2):
  - the first stage problem leads to a trivial optimization problem that can be solved in closed-form (Corollary 4.3),
  - we prove in Section 4.2 that the second stage problem is NP-hard (Corollary 4.7), which consequently proves the NP-hardness of the general $\min\max$ generalization problem (Theorem 4.8),
- we then describe in Section 5 the two relaxation schemes that we propose for the second stage problem:
  - the trust-region relaxation scheme (Section 5.1),
  - the Lagrangian relaxation scheme (Section 5.2), which is shown to be a conic-quadratic problem (Theorem 5.4),
- we prove in Section 5.3.1 that the first relaxation scheme gives better results than CGRL (Theorem 5.9),
- we show in Section 5.3.2 that the second relaxation scheme povides better results than the first relaxation scheme (Theorem 5.13), and consequently better results than CGRL (Theorem 5.14),
- we analyze in Section 5.4 the asymptotic behavior of the relaxation schemes as a function of the sample dispersion:
  - we show that the the bounds provided by the relaxtion schemes converge to-

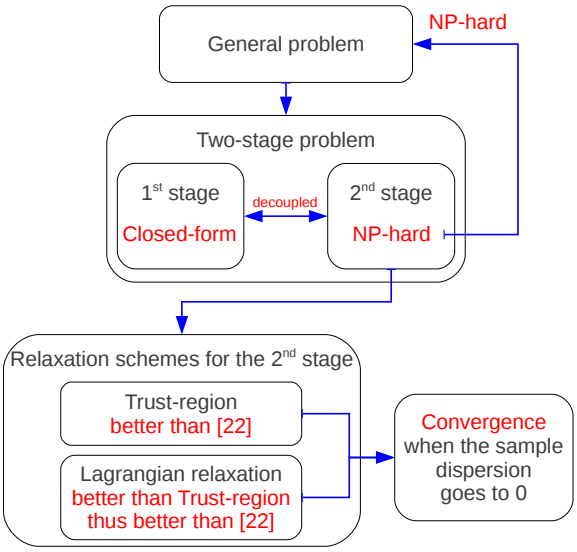

FIG. 1.1. *Main results of the paper.*

wards the actual return when the sample dispersion decreases towards zero (Theorem 5.17),
  - we show that the sequences of actions maximizing such bounds converge towards optimal sequences of actions when the sample dispersion decreases towards zero (Theorem 5.20),
- Section 6 illustrates the relaxation schemes on an academic benchmark,
- Section 7 concludes.

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

2. Related Work. Several works have already been built upon $\min\max$ paradigms for computing policies in a RL setting. In stochastic frameworks, $\min\max$ approaches are often successful for deriving robust solutions with respect to uncertainties in the (parametric) representation of the probability distributions associated with the environment [16]. 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. [29, 43]. They have also received some attention for solving partially observable Markov decision processes [30, 27].

The $\min\max$ approach towards generalization, originally introduced in [22], 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 [33, 41, 39].

Other fields of research have proposed $\min\max$-type strategies for computing control policies. This includes Robust Control theory [24] with H_∞ methods [2], but also Model Predictive Control (MPC) theory - where usually the environment is supposed to be fully known [12, 18] - for which $\min\max$ approaches have been used to determine an optimal sequence of actions with respect to the "worst case" disturbance sequence occurring [44, 4]. Finally, there is a broad stream of works in the field of Stochastic Programming [7] that

have addressed the problem of safely planning under uncertainties, mainly known as "robust stochastic programming" or "risk-averse stochastic programming" [15, 45, 46, 36]. In this field, the two-stage case has also been particularly well-studied [23, 14].

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

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

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

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

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

DEFINITION 3.1 ($T-$stage Return).

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

where

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

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).

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

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

1. The system dynamics f and the reward function ρ are *unknown*;
2. For each action $u \in \mathcal{U}$, a set of $n^{(u)} \in \mathbb{N}$ one-step system transitions

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

is known where each one-step transition is such that:

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

3. We assume that every set $\mathcal{F}^{(u)}$ contains at least one element:

$$\forall u \in \mathcal{U}, \qquad n^{(u)} > 0 \ .$$

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

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

Under those assumptions, batch mode reinforcement learning (BMRL) techniques propose to infer from the sample of one-step system transitions $\mathcal{F}$ a high-performance sequence of actions, i.e. a sequence of actions $\left(\tilde{u}_0^*, \ldots, \tilde{u}_{T-1}^*\right) \in \mathcal{U}^T$ such that $J_T^{(\tilde{u}_0^*, \ldots, \tilde{u}_{T-1}^*)}$ 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 [22].

We first assume that the system dynamics f and the reward function ρ are assumed to be Lipschitz continuous. There exist finite constants $L_f, L_\rho \in \mathbb{R}$ such that:

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

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

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 [22], 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 in determining a most pessimistic trajectory of the system that is still compliant with the sample of data and the Lipschitz continuity assumptions. More specifically, for a given sequence of actions $(u_0, \ldots, u_{T-1}) \in \mathcal{U}^T$, some given constants L_f and L_ρ, a given initial state $x_0 \in \mathcal{X}$ and a given sample of transitions $\mathcal{F}$, this optimization problem writes:

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

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

subject to

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

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

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

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

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

Note that, throughout the paper, optimization variables will be written in bold.

The min max approach to generalization aims at identifying which sequence of actions maximizes its worst possible return, that is which sequence of actions leads to the highest value of $(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1}))$.

We focus in this paper on the design of resolution schemes for solving the program $(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1}))$. These schemes can afterwards be used for solving the $\min\max$ problem through exhaustive search over the set of all sequences of actions.

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

4. The Two-stage Case. In this section, we restrict ourselves to the case where the time horizon contains only two steps, i.e. $T = 2$, which is an important particular case of $\left(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1})\right)$. Many works in optimal sequential decision making have considered the two-stage case [23, 14], which relates to many applications, such as for instance medical applications where one wants to infer "safe" clinical decision rules from batch collections of clinical data [1, 31, 32, 49].

In Section 4.1, we show that this problem can be decoupled into two subproblems. While the first subproblem is straightforward to solve, we prove in Section 4.2 that the second one is NP-hard, which proves that the two-stage problem as well as the generalized $T-$stage problem $(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1}))$ are also NP-hard.

Given a two-stage sequence of actions $(u_0, u_1) \in \mathcal{U}^2$, the two-stage version of the problem $\left(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1})\right)$ writes as follows:

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

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

subject to

$$\left|\hat{\mathbf{r}}_\mathbf{0} - r^{(u_0),k_0}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_\mathbf{0} - x^{(u_0),k_0}\right\|^2 , \forall k_0 \in \left\{1, \ldots, n^{(u_0)}\right\} , \tag{4.1}$$

$$\left|\hat{\mathbf{r}}_\mathbf{1} - r^{(u_1),k_1}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_\mathbf{1} - x^{(u_1),k_1}\right\|^2 , \forall k_1 \in \left\{1, \ldots, n^{(u_1)}\right\} , \tag{4.2}$$

$$\left\|\hat{\mathbf{x}}_\mathbf{1} - y^{(u_0),k_0}\right\|^2 \leq L_f^2 \left\|\hat{\mathbf{x}}_\mathbf{0} - x^{(u_0),k_0}\right\|^2 , \forall k_0 \in \left\{1, \ldots, n^{(u_0)}\right\} , \tag{4.3}$$

$$|\hat{\mathbf{r}}_\mathbf{0} - \hat{\mathbf{r}}_\mathbf{1}|^2 \leq L_\rho^2 \|\hat{\mathbf{x}}_\mathbf{0} - \hat{\mathbf{x}}_\mathbf{1}\|^2 \text{ if } u_0 = u_1 , \tag{4.4}$$

$$\hat{\mathbf{x}}_\mathbf{0} = x_0 . \tag{4.5}$$

For a matter of simplicity, we will often drop the arguments in the definition of the optimization problem and refer $\left(\mathcal{P}_2(\mathcal{F}, L_f, L_\rho, x_0, u_0, u_1)\right)$ as $(\mathcal{P}_2^{(u_0,u_1)})$. We denote by $B_2^{(u_0,u_1)}(\mathcal{F})$ the lower bound associated with an optimal solution of $(\mathcal{P}_2^{(u_0,u_1)})$:

DEFINITION 4.1 (Optimal Value $B_2^{(u_0,u_1)}(\mathcal{F})$). *Let $(u_0, u_1) \in \mathcal{U}^2$, and let $(\hat{\mathbf{r}}_\mathbf{0}^*, \hat{\mathbf{r}}_\mathbf{1}^*, \hat{\mathbf{x}}_\mathbf{0}^*, \hat{\mathbf{x}}_\mathbf{1}^*)$ be an optimal solution to $\left(\mathcal{P}_2^{(u_0,u_1)}\right)$. Then,*

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

4.1. Decoupling Stages. Let $(\mathcal{P}_2'^{(u_0,u_1)})$ and $(\mathcal{P}_2''^{(u_0,u_1)})$ be the two following subproblems:

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

$$\begin{array}{ll} \min & \hat{\mathbf{r}}_\mathbf{0} \\ \hat{\mathbf{r}}_\mathbf{0} \in \mathbb{R} & \\ \hat{\mathbf{x}}_\mathbf{0} \in \mathcal{X} & \end{array}$$

subject to

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

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

$$\begin{array}{ll} \min & \hat{\mathbf{r}}_\mathbf{1} \\ \hat{\mathbf{r}}_\mathbf{1} \in \mathbb{R} & \\ \hat{\mathbf{x}}_\mathbf{1} \in \mathcal{X} & \end{array} \tag{4.6}$$

subject to

$$\left|\hat{\mathbf{r}}_\mathbf{1} - r^{(u_1),k_1}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_\mathbf{1} - x^{(u_1),k_1}\right\|^2 , \forall k_1 \in \left\{1,\ldots,n^{(u_1)}\right\} , \tag{4.7}$$
$$\left\|\hat{\mathbf{x}}_\mathbf{1} - y^{(u_0),k_0}\right\|^2 \leq L_f^2 \left\|x_0 - x^{(u_0),k_0}\right\|^2 , \forall k_0 \in \left\{1,\ldots,n^{(u_0)}\right\} . \tag{4.8}$$

We show in this section that an optimal solution to $(\mathcal{P}_2^{(u_0,u_1)})$ can be obtained by solving the two subproblems $(\mathcal{P}_2'^{(u_0,u_1)})$ and $(\mathcal{P}_2''^{(u_0,u_1)})$ corresponding to the first stage and the second stage. Indeed, one can see that the stages $t = 0$ and $t = 1$ are theoretically coupled by constraint (4.4), except in the case where the two actions u_0 and u_1 are different for which $(\mathcal{P}_2^{(u_0,u_1)})$ is trivially decoupled. We prove in the following that, even in the case $u_0 = u_1$, optimal solutions to the two decoupled problems $(\mathcal{P}_2'^{(u_0,u_1)})$ and $(\mathcal{P}_2''^{(u_0,u_1)})$ also satisfy constraint (4.4). Additionally, we provide the solution of $(\mathcal{P}_2'^{(u_0,u_1)})$.

THEOREM 4.2. *Let* $(u_0, u_1) \in \mathcal{U}^2$. *If* $(\hat{\mathbf{r}}_\mathbf{0}^*, \hat{\mathbf{x}}_\mathbf{0}^*)$ *is an optimal solution to* $\left(\mathcal{P}_2'^{(u_0,u_1)}\right)$ *and* $(\hat{\mathbf{r}}_\mathbf{1}^*, \hat{\mathbf{x}}_\mathbf{1}^*)$ *is an optimal solution to* $\left(\mathcal{P}_2''^{(u_0,u_1)}\right)$, *then* $(\hat{\mathbf{r}}_\mathbf{0}^*, \hat{\mathbf{r}}_\mathbf{1}^*, \hat{\mathbf{x}}_\mathbf{0}^*, \hat{\mathbf{x}}_\mathbf{1}^*)$ *is an optimal solution to* $\left(\mathcal{P}_2^{(u_0,u_1)}\right)$.

Proof.

- First case: $u_0 \neq u_1$.

The constraint (4.4) drops and the theorem is trivial.

- Second case: $u_0 = u_1$.

The rationale of the proof is the following. We first relax constraint (4.4), and consider the two problems $(\mathcal{P}_2'^{(u_0,u_1)})$ and $(\mathcal{P}_2''^{(u_0,u_1)})$. Then, we show that optimal solutions of $(\mathcal{P}_2'^{(u_0,u_1)})$ and $(\mathcal{P}_2''^{(u_0,u_1)})$ also satisfy constraint (4.4).

About $(\mathcal{P}_2'^{(u_0,u_1)})$. The problem $(\mathcal{P}_2'^{(u_0,u_1)})$ consists in the minimization of $\hat{\mathbf{r}}_\mathbf{0}$ under the intersection of interval constraints. It is therefore straightforward to solve. In particular the optimal solution $\hat{\mathbf{r}}_\mathbf{0}^*$ lies at the lower value of one of the intervals. Therefore there exists

$\left(x^{(u_0),k_0^*}, r^{(u_0),k_0^*}, y^{(u_0),k_0^*}\right) \in \mathcal{F}^{(u_0)}$ such that

$$\hat{\mathbf{r}}_{\mathbf{0}}^* = r^{(u_0),k_0^*} - L_\rho \left\| x_0 - x^{(u_0),k_0^*} \right\|. \tag{4.9}$$

Furthermore $\hat{\mathbf{r}}_{\mathbf{0}}^*$ must belong to all intervals. We therefore have that

$$\hat{\mathbf{r}}_{\mathbf{0}}^* \geq r^{(u_0),k_0} - L_\rho \left\| x_0 - x^{(u_0),k_0} \right\| , \qquad \forall k_0 \in \left\{1, \ldots, n^{(u_0)}\right\}. \tag{4.10}$$

In other words,

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

About $(\mathcal{P}_2''^{(u_0,u_1)})$. Again we observe that it is the minimization of $\hat{\mathbf{r}}_{\mathbf{1}}$ under the intersection of interval constraints as well. The sizes of the intervals are however not fixed but determined by the variable $\hat{\mathbf{x}}_{\mathbf{1}}$. If we denote the optimal solution of $(\mathcal{P}_2''^{(u_0,u_1)})$ by $\hat{\mathbf{r}}_{\mathbf{1}}^*$ and $\hat{\mathbf{x}}_{\mathbf{1}}^*$, we know that $\hat{\mathbf{r}}_{\mathbf{1}}^*$ also lies at the lower value of one of the intervals. Hence there exists $\left(x^{(u),k_1^*}, r^{(u),k_1^*}, y^{(u),k_1^*}\right) \in \mathcal{F}^{(u)}$ such that

$$\hat{\mathbf{r}}_{\mathbf{1}}^* = r^{(u),k_1^*} - L_\rho \left\| \hat{\mathbf{x}}_{\mathbf{1}}^* - x^{(u),k_1^*} \right\|. \tag{4.11}$$

Furthermore $\hat{\mathbf{r}}_{\mathbf{1}}^*$ must belong to all intervals. We therefore have that

$$\hat{\mathbf{r}}_{\mathbf{1}}^* \geq r^{(u),k_1} - L_\rho \left\| \hat{\mathbf{x}}_{\mathbf{1}}^* - x^{(u),k_1} \right\| , \qquad \forall k_1 \in \left\{1, \ldots, n^{(u)}\right\}. \tag{4.12}$$

We now discuss two cases depending on the sign of $\hat{\mathbf{r}}_{\mathbf{0}}^* - \hat{\mathbf{r}}_{\mathbf{1}}^*$.

– **If** $\hat{\mathbf{r}}_{\mathbf{0}}^* - \hat{\mathbf{r}}_{\mathbf{1}}^* \geq 0$
Using (4.9) and (4.12) with index k_0^*, we have

$$\hat{\mathbf{r}}_{\mathbf{0}}^* - \hat{\mathbf{r}}_{\mathbf{1}}^* \leq L_\rho \left(\left\| \hat{\mathbf{x}}_{\mathbf{1}}^* - x^{(u),k_0^*} \right\| - \left\| x_0 - x^{(u),k_0^*} \right\| \right) \tag{4.13}$$

Since $\hat{\mathbf{r}}_{\mathbf{0}}^* - \hat{\mathbf{r}}_{\mathbf{1}}^* \geq 0$, we therefore have

$$|\hat{\mathbf{r}}_{\mathbf{0}}^* - \hat{\mathbf{r}}_{\mathbf{1}}^*| \leq L_\rho \left(\left\| \hat{\mathbf{x}}_{\mathbf{1}}^* - x^{(u),k_0^*} \right\| - \left\| x_0 - x^{(u),k_0^*} \right\| \right) . \tag{4.14}$$

Using the triangle inequality we can write

$$\left\| \hat{\mathbf{x}}_{\mathbf{1}}^* - x^{(u),k_0^*} \right\| \leq \| \hat{\mathbf{x}}_{\mathbf{1}}^* - x_0 \| + \left\| x_0 - x^{(u),k_0^*} \right\| . \tag{4.15}$$

Replacing (4.15) in (4.14) we obtain

$$|\hat{\mathbf{r}}_{\mathbf{1}}^* - \hat{\mathbf{r}}_{\mathbf{0}}^*| \leq L_\rho \| \hat{\mathbf{x}}_{\mathbf{1}}^* - x_0 \|$$

which shows that $\hat{\mathbf{r}}_{\mathbf{0}}^*$ and $\hat{\mathbf{r}}_{\mathbf{1}}^*$ satisfy constraint (4.4).

– **If** $\hat{\mathbf{r}}_{\mathbf{0}}^* - \hat{\mathbf{r}}_{\mathbf{1}}^* < 0$
Using (4.11) and (4.10) with index k_1^*, we have

$$\hat{\mathbf{r}}_{\mathbf{1}}^* - \hat{\mathbf{r}}_{\mathbf{0}}^* \leq L_\rho \left(\left\| x_0 - x^{(u),k_1^*} \right\| - \left\| \hat{\mathbf{x}}_{\mathbf{1}}^* - x^{(u),k_1^*} \right\| \right)$$

and since $\hat{\mathbf{r}}_\mathbf{0}^* - \hat{\mathbf{r}}_\mathbf{1}^* < 0$,

$$|\hat{\mathbf{r}}_\mathbf{1}^* - \hat{\mathbf{r}}_\mathbf{0}^*| \leq L_\rho \left(\left\| x_0 - x^{(u),k_1^*} \right\| - \left\| \hat{\mathbf{x}}_\mathbf{1}^* - x^{(u),k_1^*} \right\| \right). \tag{4.16}$$

Using the triangle inequality we can write

$$\left\| x_0 - x^{(u),k_1^*} \right\| \leq \|x_0 - \hat{\mathbf{x}}_\mathbf{1}^*\| + \left\| \hat{\mathbf{x}}_\mathbf{1}^* - x^{(u),k_1^*} \right\|. \tag{4.17}$$

Replacing (4.17) in (4.16) yields

$$|\hat{\mathbf{r}}_\mathbf{1}^* - \hat{\mathbf{r}}_\mathbf{0}^*| \leq L_\rho \|x_0 - \hat{\mathbf{x}}_\mathbf{1}^*\| \ ,$$

which again shows that $\hat{\mathbf{r}}_\mathbf{0}^*$ and $\hat{\mathbf{r}}_\mathbf{1}^*$ satisfy constraint (4.4).
In both cases $\hat{\mathbf{r}}_\mathbf{0}^* - \hat{\mathbf{r}}_\mathbf{1}^* \geq 0$ and $\hat{\mathbf{r}}_\mathbf{0}^* - \hat{\mathbf{r}}_\mathbf{1}^* < 0$, we have shown that constraint (4.4) is satisfied. □

In the following of the paper, we focus on the two subproblems $(\mathcal{P}_2'^{(u_0,u_1)})$ and $(\mathcal{P}_2''^{(u_0,u_1)})$ rather than on $(\mathcal{P}_2^{(u_0,u_1)})$. From the proof of Theorem 4.2 given above, we can directly obtain the solution of $(\mathcal{P}_2'^{(u_0,u_1)})$:

COROLLARY 4.3. *The solution of the problem* $(\mathcal{P}_2'^{(u_0,u_1)})$ *is*

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

4.2. Complexity of $(\mathcal{P}_2''^{(u_0,u_1)})$. The problem $(\mathcal{P}_2'^{(u_0,u_1)})$ being solved, we now focus in this section on the resolution of $(\mathcal{P}_2''^{(u_0,u_1)})$. In particular, we show that it is NP-hard, even in the particular case where there is only one element in the sample $\mathcal{F}^{(u_1)} = \left\{ \left(x^{(u_1),1}, r^{(u_1),1}, y^{(u_1),1} \right) \right\}$. In this particular case, the problem $(\mathcal{P}_2''^{(u_0,u_1)})$ amounts to maximizing of the distance $\left\| \hat{\mathbf{x}}_\mathbf{1} - x^{(u_1),1} \right\|$ under an intersection of balls as we show in the following lemma.

LEMMA 4.4. *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 $\left(\mathcal{P}_2''^{(u_0,u_1)} \right)$ *satisfies*

$$\hat{\mathbf{r}}_\mathbf{1}^* = r^{(u_1),1} - L_\rho \left\| \hat{\mathbf{x}}_\mathbf{1}^* - x^{(u_1),1} \right\|$$

where $\hat{\mathbf{x}}_\mathbf{1}^*$ *maximizes* $\left\| \hat{\mathbf{x}}_\mathbf{1} - x^{(u_1),1} \right\|$ *subject to*

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

Proof. The unique constraint concerning $\hat{\mathbf{r}}_\mathbf{1}$ is an interval. Therefore $\hat{\mathbf{r}}_\mathbf{1}^*$ takes the value of the lower bound of the interval. In order to obtain the lowest such value, the right-hand-side of (4.7) 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 $(\mathcal{P}_2^{(u_0,u_1)})$ can be solved exactly, as we will later show in Corollary 5.3. But, in the general case where $n^{(u_0)} > 1$, 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 4.5 (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 4.6. *MNBC is NP-hard.*

Proof. To prove it, we will do a reduction from the $\{0,1\}-$programming feasibility problem [40]. 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

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

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

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

that we can write algebraically

$$\sum_{j\neq k}(x_j-1)^2+(x_k-1)^2\geq d \tag{4.18}$$

$$\sum_{j\neq k}(x_j-1)^2+x_k^2\leq d+3 \tag{4.19}$$

$$\sum_{j\neq k}(x_j-1)^2+(x_k-2)^2\leq d+3. \tag{4.20}$$

By computing (4.19) – (4.18) and (4.20) – (4.18), we obtain $x_k \leq 2$ and $x_k \geq 0$ respectively. This implies that

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

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 $\sum_{j=1}^{d} a_j x_j \leq b$. 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 exists $y \in \mathbb{Q}^d$ and $\gamma \in \mathbb{Q}$ such that

$$\left\{x \in \{0,2\}^d \mid a^T x \leq b\right\} = \left\{x \in \{0,2\}^d \mid \|x-y\|^2 \leq \gamma\right\}. \tag{4.21}$$

Let $\bar{y} \in \mathbb{R}^d$ be the intersection point of the hyperplane $a^T x = b+1$ and the line $(1\ \cdots\ 1)^T + \lambda(a_1\ \cdots\ a_d)^T, \lambda \in \mathbb{R}$. Let r be defined as follows:

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

We claim that choosing $\gamma \triangleq r^2$ and $y \triangleq \bar{y} - ra$ allows us to obtain (4.21). 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 $a^T x \leq b$. Let $\bar{x} \in \{0,2\}^d$. There are two cases to consider:

- Suppose first that $a^T \bar{x} \geq b+2$.

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:

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

- Suppose now that $a^T \bar{x} \leq b$ 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 $\left((1\cdots 1)^T, \tilde{y}, \bar{x}\right)$ form a right triangle with the right angle in $\tilde{y}$ and since $\|(1\cdots 1)^T - \bar{x}\|^2 \leq d$, we have

$$\|\tilde{y}-\bar{x}\|^2 \leq d. \tag{4.22}$$

By definition of y, we have:

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

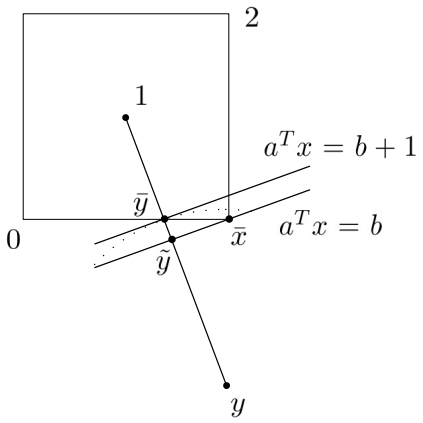

FIG. 4.1. *The case when $a^T \bar{x} \leq b$.*

and by 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

$$\|y - \tilde{y}\| \leq r - \frac{1}{\sqrt{\sum_{j=1}^{d} a_j^2}}. \tag{4.23}$$

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

$$\begin{aligned}
\|\bar{x} - y\|^2 &= \|y - \tilde{y}\|^2 + \|\bar{x} - \tilde{y}\|^2 \\
&\leq \left(r - \frac{1}{\sqrt{\sum_{j=1}^{d} a_j^2}} \right)^2 + d \quad \text{using (4.22), (4.23)} \\
&= r^2 - \frac{2r}{\sqrt{\sum_{j=1}^{d} a_j^2}} + \frac{1}{\sum_{j=1}^{d} a_j^2} + d.
\end{aligned}$$

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 $\left\{ x \in \mathbb{R}^d \mid \|x - y\|^2 \leq r^2 \right\}$ 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 from the choice of the norm used over the state space $\mathcal{X}$. The two results follow:

COROLLARY 4.7. *$\left(\mathcal{P}_2''^{(u_0,u_1)}\right)$ is NP-hard.*

THEOREM 4.8. *The two-stage problem $\left(\mathcal{P}_2^{(u_0,u_1)}\right)$ and the generalized $T-$stage problem $(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1}))$ are NP-hard.*

5. Relaxation Schemes for the Two-stage Case. 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 (except if the cardinality of $n^{(u_0)}$ of $\mathcal{F}^{(u_0)}$ is also equal to 1, in this case $(\mathcal{P}_2^{(u_0,u_1)})$ is solvable in polynomial time as we will see later in Corollary 5.3). It is therefore unlikely that one can design an algorithm that optimally solves the general two-stage case in polynomial time (unless P = NP). The aim of the $\min\max$ optimization problem is to obtain a sequence of actions that has a performance guarantee. Therefore solving the optimization problem approximately or obtaining an upper bound would be irrelevant. Instead we want to propose some relaxation schemes that are computationally more tractable, and that are still leading to lower bounds on the actual return of the sequences of actions.

The first relaxation scheme works by dropping some constraints in order to obtain a problem that is solvable in polynomial time. We show that this scheme provides bounds that are greater or equal to the CGRL bound introduced in [22]. The second relaxation scheme is based on a Lagrangian relaxation where all constraints are dualized. Solving the Lagrangian dual is shown to be a conic-quadratic problem that can be solved 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 [22]. We also prove that the bounds computed from these relaxation schemes converge towards the actual return of the sequence (u_0, u_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 two-stage problem $(\mathcal{P}_2^{(u_0,u_1)})$ can be decoupled into two subproblems $(\mathcal{P}_2'^{(u_0,u_1)})$ and $(\mathcal{P}_2''^{(u_0,u_1)})$, where $(\mathcal{P}_2'^{(u_0,u_1)})$ can be solved straightforwardly (cf Theorem 4.2). We therefore only focus on relaxing the subproblem $(\mathcal{P}_2''^{(u_0,u_1)})$:

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

$$\text{subject to} \quad \left|\hat{\mathbf{r}}_\mathbf{1} - r^{(u_1),k_1}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_\mathbf{1} - x^{(u_1),k_1}\right\|^2 \quad \forall k_1 \in \left\{1, \ldots, n^{(u_1)}\right\} \tag{5.1}$$

$$\left\|\hat{\mathbf{x}}_\mathbf{1} - y^{(u_0),k_0}\right\|^2 \leq L_f^2 \left\|x_0 - x^{(u_0),k_0}\right\|^2 \quad \forall k_0 \in \left\{1, \ldots, n^{(u_0)}\right\} \tag{5.2}$$

5.1. The Trust-region Subproblem Relaxation Scheme. An easy way to obtain a relaxation from an optimization problem is to drop some constraints. We therefore suggest to drop all constraints (5.1) but one, indexed by k_1. Similarly we drop all constraints (5.2) but one, indexed by k_0. The following problem is therefore a relaxation of $(\mathcal{P}_2''^{(u_0,u_1)})$:

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

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

subject to

$$\left|\hat{\mathbf{r}}_\mathbf{1} - r^{(u_1),k_1}\right|^2 \leq L_\rho^2 \left\|\hat{\mathbf{x}}_\mathbf{1} - x^{(u_1),k_1}\right\|^2, \tag{5.3}$$

$$\left\|\hat{\mathbf{x}}_\mathbf{1} - y^{(u_0),k_0}\right\|^2 \leq L_f^2 \left\|x_0 - x^{(u_0),k_0}\right\|^2 . \tag{5.4}$$

We then have the following theorem:

THEOREM 5.1. *Let us denote by $B_{TR}^{\prime\prime(u_0,u_1),k_0,k_1}(\mathcal{F})$ the bound given by the resolution of $(\mathcal{P}_{TR}^{\prime\prime(u_0,u_1)}(k_0,k_1))$. We have:*

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

where

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

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

Proof. Observe that it consists in the minimization of $\hat{\mathbf{r}}_\mathbf{1}$ under one interval constraint for $\hat{\mathbf{r}}_\mathbf{1}$ where the size of the interval is determined through the constraint (5.4). The problem is therefore equivalent to finding the largest right-hand-side of (5.3) under constraint (5.4). An equivalent problem is therefore

$$\begin{array}{ll} \max\limits_{\hat{\mathbf{x}}_\mathbf{1} \in \mathcal{X}} & \left\|\hat{\mathbf{x}}_\mathbf{1} - x^{(u_1),k_1}\right\|^2 \\ \text{subject to} & \left\|\hat{\mathbf{x}}_\mathbf{1} - y^{(u_0),k_0}\right\| \leq L_f \left\|x_0 - x^{(u_0),k_0}\right\|. \end{array}$$

This is the maximization of a quadratic function under a norm constraint. This problem is referred to in the literature as the *trust-region subproblem* [13]. In our case, the optimal value for $\hat{\mathbf{x}}_\mathbf{1}$ - denoted by $\hat{\mathbf{x}}_\mathbf{1}^*(k_0,k_1)$ - lies on the same line as $x^{(u_1),k_1}$ and $y^{(u_0),k_0}$, with $y^{(u_0),k_0}$ lying in between $x^{(u_1),k_1}$ and $\hat{\mathbf{x}}_\mathbf{1}^*(k_0,k_1)$, the distance between $y^{(u_0),k_0}$ and $\hat{\mathbf{x}}_\mathbf{1}^*(k_0,k_1)$ being exactly equal to the distance between x_0 and $x^{(u_0),k_0}$. An illustration is given in Figure 5.1. □

Solving $(\mathcal{P}_{TR}^{\prime\prime(u_0,u_1)}(k_0,k_1))$ provides us with a family of relaxations for our initial problem by considering any combination (k_0,k_1) of two non-relaxed constraints. Taking the maximum out of these lower bounds yields the best possible bound out of this family of relaxations. Finally, if we denote by $B_{TR}^{(u_0,u_1)}(\mathcal{F})$ the bound made of the sum of the solution of $(\mathcal{P}_2^{\prime(u_0,u_1)})$ and the maximal Trust-region relaxation of the problem $(\mathcal{P}_2^{\prime\prime(u_0,u_1)})$ over all possible couples of constraints, we have:

DEFINITION 5.2 (Trust-region Bound $B_{TR}^{(u_0,u_1)}(\mathcal{F})$).

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

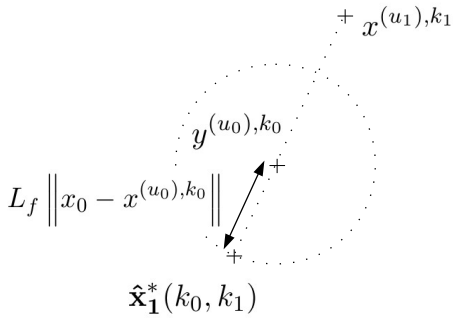

FIG. 5.1. *A simple geometric algorithm to solve* $(\mathcal{P}''_{TR}(k_0, k_1))$.

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

COROLLARY 5.3.

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

5.2. The Lagrangian Relaxation. Another way to obtain a lower bound on the value of a minimization problem is to consider a Lagrangian relaxation. In this section, we show that the Lagrangian relaxation of the second stage problem is a conic quadratic optimization program. Consider again the optimization problem $(\mathcal{P}_2''^{(u_0,u_1)})$. If we multiply the constraints (5.1) by dual variables $\mu_1, \ldots, \mu_{k_1}, \ldots, \mu_{n^{(u_1)}} \geq 0$ and the constraints (5.2) by dual variables $\lambda_1, \ldots, \lambda_{k_0}, \ldots, \lambda_{n^{(u_0)}} \geq 0$, we obtain the Lagrangian dual:

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

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

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

$$+ \sum_{k_0=1}^{n^{(u_0)}} \lambda_{k_0} \left(\left\| \hat{\mathbf{x}}_\mathbf{1} - y^{(u_0),k_0} \right\|^2 - L_f^2 \left\| x_0 - x^{(u_0),k_0} \right\|^2 \right) . \tag{5.5}$$

Observe that the optimal value of $(\mathcal{P}_{LD}^{\prime\prime(u_0,u_1)})$ is known to provide a lower bound on the optimal value of $(\mathcal{P}_2^{\prime\prime(u_0,u_1)})$ [25].

THEOREM 5.4. $\left(\mathcal{P}_{LD}^{\prime\prime(u_0,u_1)}\right)$ *is a conic quadratic program.*

Proof. In (5.5), we can decompose the squared norms and obtain

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

$$\max_{\substack{\lambda_1,\ldots,\lambda_{n^{(u_0)}}\in\mathbb{R}_+\\ \mu_1,\ldots,\mu_{n^{(u_1)}}\in\mathbb{R}_+}} \min_{\substack{\hat{\mathbf{r}}_\mathbf{1}\in\mathbb{R}\\ \hat{\mathbf{x}}_\mathbf{1}\in\mathcal{X}}} \hat{\mathbf{r}}_\mathbf{1}^2\left(\sum_{k_1=1}^{n^{(u_1)}}\mu_{k_1}\right)+\|\hat{\mathbf{x}}_\mathbf{1}\|^2\left(-L_\rho^2\sum_{k_1=1}^{n^{(u_1)}}\mu_{k_1}+\sum_{k_0=1}^{n^{(u_0)}}\lambda_{k_0}\right) \tag{5.6}$$

$$+\hat{\mathbf{r}}_\mathbf{1}\left(1-2\sum_{k_1=1}^{n^{(u_1)}}r^{(u_1),k_1}\right)+\sum_{k_1=1}^{n^{(u_1)}}2L_\rho^2\mu_{k_1}\left\langle\hat{\mathbf{x}}_\mathbf{1},x^{(u_1),k_1}\right\rangle-\sum_{k_0=1}^{n^{(u_0)}}2\lambda_{k_0}\left\langle\hat{\mathbf{x}}_\mathbf{1},y^{(u_0),k_0}\right\rangle \tag{5.7}$$

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

$$+\sum_{k_0=1}^{n^{(u_0)}}\lambda_{k_0}\left(\left\|y^{(u_0),k_0}\right\|^2-L_f^2\left\|x^{(u_0),k_0}-x_0\right\|^2\right), \tag{5.8}$$

where $\langle a,b\rangle$ denotes the inner product of a and b. We observe that the minimization problem in $\hat{\mathbf{r}}_\mathbf{1}$ and $\hat{\mathbf{x}}_\mathbf{1}$ contains a quadratic part (5.6), a linear part (5.7) and a constant part (5.8) once we fix λ_{k_0} and μ_{k_1}. In particular, observe that the optimal solution of the minimization problem is $-\infty$ as soon as the quadratic term is negative, i.e. if :

$$\sum_{k_1=1}^{n^{(u_1)}}\mu_{k_1}\leq 0 \tag{5.9}$$

or

$$\left(-L_\rho^2\sum_{k_1=1}^{n^{(u_1)}}\mu_{k_1}+\sum_{k_0=1}^{n^{(u_0)}}\lambda_{k_0}\right)\leq 0. \tag{5.10}$$

Since we want to find the maximum of this series of optimization problems, we are only interested in the problems for which the solution is finite. Observe that, since $\mu_{k_1}\geq 0$ for all k_1, the inequality (5.9) is never satisfied, unless if $\mu_{k_1}=0$ for all k_1. Therefore in the following, we will constraint λ_{k_0} and μ_{k_1} to be such that inequalities (5.9) and (5.10) are never satisfied, i.e.:

$$\sum_{k_1=1}^{n^{(u_1)}}\mu_{k_1}>0$$

$$-L_\rho^2\sum_{k_1=1}^{n^{(u_1)}}\mu_{k_1}+\sum_{k_0=1}^{n^{(u_0)}}\lambda_{k_0}>0.$$

Once that constraint is enforced, we observe that the minimization program is the minimization of a convex quadratic function for which the optimum can be found as a closed form formula. In order to simplify the rest of the proof, we introduce some useful notations:

DEFINITION 5.5 (Additional Notations).

$$M \triangleq \sum_{k_1=1}^{n^{(u_1)}} \mu_{k_1} \quad , \quad L \triangleq \sum_{k_0=1}^{n^{(u_0)}} \lambda_{k_0} \ ,$$
$$X \triangleq \left(x^{(u_1),1} \dots x^{(u_1),n^{(u_1)}}\right) \quad , \quad Y \triangleq \left(y^{(u_0),1} \dots y^{(u_0),n^{(u_0)}}\right) \ ,$$
$$\boldsymbol{\lambda} \triangleq \begin{pmatrix}\lambda_1 & \dots & \lambda_{n^{(u_0)}}\end{pmatrix}^T \quad , \quad \boldsymbol{\mu} \triangleq \begin{pmatrix}\mu_1 & \dots & \mu_{n^{(u_1)}}\end{pmatrix}^T \quad , \quad \bar{r} \triangleq \begin{pmatrix}r^{(1)} & \dots & r^{(n^{(u_1)})}\end{pmatrix}^T \ ,$$
$\forall p \in \mathbb{N}_0$, *$I_p$ is an identity matrix of size p.*

The quadratic form coming from (5.6), (5.7) and (5.8) can be written in the form

$$z^T Q z + l^T z + c$$

with

$$z \triangleq \begin{pmatrix} \hat{\mathbf{x}}_\mathbf{1} \\ \hat{\mathbf{r}}_\mathbf{1} \end{pmatrix} \in \mathbb{R}^{d+1}, \qquad Q \triangleq \begin{pmatrix} \left(-ML_\rho^2 + L\right) I_d & \\ & M \end{pmatrix}, \qquad l \triangleq \begin{pmatrix} 2L_\rho^2 X\boldsymbol{\mu} - 2Y\boldsymbol{\lambda} \\ 1 - 2\bar{r}^T\boldsymbol{\mu} \end{pmatrix}$$

and the constant term is given by (5.8). The minimum of a convex quadratic form $z^T Q z + l^T z + c$ is known to take the value $-\frac{1}{4} l^T Q^{-1} l + c$. In our case, the inverse of the matrix Q is trivial to compute and we obtain finally that $(\mathcal{P}_{LD}^{\prime\prime(u_0,u_1)})$ can be written as

$$\begin{aligned}\left(\mathcal{P}_{LD}^{\prime\prime(u_0,u_1)}\right) : \max_{\boldsymbol{\lambda}\in\mathbb{R}_+^{n^{(u_0)}}, \boldsymbol{\mu}\in\mathbb{R}_+^{n^{(u_1)}}} & \frac{-\|L_\rho^2 X\boldsymbol{\mu} - Y\boldsymbol{\lambda}\|^2}{-ML_\rho^2 + L} - \frac{\left(1 - 2\bar{r}^T\boldsymbol{\mu}\right)^2}{4M} \\ & + \sum_{k_0=1}^{n^{(u_0)}} \lambda_{k_0} \left(\left\|y^{(u_0),k_0}\right\|^2 - L_f^2 \left\|x^{(u_0),k_0} - x_0\right\|^2\right) \\ & + \sum_{k_1=1}^{n^{(u_1)}} \mu_{k_1} \left(\left(r^{(u_1),k_1}\right)^2 - L_\rho^2 \left\|x^{(u_1),k_1}\right\|^2\right) \\ \text{subject to} \quad & M > 0 \\ & L > ML_\rho^2\end{aligned} \tag{5.11}$$

The optimization problem (5.11) is in variables $\lambda_1, \dots, \lambda_{n^{(u_0)}}$ and $\mu_1, \dots, \mu_{n^{(u_1)}}$. Observe that, with our notation, M and L are linear functions of the variables. The objective function contains linear terms in $\lambda_1, \dots, \lambda_{n^{(u_0)}}$ and $\mu_1, \dots, \mu_{n^{(u_1)}}$ as well as a *fractional-quadratic* function ([5]), i.e. the quotient of a concave quadratic function with a linear function. The constraint is linear. This type of problem is known as a *rotated quadratic conic problem* and can be formulated as a conic quadratic optimization problem ([5]) that can be solved in polynomial time using interior point methods [5, 37, 9]. □

From there, we have the following corollary:

COROLLARY 5.6. $\forall (u_0, u_1) \in \mathcal{U}^2$,

$$B_{LD}^{\prime\prime(u_0,u_1)}(\mathcal{F}) \triangleq \max_{\boldsymbol{\lambda}\in\mathbb{R}_+^{n^{(u_0)}},\boldsymbol{\mu}\in\mathbb{R}_+^{n^{(u_1)}}} \frac{-\|L_\rho^2 X\boldsymbol{\mu} - Y\boldsymbol{\lambda}\|^2}{-ML_\rho^2 + L} - \frac{\left(1 - 2\bar{r}^T\boldsymbol{\mu}\right)^2}{4M} \tag{5.12}$$

$$+ \sum_{k_0=1}^{n^{(u_0)}} \lambda_{k_0} \left(\left\| y^{(u_0),k_0} \right\|^2 - L_f^2 \left\| x^{(u_0),k_0} - x_0 \right\|^2 \right) \tag{5.13}$$

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

$$\textit{subject to} \qquad M > 0$$
$$L > ML_\rho^2$$

In the following, we denote by $B_{LD}^{(u_0,u_1)}(\mathcal{F})$ the lower bound made of the sum of the solution of $(\mathcal{P}_2^{\prime(u_0,u_1)})$ and the relaxation of $(\mathcal{P}_2^{\prime\prime(u_0,u_1)})$ computed from the Lagrangian relaxation:

DEFINITION 5.7 (Lagrangian Relaxation Bound $B_{LD}^{(u_0,u_1)}(\mathcal{F})$).

$$\forall (u_0, u_1) \in \mathcal{U}^2, \qquad B_{LD}^{(u_0,u_1)}(\mathcal{F}) \triangleq \hat{\mathbf{r}}_{\mathbf{0}}^* + B_{LD}^{\prime\prime(u_0,u_1)}(\mathcal{F}) \tag{5.14}$$

5.3. Comparing the Bounds. The CGRL algorithm proposed in [21, 22] for addressing the $\min\max$ problem uses the procedure described in [20] for computing a lower bound on the return of a policy given a sample of trajectories. More specifically, for a given sequence $(u_0, u_1) \in \mathcal{U}^2$, the program $(\mathcal{P}_T(\mathcal{F}, L_f, L_\rho, x_0, u_0, \ldots, u_{T-1}))$ is replaced by a lower bound $B_{CGRL}^{(u_0,u_1)}(\mathcal{F})$. We may now wonder how this bound compares in the two-stage case with the two new bounds of $\left(\mathcal{P}_2^{(u_0,u_1)}\right)$ that we have proposed: the trust-region bound and the Lagrangian relaxation bound.

5.3.1. Trust-region Versus CGRL. We first recall the definition of the CGRL bound in the two-stage case.

DEFINITION 5.8 (CGRL Bound $B_{CGRL}^{(u_0,u_1)}(\mathcal{F})$). $\forall (u_0, u_1) \in \mathcal{U}^2$,

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

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

THEOREM 5.9.

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

Proof. Let $k_0^* \in \left\{1,\ldots,n^{(u_0)}\right\}$ and $k_1^* \in \left\{1,\ldots,n^{(u_1)}\right\}$ be such that

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

Now, let us consider the solution $B_{TR}^{\prime\prime(u_0,u_1),k_0^*,k_1^*}(\mathcal{F})$ of the problem $(\mathcal{P}_{TR}^{\prime\prime(u_0,u_1)}(k_0^*,k_1^*))$, and let us denote by $B^{(u_0,u_1),k_0^*,k_1^*}$ the bound obtained if, in the definition of the value of $\hat{r}_0^*$ given in Corollary 4.3, we fix the value of k_0' to k_0^* instead of maximizing over all possible k_0':

$$B^{(u_0,u_1),k_0^*,k_1^*} = r^{(u_0),k_0^*} - L_\rho\left\|x_0 - x^{(u_0),k_0^*}\right\| + B_{TR}^{\prime\prime(u_0,u_1),k_0^*,k_1^*}(\mathcal{F})$$

Since $r^{(u_0),k_0^*} - L_\rho\left\|x_0 - x^{(u_0),k_0^*}\right\|$ is smaller or equal to the solution $\hat{\mathbf{r}}_{\mathbf{0}}^*$ of $(\mathcal{P}_2^{\prime(u_0,u_1)})$, one has:

$$B_{TR}^{(u_0,u_1),k_0^*,k_1^*}(\mathcal{F}) \geq B^{(u_0,u_1),k_0^*,k_1^*}\,. \tag{5.15}$$

Now, observe that:

$$\begin{aligned} B^{(u_0,u_1),k_0^*,k_1^*} - B_{CGRL}^{(u_0,u_1)}(\mathcal{F}) &= L_\rho L_f\left\|x^{(u_0),k_0^*} - x_0\right\| + L_\rho\left\|y^{(u_0),k_0^*} - x^{(u_1),k_1^*}\right\| \\ &\quad - L_\rho\left\|\hat{\mathbf{x}}_{\mathbf{1}}^*(k_0^*,k_1^*) - x^{(u_1),k_1^*}\right\|\,. \end{aligned} \tag{5.16}$$

By construction, $\hat{\mathbf{x}}_{\mathbf{1}}^*(k_0^*,k_1^*)$ lies on the same line as $y^{(u_0),k_0^*}$ and $x^{(u_1),k_1^*}$ (see Figure 5.1). Furthermore

$$\left\|\hat{\mathbf{x}}_{\mathbf{1}}^*(k_0^*,k_1^*) - x^{(u_1),k_1^*}\right\| = \left\|\hat{\mathbf{x}}_{\mathbf{1}}^*(k_0^*,k_1^*) - y^{(u_0),k_0^*}\right\| + \left\|y^{(u_0),k_0^*} - x^{(u_1),k_1^*}\right\|. \tag{5.17}$$

Using (5.17) in (5.16) yields

$$\begin{aligned} &B^{(u_0,u_1),k_0^*,k_1^*} - B_{CGRL}^{(u_0,u_1),k_0^*,k_1^*}(\mathcal{F}) = L_\rho L_f\left\|x^{(u_0),k_0^*} - x_0\right\| \\ &+L_\rho\left(\left\|y^{(u_0),k_0^*} - x^{(u_1),k_1^*}\right\| - \left\|\hat{\mathbf{x}}_{\mathbf{1}}^*(k_0^*,k_1^*) - y^{(u_0),k_0^*}\right\| - \left\|y^{(u_0),k_0^*} - x^{(u_1),k_1^*}\right\|\right) \\ &= L_\rho\left(L_f\left\|x^{(u_0),k_0^*} - x_0\right\| - \left\|\hat{\mathbf{x}}_{\mathbf{1}}^*(k_0^*,k_1^*) - y^{(u_0),k_0^*}\right\|\right). \end{aligned} \tag{5.18}$$

By construction, Equation (5.18) is equal to 0 (see Figure 5.1), which proves the equality of the two bounds:

$$B^{(u_0,u_1),k_0^*,k_1^*} = B_{CGRL}^{(u_0,u_1)}(\mathcal{F})\,. \tag{5.19}$$

The final result is given by combining Equations (5.15) and (5.19). □

From the proof, one can observe that the gap between the CGRL bound and the Trust-region bound is only due to the resolution of $(\mathcal{P}_2^{\prime(u_0,u_1)})$. Note that in the case where k_0^* also belongs to the set $\underset{k_0\in\{1,\ldots,n^{(u_0)}\}}{\arg\max}\; r^{(u_0),k_0} - L_\rho\left\|x^{(u_0),k_0} - x_0\right\|$, then the bounds are equal. The two corollaries follow:

COROLLARY 5.10. *Let $(u_0,u_1) \in \mathcal{U}^2$. Let $k_0^* \in \left\{1,\ldots,n^{(u_0)}\right\}$ and $k_1^* \in \left\{1,\ldots,n^{(u_1)}\right\}$ be such that:*

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

Then,

$$\left(k_0^* \in \underset{k_0 \in \{1,\ldots,n^{(u_0)}\}}{\arg\max} \; r^{(u_0),k_0} - L_\rho \left\| x^{(u_0),k_0} - x_0 \right\| \right) \implies B_{CGRL}^{(u_0,u_1)}(\mathcal{F}) = B_{TR}^{(u_0,u_1)}(\mathcal{F}) \; .$$

COROLLARY 5.11.

$$\forall (u_0,u_1) \in \mathcal{U}^2, \qquad \left(n^{(u_0)} = 1\right) \implies B_{CGRL}^{(u_0,u_1)}(\mathcal{F}) = B_{TR}^{(u_0,u_1)}(\mathcal{F}) \; .$$

5.3.2. Lagrangian Relaxation Versus Trust-region. In this section, we prove that the lower bound obtained with the Lagrangian relaxation is always greater than or equal to the Trust-region bound. To prove this result, we give a preliminary lemma:

LEMMA 5.12. *Let* $(u_0,u_1) \in \mathcal{U}^2$ *and* $(k_0,k_1) \in \left\{1,\ldots,n^{(u_0)}\right\} \times \left\{1,\ldots,n^{(u_1)}\right\}$. *Consider again the problem* $\left(\mathcal{P}_{TR}^{\prime\prime(u_0,u_1)}(k_0,k_1)\right)$ *where all constraints are dropped except the two defined by* (k_0,k_1)*:*

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

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

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

Then, the Lagrangian relaxation of $\left(\mathcal{P}_{TR}^{\prime\prime(u_0,u_1)}(k_0,k_1)\right)$ *leads to a bound denoted by* $B_{LD}^{\prime\prime(u_0,u_1),k_0,k_1}(\mathcal{F})$ *which is equal to the Trust-region bound* $B_{TR}^{\prime\prime(u_0,u_1),k_0,k_1}(\mathcal{F})$*, i.e.*

$$B_{LD}^{\prime\prime(u_0,u_1),k_0,k_1}(\mathcal{F}) = B_{TR}^{\prime\prime(u_0,u_1),k_0,k_1}(\mathcal{F}) \; .$$

Proofs of this lemma can be found in [3] and [8], but we also provide in Appendix A a proof in our particular case. We then have the following theorem:

THEOREM 5.13.

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

Proof. Let $(u_0,u_1) \in \mathcal{U}^2$. Let $(k_0^*,k_1^*) \in \left\{1,\ldots,n^{(u_0)}\right\} \times \left\{1,\ldots,n^{(u_1)}\right\}$ be such that:

$$B_{TR}^{(u_0,u_1)}(\mathcal{F}) = \hat{\mathbf{r}}_\mathbf{0}^* + B_{TR}^{\prime\prime(u_0,u_1),k_0^*,k_1^*}(\mathcal{F}).$$

Considering $(k_0,k_1) = (k_0^*,k_1^*)$ in Lemma 5.12, we have:

$$B_{TR}^{(u_0,u_1)}(\mathcal{F}) = \hat{\mathbf{r}}_\mathbf{0}^* + B_{LD}^{\prime\prime(u_0,u_1),k_0^*,k_1^*}(\mathcal{F}) \tag{5.20}$$

Then, one can observe that the Lagrangian relaxation of the problem $(\mathcal{P}_{TR}^{\prime\prime(u_0,u_1)}(k_0^*,k_1^*))$ - from which $B_{LD}^{\prime\prime(u_0,u_1),k_0^*,k_1^*}(\mathcal{F})$ is computed - is also a relaxation of the problem $(\mathcal{P}_{LD}^{\prime\prime(u_0,u_1)})$

for which all the dual variables corresponding to constraints that are not related with the system transitions $\left(x^{(u_0),k_0^*}, r^{(u_0),k_0^*}, y^{(u_0),k_0^*}\right)$ and $\left(x^{(u_1),k_1^*}, r^{(u_1),k_1^*}, y^{(u_1),k_1^*}\right)$ would be forced to zero, i.e.

$$
\begin{aligned}
B_{LD}^{\prime\prime(u_0,u_1),k_0^*,k_1^*}(\mathcal{F}) = \max_{\boldsymbol{\lambda}\in\mathbb{R}_+^{n^{(u_0)}},\boldsymbol{\mu}\in\mathbb{R}_+^{n^{(u_1)}}} & \frac{-\|L_\rho^2 X\boldsymbol{\mu} - Y\boldsymbol{\lambda}\|^2}{-ML_\rho^2 + L} - \frac{\left(1 - 2\bar{r}^T\boldsymbol{\mu}\right)^2}{4M} \\
& + \sum_{k_0=1}^{n^{(u_0)}} \lambda_{k_0}\left(\left\|y^{(u_0),k_0}\right\|^2 - L_f^2\left\|x^{(u_0),k_0} - \hat{x}_0\right\|^2\right) \\
& + \sum_{k_1=1}^{n^{(u_1)}} \mu_{k_1}\left(\left(r^{(u_1),k_1}\right)^2 - L_\rho^2\left\|x^{(u_1),k_1}\right\|^2\right) \\
\text{subject to} \quad & M > 0\,, \\
& L > ML_\rho^2\,, \\
& \lambda_{k_0} = 0 \text{ if } k_0 \neq k_0^*, \forall k_0 \in \left\{1,\ldots,n^{(u_0)}\right\}\,, \\
& \mu_{k_1} = 0 \text{ if } k_1 \neq k_1^*, \forall k_1 \in \left\{1,\ldots,n^{(u_1)}\right\}\,.
\end{aligned}
$$

We therefore have:

$$
B_{LD}^{\prime\prime(u_0,u_1),k_0^*,k_1^*}(\mathcal{F}) \leq B_{LD}^{\prime\prime(u_0,u_1)}(\mathcal{F})\,. \tag{5.21}
$$

By definition of the Lagrangian relaxation bound $B_{LD}^{(u_0,u_1)}(\mathcal{F})$, we have:

$$
B_{LD}^{(u_0,u_1)}(\mathcal{F}) = \hat{\mathbf{r}}_{\mathbf{0}}^* + B_{LD}^{\prime\prime(u_0,u_1)}(\mathcal{F}) \tag{5.22}
$$

Equations (5.20), (5.21) and (5.22) finally give:

$$
B_{TR}^{(u_0,u_1)}(\mathcal{F}') = B_{LD}^{(u_0,u_1)}(\mathcal{F})\,.
$$

□

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

THEOREM 5.14. $\forall\,(u_0,u_1) \in \mathcal{U}^2$,

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

Proof. Let $(u_0,u_1) \in \mathcal{U}^2$. The inequality

$$
B_{CGRL}^{(u_0,u_1)}(\mathcal{F}) \leq B_{TR}^{(u_0,u_1)}(\mathcal{F}) \leq B_{LD}^{(u_0,u_1)}(\mathcal{F}) \tag{5.23}
$$

is a straightforward consequence of Theorems 5.9 and 5.13. The inequality

$$
B_{LD}^{(u_0,u_1)}(\mathcal{F}) \leq B_2^{(u_0,u_1)}(\mathcal{F}) \tag{5.24}
$$

is a property of the Lagrangian relaxation, and the inequality

$$
B_2^{(u_0,u_1)}(\mathcal{F}) \leq J_2^{(u_0,u_1)}
$$

is derived from the formalization of the $\min\max$ generalization problem introduced in [22].
□

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 5.15 (Sample Dispersion). *Since $\mathcal{X}$ is bounded, one has:*

$$\exists\, \alpha > 0 : \forall u \in \mathcal{U}\ , \qquad \sup_{x \in \mathcal{X}} \min_{k \in \left\{1,\ldots,n^{(u)}\right\}} \left\| x^{(u),k} - x \right\| \leq \alpha\ . \tag{5.25}$$

The smallest α which satisfies equation (5.25) 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 non-visited state space area.

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

LEMMA 5.16.
$\exists\, C > 0 : \forall (u_0, u_1) \in \mathcal{U}^2, \forall B^{(u_0,u_1)}(\mathcal{F}) \in \left\{ B^{(u_0,u_1)}_{CGRL}(\mathcal{F}), B^{(u_0,u_1)}_{TR}(\mathcal{F}), B^{(u_0,u_1)}_{LD}(\mathcal{F}) \right\},$

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

Proof. The proof for the case where $B^{(u_0,u_1)}(\mathcal{F}) = B^{(u_0,u_1)}_{CGRL}(\mathcal{F})$ is given in [21]. According to Theorem 5.14, one has:

$$\forall\, (u_0, u_1) \in \mathcal{U}^2, \qquad B^{(u_0,u_1)}_{CGRL}(\mathcal{F}) \leq B^{(u_0,u_1)}_{TR}(\mathcal{F}) \leq B^{(u_0,u_1)}_{LD}(\mathcal{F}) \leq J_2^{(u_0,u_1)},$$

which ends the proof. □

We therefore have the following theorem:

THEOREM 5.17.
$\forall (u_0, u_1) \in \mathcal{U}^2, \forall B^{(u_0,u1)}(\mathcal{F}) \in \left\{ B^{(u_0,u_1)}_{CGRL}(\mathcal{F}), B^{(u_0,u_1)}_{TR}(\mathcal{F}), B^{(u_0,u_1)}_{LD}(\mathcal{F}) \right\},$

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

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

DEFINITION 5.18 (Maximal Bounds).

$$B^{(*)}_{CGRL}(\mathcal{F}) \triangleq \max_{(u_0,u_1) \in \mathcal{U}^2} B^{(u_0,u_1)}_{CGRL}(\mathcal{F})\ ,$$
$$B^{(*)}_{TR}(\mathcal{F}) \triangleq \max_{(u_0,u_1) \in \mathcal{U}^2} B^{(u_0,u_1)}_{TR}(\mathcal{F})\ ,$$
$$B^{(*)}_{LD}(\mathcal{F}) \triangleq \max_{(u_0,u_1) \in \mathcal{U}^2} B^{(u_0,u_1)}_{LD}(\mathcal{F})\ .$$

We also denote by $(u_0,u_1)_{\mathcal{F}}^{CGRL}$ (resp. $(u_0,u_1)_{\mathcal{F}}^{TR}$ and $(u_0,u_1)_{\mathcal{F}}^{LD}$) three sequences of actions that maximize the bounds:

DEFINITION 5.19 (Bound-optimal Sequences of Actions).

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

We finally give in this section a last theorem that shows the convergence of the sequences of actions $(u_0,u_1)_{\mathcal{F}}^{CGRL}$, $(u_0,u_1)_{\mathcal{F}}^{TR}$ and $(u_0,u_1)_{\mathcal{F}}^{LD}$ towards optimal sequences of actions - i.e. sequences of actions that lead to an optimal return J_2^* - when the sample dispersion $\alpha^*(\mathcal{F})$ decreases towards zero.

THEOREM 5.20. *Let $\mathfrak{J}^*$ be the set of optimal two-stage sequences of actions:*

$$\mathfrak{J}_2^* \triangleq \left\{(u_0,u_1)\in\mathcal{U}^2 | J_2^{(u_0,u_1)} = J_2^*\right\},$$

and let us suppose that $\mathfrak{J}_2^* \neq \mathcal{U}^2$ (if $\mathfrak{J}_2^* = \mathcal{U}^2$, the search for an optimal sequence of actions is indeed trivial). We define

$$\epsilon \triangleq \min_{(u_0,u_1)\in\mathcal{U}^2\setminus\mathfrak{J}_2^*} \left\{J_2^* - J_2^{(u_0,u_1)}\right\}.$$

Then, $\forall (\tilde{u}_0,\tilde{u}_1)_{\mathcal{F}} \in \left\{(u_0,u_1)_{\mathcal{F}}^{CGRL}, (u_0,u_1)_{\mathcal{F}}^{TR}, (u_0,u_1)_{\mathcal{F}}^{LD}\right\}$,

$$\left(C\alpha^*(\mathcal{F}) < \epsilon\right) \implies (\tilde{u}_0,\tilde{u}_1)_{\mathcal{F}} \in \mathfrak{J}_2^*. \tag{5.26}$$

Proof. Let us prove the theorem by contradiction. Let us assume that $C\alpha^*(\mathcal{F}) < \epsilon$. Let $B^{(u_0,u1)}(\mathcal{F}) \in \left\{B_{CGRL}^{(u_0,u_1)}(\mathcal{F}), B_{TR}^{(u_0,u_1)}(\mathcal{F}), B_{LD}^{(u_0,u_1)}(\mathcal{F})\right\}$, and let $(\tilde{u}_0,\tilde{u}_1)_{\mathcal{F}}$ be a sequence such that

$$(\tilde{u}_0,\tilde{u}_1)_{\mathcal{F}} \in \underset{(u_0,u_1)\in\mathcal{U}^2}{\arg\max}\, B^{(u_0,u_1)}(\mathcal{F})$$

and let us assume that $(\tilde{u}_0,\tilde{u}_1)_{\mathcal{F}}$ is not optimal. This implies that

$$J_2^{(\tilde{u}_0,\tilde{u}_1)_{\mathcal{F}}} \leq J_2^* - \epsilon.$$

Now, let us consider a sequence $(u_0^*,u_1^*) \in \mathfrak{J}_2^*$. Then

$$J_2^{(u_0^*,u_1^*)} = J_2^*.$$

The lower bound $B^{(u_0^*,u_1^*)}(\mathcal{F})$ satisfies the relationship

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

Knowing that $C\alpha^*(\mathcal{F}) < \epsilon$, we have

$$B^{(u_0^*,u_1^*)}(\mathcal{F}) > J_2^* - \epsilon.$$

Since

$$J_2^{(\tilde{u}_0,\tilde{u}_1)_{\mathcal{F}}} \geq B^{(\tilde{u}_0,\tilde{u}_1)_{\mathcal{F}}}(\mathcal{F}),$$

we have

$$B^{(u_0^*,u_1^*)}(\mathcal{F}) > B^{(\tilde{u}_0,\tilde{u}_1)_{\mathcal{F}}}(\mathcal{F})$$

which contradicts the fact that $(\tilde{u}_0,\tilde{u}_1)_{\mathcal{F}}$ belongs to the set $\underset{(u_0,u_1)\in\mathcal{U}^2}{\arg\max}\ B^{(u_0,u_1)}(\mathcal{F})$. This ends the proof. □

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), 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, Trust-region and Lagrangian relaxation 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. 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 $i-$th 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 to solve a conic-quadratic program, are done using SeDuMi [47].

6.2. Protocol and Results.

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

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

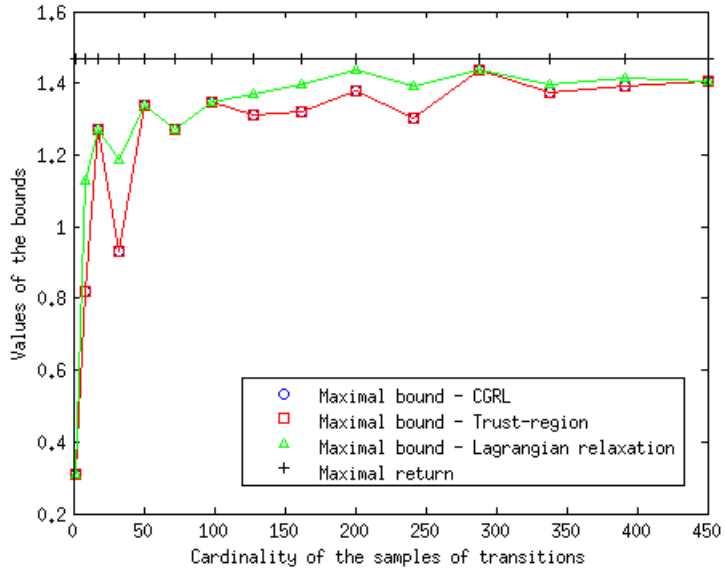

FIG. 6.1. *Bounds* $B^{(*)}_{CGRL}(\mathcal{F}_{c_i})$, $B^{(*)}_{TR}(\mathcal{F}_{c_i})$ *and* $B^{(*)}_{LD}(\mathcal{F}_{c_i})$ *computed from all samples of transitions* $\mathcal{F}_{c_i}$ $i \in \{1,\ldots,15\}$ *of cardinality* $c_i = 2i^2$.

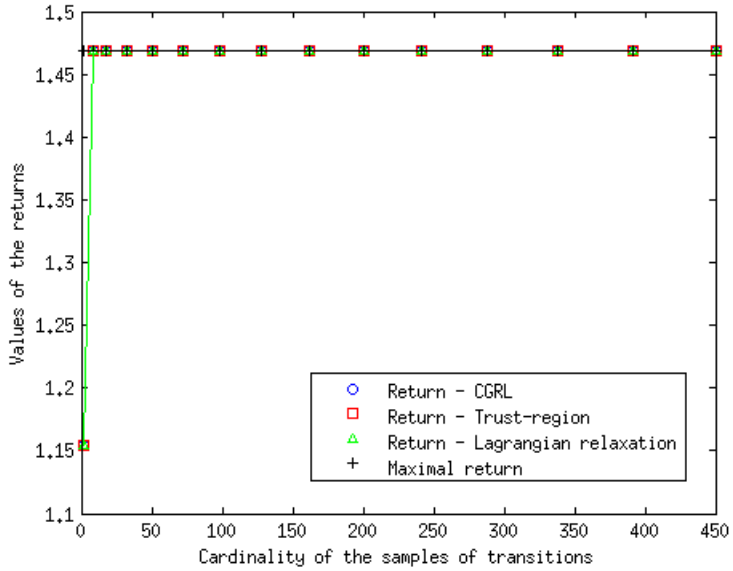

FIG. 6.2. *Returns of the sequences* $(u_0,u_1)^{CGRL}_{\mathcal{F}_{c_i}}$, $(u_0,u_1)^{TR}_{\mathcal{F}_{c_i}}$ *and* $(u_0,u_1)^{LD}_{\mathcal{F}_{c_i}}$ *computed from all samples of transitions* $\mathcal{F}_{c_i}$ $i \in \{1,\ldots,15\}$ *of cardinality* $c_i = 2i^2$.

and

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

We report in Figure 6.1 the values of the maximal CGRL bound $B^{(*)}_{CGRL}(\mathcal{F}_{c_i})$, the maximal Trust-region bound $B^{(*)}_{TR}(\mathcal{F}_{c_i})$ and the maximal Lagrangian relaxation bound $B^{(*)}_{LD}(\mathcal{F}_{c_i})$ as 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 $J_2^{(u_0,u_1)^{CGRL}_{\mathcal{F}_{c_i}}}$, $J_2^{(u_0,u_1)^{TR}_{\mathcal{F}_{c_i}}}$ and $J_2^{(u_0,u_1)^{LD}_{\mathcal{F}_{c_i}}}$ of the bound-optimal sequences of actions $(u_0,u_1)^{CGRL}_{\mathcal{F}_{c_i}}$, $(u_0,u_1)^{TR}_{\mathcal{F}_{c_i}}$ 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 or equal to the Trust-region bound, which is also greater or equal to the CGRL bound as predicted by Theorem 5.14. On the other hand, no difference were observed in terms of return of the bound-optimal sequences of actions.

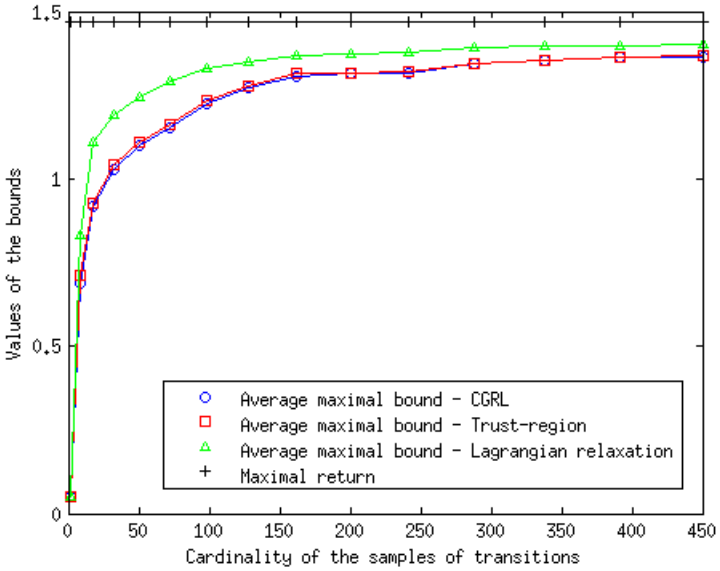

FIG. 6.3. *Average values $A_{CGRL}(c_i)$, $A_{TR}(c_i)$ and $A_{LD}(c_i)$ of the bounds computed from all samples of transitions $\mathcal{F}_{c_i,k}$ $k \in \{1,\ldots,100\}$ of cardinality $c_i = 2i^2$.*

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 $c_i = 2i^2, i = 1,\ldots,15$, 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 Trust-region bound $B^{(*)}_{TR}(\mathcal{F}_{c_i,k})$ and the maximal Lagrangian relaxation bound $B^{(*)}_{LD}(\mathcal{F}_{c_i,k})$. We then compute

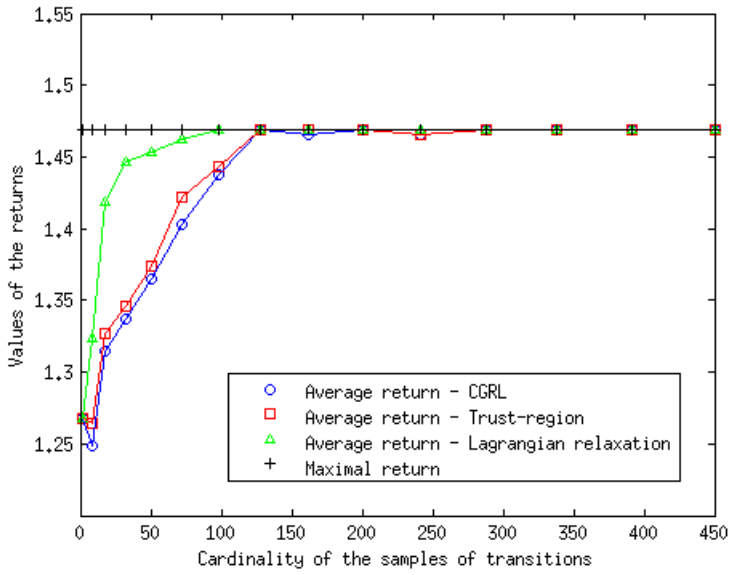

FIG. 6.4. *Average values J_{CGRL}, J_{TR} and J_{LD} of the return of the bound-optimal sequences of actions computed from all samples of transitions $\mathcal{F}_{c_i,k}$ $k \in \{1,\ldots,100\}$ of cardinality $c_i = 2i^2$.*

the average values of the maximal CGRL, Trust-region and Lagrangian relaxation bounds :

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

$$A_{TR}(c_i) = \frac{1}{100}\sum_{k=1}^{100} B^{(*)}_{TR}\left(\mathcal{F}_{c_i,k}\right)$$

$$A_{LD}(c_i) = \frac{1}{100}\sum_{k=1}^{100} B^{(*)}_{LD}\left(\mathcal{F}_{c_i,k}\right)$$

and we report in Figure 6.3 the values $A_{CGRL}(c_i)$ (resp. $A_{TR}(c_i)$ 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 $(u_0,u_1)^{CGRL}_{\mathcal{F}_{c_i,k}}$, $(u_0,u_1)^{TR}_{\mathcal{F}_{c_i,k}}$ and $(u_0,u_1)^{LD}_{\mathcal{F}_{c_i,k}}$:

$$\forall i \in \{1,\ldots,15\}, \qquad J_{CGRL}(c_i) = \frac{1}{100}\sum_{k=1}^{100} J_2^{(u_0,u_1)^{CGRL}_{\mathcal{F}_{c_i,k}}}$$

$$J_{TR}(c_i) = \frac{1}{100}\sum_{k=1}^{100} J_2^{(u_0,u_1)^{TR}_{\mathcal{F}_{c_i,k}}}$$

$$J_{LD}(c_i) = \frac{1}{100}\sum_{k=1}^{100} J_2^{(u_0,u_1)^{LD}_{\mathcal{F}_{c_i,k}}} \ .$$

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

We observe that, on average, the Lagrangian relaxation bound is much tighter that the Trust-region and the CGRL bounds. The CGRL bound and the Trust-region 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 $(u_0, u_1)^{LD}_{\mathcal{F}_{c_i,k}}$ better perform on average.

7. Conclusions. We have considered in this paper the problem of computing $\min\max$ policies for deterministic, Lipschitz continuous batch mode reinforcement learning. First, we have shown that this $\min\max$ problem is NP-hard. Afterwards, we have proposed for the two-stage case 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.

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

Acknowledgements. Raphael Fonteneau is a Post-doctoral fellow of the FRS-FNRS (Funds for Scientific Research). This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control and Optimization) funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The authors thank Yurii Nesterov for pointing out the idea of using Lagrangian relaxation. The scientific responsibility rests with its authors.

Appendix A. Proof of Lemma 5.12.

Proof. For conciseness, we denote $\left(x^{(u_0),k_0}, r^{(u_0),k_0}, y^{(u_0),k_0}\right)$ (resp. $\left(x^{(u_1),k_1}, r^{(u_1),k_1}, y^{(u_1),k_1}\right)$) by $\left(x^0, r^0, y^0\right)$ (resp. $\left(x^1, r^1, y^1\right)$), and λ_1 (resp. μ_1) by λ (resp. μ). We assume that $x_0 \neq x^0$ and $x^1 \neq y^0$ otherwise the problem is trivial.

- Trust-region solution.

According to Definition 5.2, we have:

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

where

$$\hat{\mathbf{x}}^*_{\mathbf{1}}(k_0,k_1) = y^0 + L_f \frac{\left\|x_0 - x^0\right\|}{\left\|y^0 - x^1\right\|}\left(y^0 - x^1\right),$$

which writes

$$\begin{aligned} B''^{(u_0,u_1),k_0,k_1}_{TR}(\mathcal{F}) &= r^1 - L_\rho \left\| y^0 + L_f \frac{\left\|x_0 - x^0\right\|}{\left\|y^0 - x^1\right\|}\left(y^0 - x^1\right) - x^1 \right\|, \\ &= r^1 - L_\rho \left\|y^0 - x^1\right\| \left(1 + L_f \frac{\left\|x_0 - x^0\right\|}{\left\|y^0 - x^1\right\|}\right) \\ &= r^1 - L_\rho \left\|y^0 - x^1\right\| - L_\rho L_f \left\|x_0 - x^0\right\| \end{aligned}$$

- Lagrangian relaxation based solution.

According to Equation (5.11), we can write:

$$B_{LD}^{\prime\prime(u_0,u_1),k_0,k_1}(\mathcal{F}) = \max_{\lambda\in\mathbb{R}_+,\mu\in\mathbb{R}_+} \frac{-\left\|L_\rho^2 x^1\mu - y^0\lambda\right\|^2}{-\mu L_\rho^2+\lambda} - \frac{\left(1-2r^1\mu\right)^2}{4\mu} \\ +\lambda\left(\left\|y^0\right\|^2 - L_f^2\left\|x^0-x_0\right\|^2\right) \quad +\mu\left(\left(r^1\right)^2 - L_\rho^2\left\|x^1\right\|^2\right)$$

subject to

$$\mu > 0 \\ \lambda > \mu L_\rho^2$$

We denote by $\mathcal{L}(\lambda,\mu)$ the quantity:

$$\mathcal{L}(\lambda,\mu) = \frac{-\left\|L_\rho^2 x^1\mu - y^0\lambda\right\|^2}{-\mu L_\rho^2+\lambda} - \frac{\left(1-2r^1\mu\right)^2}{4\mu} + \lambda\left(\left\|y^0\right\|^2 - L_f^2\left\|x^0-x_0\right\|^2\right) \\ +\mu\left(\left(r^1\right)^2 - L_\rho^2\left\|x^1\right\|^2\right)$$

Let λ and μ be such that $\lambda > \mu L_\rho^2$. Since the Trust-region solution to $(\mathcal{P}_{TR}^{\prime\prime(u_0,u_1)}(k_0,k_1))$ is optimal, and by property of the Lagrangian relaxation [25], one has:

$$\mathcal{L}(\lambda,\mu) \leq B_{TR}^{\prime\prime(u_0,u_1),k_0,k_1}(\mathcal{F})\,. \tag{A.1}$$

In order to prove the lemma, it is therefore sufficient to determine two values λ_0 and μ_0 such that the inequality (A.1) is an equality. By differentiating $\mathcal{L}(\lambda,\mu)$, we obtain, after a long calculation (that we omit here):

$$\left(\begin{cases} \frac{\partial\mathcal{L}(\lambda,\mu)}{\partial\lambda} = 0 \\ \frac{\partial\mathcal{L}(\lambda,\mu)}{\partial\mu} = 0 \\ \lambda > \mu L_\rho^2 \end{cases}\right) \implies \begin{cases} \lambda = \frac{L_\rho}{2L_f\|x^0-x_0\|}\,, \\ \mu = \frac{1}{2L_\rho(\|y^0-x^1\|+L_f\|x^0-x_0\|)}\,. \end{cases}$$

We denote by λ_0 and μ_0 the following values for the dual variables:

$$\lambda_0 \triangleq \frac{L_\rho}{2L_f\left\|x^0-x_0\right\|}\,, \\ \mu_0 \triangleq \frac{1}{2L_\rho\left(\left\|y^0-x^1\right\|+L_f\left\|x^0-x_0\right\|\right)}\,.$$

We have:

$$\mu_0 = \frac{1}{2L_\rho\left(\left\|y^0-x^1\right\|+L_f\left\|x^0-x_0\right\|\right)} > 0 \\ \frac{\lambda_0}{\mu_0} = L_\rho^2\left(1+\frac{\left\|y^0-x^1\right\|}{L_f\left\|x^0-x_0\right\|}\right) > L_\rho^2\,.$$

In the following, we denote $\left(1+\frac{\|y^0-x^1\|}{L_f\|x^0-x_0\|}\right)$ by K. We now give the expression of $\mathcal{L}(\lambda_0,\mu_0)$ using only μ_0 and K:

$$\begin{aligned}\mathcal{L}(\lambda_0,\mu_0) = &-\frac{L_\rho^4\mu_0^2\left\|x^1-Ky^0\right\|^2}{\mu_0 L_\rho^2(-1+K)} - \frac{1}{4\mu_0} + r^1 - \left(r^1\right)^2\mu_0 \\ &+\mu_0 K L_\rho^2\left(\left\|y^0\right\|^2 - L_f^2\left\|x^0-x_0\right\|^2\right) + \mu_0\left(\left(r^1\right)^2 - L_\rho^2\left\|x^1\right\|^2\right) \\ = &-\frac{L_\rho^2\mu_0\left\|x^1-Ky^0\right\|^2}{K-1} - \frac{1}{4\mu_0} + r^1 \\ &+L_\rho^2\mu_0 K\left(\left\|y^0\right\|^2 - L_f^2\left\|x^0-x_0\right\|^2\right) - L_\rho^2\mu_0\left\|x^1\right\|^2 \ .\end{aligned}$$

Using the fact that $x^1 - Ky^0 = x^1 - y^0 - (K-1)y^0$, we can write:

$$\begin{aligned}\mathcal{L}(\lambda_0,\mu_0) = &-\frac{L_\rho^2\mu_0}{K-1}\left(\left\|x^1-y^0\right\|^2 + (K-1)^2\left\|y^0\right\|^2 - 2(K-1)(x^1-y^0)^T y^0\right) \\ &-\frac{1}{4\mu_0} + r^1 + L_\rho^2\mu_0 K\left(\left\|y^0\right\|^2 - L_f^2\left\|x^0-x_0\right\|^2\right) - L_\rho^2\left\|x^1\right\|^2\end{aligned}$$

and

$$\begin{aligned}\mathcal{L}(\lambda_0,\mu_0) = &-\frac{L_\rho^2\mu_0}{K-1}\left\|x^1-y^0\right\|^2 - L_\rho^2\mu_0(K-1)\left\|y^0\right\|^2 + 2L_\rho^2\mu_0(x^1-y^0)^T y^0 \\ &-\frac{1}{4\mu_0} + r^1 + L_\rho^2\mu_0 K\left(\left\|y^0\right\|^2 - L_f^2\left\|x^0-x_0\right\|^2\right) - L_\rho^2\left\|x^1\right\|^2 \ .\end{aligned}$$

From there,

$$\begin{aligned}\mathcal{L}(\lambda_0,\mu_0) = &-\frac{L_\rho^2\mu_0}{K-1}\left\|x^1-y^0\right\|^2 + \left\|y^0\right\|^2\left(-L_\rho^2\mu_0(K-1) - 2L_\rho^2\mu_0 + L_\rho^2\mu_0 K\right) \\ &-2L_\rho^2\mu_0(x^1)^T y^0 - \frac{1}{4\mu_0} + r^1 + L_\rho^2\mu_0 K\left(-L_f^2\left\|x^0-x_0\right\|^2\right) - L_\rho^2\left\|x^1\right\|^2\end{aligned}$$

and, since $\left(-L_\rho^2\mu_0(K-1) - 2L_\rho^2\mu_0 + L_\rho^2\mu_0 K\right) = -L_\rho^2\mu_0$, we have that

$$\begin{aligned}\mathcal{L}(\lambda_0,\mu_0) = &-\frac{L_\rho^2\mu_0}{K-1}\left\|x^1-y^0\right\|^2 - L_\rho^2\mu_0\left\|y^0\right\|^2 - 2L_\rho^2\mu_0(x^1)^T y^0 \\ &-\frac{1}{4\mu_0} + r^1 + L_\rho^2\mu_0 K\left(-L_f^2\left\|x^0-x_0\right\|^2\right) - L_\rho^2\left\|x^1\right\|^2\end{aligned}$$

and

$$\begin{aligned}\mathcal{L}(\lambda_0,\mu_0) = &-\frac{L_\rho^2\mu_0}{K-1}\left\|x^1-y^0\right\|^2 - L_\rho^2\mu_0\left(\left\|y^0\right\|^2 + \left\|x^1\right\|^2 - 2(x^1)^T y^0\right) \\ &-\frac{1}{4\mu_0} + r^1 + L_\rho^2\mu_0 K\left(-L_f^2\left\|x^0-x_0\right\|^2\right) \ .\end{aligned}$$

Since $\left(\left\|y^0\right\|^2 + \left\|x^1\right\|^2 - 2(x^1)^T y^0\right) = \left\|x^1 - y^0\right\|^2$, we have:

$$\begin{aligned}
\mathcal{L}(\lambda_0, \mu_0) &= -\frac{L_\rho^2 \mu_0}{K-1} \left\|x^1 - y^0\right\|^2 - L_\rho^2 \mu_0 \left\|x^1 - y^0\right\|^2 \\
&\quad -\frac{1}{4\mu_0} + r^1 + L_\rho^2 \mu_0 K \left(-L_f^2 \left\|x^0 - x_0\right\|^2\right) \\
&= -L_\rho^2 \mu_0 \left\|x^1 - y^0\right\|^2 \frac{K}{K-1} \\
&\quad -\frac{1}{4\mu_0} + r^1 + L_\rho^2 \mu_0 K \left(-L_f^2 \left\|x^0 - x_0\right\|^2\right) \\
&= -K L_\rho^2 \mu_0 \left(\frac{\left\|x^1 - y^0\right\|^2}{K-1} + L_f^2 \left\|x^0 - x_0\right\|^2\right) - \frac{1}{4\mu_0} + r^1 \, .
\end{aligned}$$

Since $K \triangleq \left(1 + \frac{\left\|y^0 - x^1\right\|}{L_f \left\|x^0 - x_0\right\|}\right)$, we have:

$$\begin{aligned}
\mathcal{L}(\lambda_0, \mu_0) &= -\left(1 + \frac{\left\|y^0 - x^1\right\|}{L_f \left\|x^0 - \hat{x}_0\right\|}\right) L_\rho^2 \mu_0 \left(\frac{\left\|x^1 - y^0\right\|^2}{\left(1 + \frac{\left\|y^0 - x^1\right\|}{L_f \left\|x^0 - x_0\right\|}\right) - 1} + L_f^2 \left\|x^0 - x_0\right\|^2\right) \\
&\quad -\frac{1}{4\mu_0} + r^1 \\
&= -\frac{L_\rho^2 \mu_0 \left(L_f \left\|x^0 - x_0\right\| + \left\|y^0 - x^1\right\|\right)}{L_f \left\|x^0 - x_0\right\|} \left(\frac{L_f \left\|x^0 - x_0\right\| \left\|x^1 - y^0\right\|^2}{\left\|x^1 - y^0\right\|} + L_f^2 \left\|x^0 - x_0\right\|^2\right) \\
&\quad -\frac{1}{4\mu_0} + r^1 \\
&= -L_\rho^2 \mu_0 \left(L_f \left\|x^0 - x_0\right\| + \left\|y^0 - x^1\right\|\right) \left(\left\|x^1 - y^0\right\| + L_f \left\|x^0 - x_0\right\|\right) - \frac{1}{4\mu_0} + r^1 \, .
\end{aligned}$$

Since $\mu_0 = \frac{1}{2L_\rho(\|y^0 - x^1\| + L_f \|x^0 - x_0\|)}$, we finally obtain:

$$\begin{aligned}
\mathcal{L}(\lambda_0, \mu_0) &= -\frac{L\rho}{2} \left(\left\|y^0 - x^1\right\| + L_f \left\|x^0 - x_0\right\|\right) - \frac{L\rho}{2} \left(\left\|y^0 - x^1\right\| + L_f \left\|x^0 - x_0\right\|\right) + r^1 \\
&= r^1 - L_\rho \left\|y^0 - x^1\right\| - L_\rho L_f \left\|x^0 - x_0\right\| \\
&= B_{TR}^{\prime\prime(u_0,u_1),k_0,k_1}(\mathcal{F}) \, ,
\end{aligned}$$

which ends the proof. □

REFERENCES

[1] A. BANERJEE AND A.A. TSIATIS, *Adaptive two-stage designs in phase ii clinical trials*, Statistics in medicine, 25 (2006), pp. 3382–3395.

[2] T. BAŞAR AND P. BERNHARD, *H_∞-optimal control and related minimax design problems: a dynamic game approach*, vol. 5, Birkhauser, 1995.

[3] A. BECK AND Y.C. ELDAR, *Strong duality in nonconvex quadratic optimization with two quadratic constraints*, SIAM Journal on Optimization, 17 (2007), pp. 844–860.

[4] A. BEMPORAD AND M. MORARI, *Robust model predictive control: A survey*, Robustness in Identification and Control, 245 (1999), pp. 207–226.

[5] A. BEN-TAL AND A.S. NEMIROVSKI, *Lectures on Modern Convex Optimization*, Siam, 2001.

[6] D.P. BERTSEKAS AND J.N. TSITSIKLIS, *Neuro-Dynamic Programming*, Athena Scientific, 1996.

[7] J.R. BIRGE AND F. LOUVEAUX, *Introduction to Stochastic Programming*, Springer Verlag, 1997.
[8] JF BONNANS, J.C. GILBERT, C. LEMARÉCHAL, AND C. SAGASTIZÁBAL, *Numerical optimization, theoretical and numerical aspects*, 2006.
[9] S.P. BOYD AND L. VANDENBERGHE, *Convex Optimization*, Cambridge Univ Pr, 2004.
[10] S.J. BRADTKE AND A.G. BARTO, *Linear least-squares algorithms for temporal difference learning*, Machine Learning, 22 (1996), pp. 33–57.
[11] L. BUSONIU, R. BABUSKA, B. DE SCHUTTER, AND D. ERNST, *Reinforcement Learning and Dynamic Programming using Function Approximators*, Taylor & Francis CRC Press, 2010.
[12] E.F. CAMACHO AND C. BORDONS, *Model Predictive Control*, Springer, 2004.
[13] A.R. CONN, N.I.M. GOULD, AND P.L. TOINT, *Trust-region Methods*, vol. 1, Society for Industrial Mathematics, 2000.
[14] K. DARBY-DOWMAN, S. BARKER, E. AUDSLEY, AND D. PARSONS, *A two-stage stochastic programming with recourse model for determining robust planting plans in horticulture*, Journal of the Operational Research Society, (2000), pp. 83–89.
[15] B. DEFOURNY, D. ERNST, AND L. WEHENKEL, *Risk-aware decision making and dynamic programming*, Selected for oral presentation at the NIPS-08 Workshop on Model Uncertainty and Risk in Reinforcement Learning, Whistler, Canada, (2008).
[16] E. DELAGE AND S. MANNOR, *Percentile optimization for Markov decision processes with parameter uncertainty*, Operations Research, 58 (2010), pp. 203–213.
[17] D. ERNST, P. GEURTS, AND L. WEHENKEL, *Tree-based batch mode reinforcement learning*, Journal of Machine Learning Research, 6 (2005), pp. 503–556.
[18] D. ERNST, M. GLAVIC, F. CAPITANESCU, AND L. WEHENKEL, *Reinforcement learning versus model predictive control: a comparison on a power system problem*, IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics, 39 (2009), pp. 517–529.
[19] R. FONTENEAU, *Contributions to Batch Mode Reinforcement Learning*, PhD thesis, University of Liège, 2011.
[20] R. FONTENEAU, S. MURPHY, L. WEHENKEL, AND D. ERNST, *Inferring bounds on the performance of a control policy from a sample of trajectories*, in Proceedings of the 2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (IEEE ADPRL 09), Nashville, TN, USA, 2009.
[21] R. FONTENEAU, S.A. MURPHY, L. WEHENKEL, AND D. ERNST, *A cautious approach to generalization in reinforcement learning*, in Proceedings of the Second International Conference on Agents and Artificial Intelligence (ICAART 2010), Valencia, Spain, 2010.
[22] R. FONTENEAU, S. A. MURPHY, L. WEHENKEL, AND D. ERNST, *Towards min max generalization in reinforcement learning*, in Agents and Artificial Intelligence: International Conference, ICAART 2010, Valencia, Spain, January 2010, Revised Selected Papers. Series: Communications in Computer and Information Science (CCIS), vol. 129, Springer, Heidelberg, 2011, pp. 61–77.
[23] K. FRAUENDORFER, *Stochastic Two-stage Programming*, Springer, 1992.
[24] L.P. HANSEN AND T.J. SARGENT, *Robust Control and Model Uncertainty*, American Economic Review, (2001), pp. 60–66.
[25] J.B. HIRIART-URRUTY AND C. LEMARÉCHAL, *Convex Analysis and Minimization Algorithms: Fundamentals*, vol. 305, Springer-Verlag, 1996.
[26] J.E. INGERSOLL, *Theory of Financial Decision Making*, Rowman and Littlefield Publishers, Inc., 1987.
[27] S. KOENIG, *Minimax real-time heuristic search*, Artificial Intelligence, 129 (2001), pp. 165–197.
[28] M.G. LAGOUDAKIS AND R. PARR, *Least-squares policy iteration*, Jounal of Machine Learning Research, 4 (2003), pp. 1107–1149.
[29] M. L. LITTMAN, *Markov games as a framework for multi-agent reinforcement learning*, in Proceedings of the Eleventh International Conference on Machine Learning (ICML 1994), New Brunswick, NJ, USA, 1994.
[30] ———, *A tutorial on partially observable markov decision processes*, Journal of Mathematical Psychology, 53 (2009), pp. 119 – 125. Special Issue: Dynamic Decision Making.
[31] Y. LOKHNYGINA AND A.A. TSIATIS, *Optimal two-stage group-sequential designs*, Journal of Statistical Planning and Inference, 138 (2008), pp. 489–499.
[32] J.K. LUNCEFORD, M. DAVIDIAN, AND A.A. TSIATIS, *Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials*, Biometrics, (2002), pp. 48–57.
[33] S. MANNOR, D. SIMESTER, P. SUN, AND J.N. TSITSIKLIS, *Bias and variance in value function estimation*, in Proceedings of the Twenty-first International Conference on Machine Learning (ICML 2004), Banff, Alberta, Canada, 2004.
[34] S.A. MURPHY, *Optimal dynamic treatment regimes*, Journal of the Royal Statistical Society, Series B, 65(2) (2003), pp. 331–366.
[35] S.A. MURPHY, *An experimental design for the development of adaptive treatment strategies*, Statistics in Medicine, 24 (2005), pp. 1455–1481.
[36] A. NEMIROVSKI, A. JUDITSKY, G. LAN, AND A. SHAPIRO, *Robust stochastic approximation approach to*

stochastic programming, SIAM Journal on Optimization, 19 (2009), pp. 1574–1609.

[37] Y. NESTEROV AND A. NEMIROVSKI, *Interior point polynomial methods in convex programming*, Studies in applied mathematics, 13 (1994).

[38] D. ORMONEIT AND S. SEN, *Kernel-based reinforcement learning*, Machine Learning, 49 (2002), pp. 161–178.

[39] C. PADURARU, D. PRECUP, AND J. PINEAU, *A framework for computing bounds for the return of a policy*, in Ninth European Workshop on Reinforcement Learning (EWRL9), 2011.

[40] C.H. PAPADIMITRIOU, *Computational Complexity*, John Wiley and Sons Ltd., 2003.

[41] M. QIAN AND S.A. MURPHY, *Performance guarantees for individualized treatment rules*, Tech. Report 498, Department of Statistics, University of Michigan, 2009.

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

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

[44] P. SCOKAERT AND D. MAYNE, *Min-max feedback model predictive control for constrained linear systems*, IEEE Transactions on Automatic Control, 43 (1998), pp. 1136–1142.

[45] A. SHAPIRO, *A dynamic programming approach to adjustable robust optimization*, Operations Research Letters, 39 (2011), pp. 83–87.

[46] ———, *Minimax and risk averse multistage stochastic programming*, tech. report, School of Industrial & Systems Engineering, Georgia Institute of Technology, 2011.

[47] J.F. STURM, *Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones*, Optimization methods and software, 11 (1999), pp. 625–653.

[48] R.S. SUTTON AND A.G. BARTO, *Reinforcement Learning*, MIT Press, 1998.

[49] A.S. WAHED AND A.A. TSIATIS, *Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials*, Biometrics, 60 (2004), pp. 124–133.