Augmented Lagrangian Method for Instantaneously Constrained Reinforcement Learning Problems

Jingqi Li, David Fridovich-Keil, Somayeh Sojoudi, and Claire J. Tomlin

***Abstract*— In this paper, we study the Instantaneously Constrained Reinforcement Learning (ICRL) problem, in which we are tasked to find a policy maximizing a reward while satisfying certain constraints at each time step. We first extend a result on strong duality of Constrained Markov Decision Process (CMDP) in the literature and propose a sufficient condition for strong duality of the ICRL problem. Inspired by the Augmented Lagrangian Method in constrained optimization, we propose a new surrogate objective function for ICRL, which could be efficiently optimized by common policy-gradient based RL algorithms. We show theoretically that a feasible and optimal policy could be obtained by optimizing this surrogate function, under certain conditions related to the feasible policy set. Our empirical results on a tabular Markov Decision Process and two nonlinear optimal control problems, a constrained pendulum and a constrained half-cheetah, justify our analysis, and suggest that our method could promote safety during learning and converge in a smaller number of iterations compared to the existing algorithms.**

I. Introduction

Deep reinforcement learning algorithms have achieved state-of-the-art performance in many domains [1–3]. In standard reinforcement learning (RL), the ultimate goal is to optimize the expected sum of rewards or costs, and the agent can freely explore in order to improve the current policy. RL methods have been widely used to learn optimal policies for agents with complicated or even unknown dynamics. RL has successfully solved a wide range of tasks, including the game of Go [4], robotics control [5], and traffic control [6].

There is a well-known trade-off between exploration and exploitation in RL. To optimize the overall reward, the agent must balance whether to take a sequence of actions similar to what it has already tried (i.e., exploitation) or to try a new combination of actions (i.e., exploration). Since most RL problems are non-convex, pure exploitation leads to a suboptimal policy leading to a poor local maximum of the reward function. To encourage the agent to find a better policy, various methods have been proposed for promoting exploration, such as using an Upper Confidence Bound [7], Inverse Entropy [8], or designing a Variational Auto-encoder [9]. Nevertheless, in many applications such as autonomous driving [10] and surgical robotics [11], exploration can be dangerous because violating certain constraints even by a small amount may have significant consequences. Thus, ensuring safety is of great importance in real-world applications.

A natural way of encoding safety in RL is through constraints. Here, there are two types of constraints: cumulative constraints (e.g., average vehicle speed) and instantaneous constraints (e.g., collision avoidance at each time). A cumulative constraint requires that an infinite-horizon or a finite-horizon discounted sum of a constraint cost function lie within a certain bound. By contrast, an instantaneous constraint must hold at all time instants. For both problems, the horizon could be either infinite or finite.

One common formulation of RL with cumulative constraints is the Constrained Markov Decision Process (CMDP) framework [12], where the agent optimizes an objective while satisfying constraints on the expectation of an infinite-horizon discounted sum of auxiliary costs. A classical approach to solving CMDPs is the Lagrangian dual method [12]. The Lagrangian approach allows us to transform a constrained control problem to an equivalent minmax unconstrained control problem. Recently, it has been shown that under certain regularity conditions there is no duality gap for infinite-horizon RL problems with cumulative constraints, despite their non-convex nature [13]. This result theoretically justifies the effectiveness of popular Lagrangian-relaxation-based CMDP algorithms, such as Constrained Policy Optimization (CPO) [14], Primal-Dual Policy Optimization (PDO) [15] and Lyapunov-based safe learning [16].

As will be shown in Section II, the satisfaction of cumulative constraints may not lead to the satisfaction of instantaneous constraints. Therefore, it is crucial to develop methods for solving instantaneously constrained RL problems. The authors of [17] propose to solve instantaneously constrained RL problems by optimizing a smoothed version of the worst constraint violation rather than an explicitly constrained objective. One line of work devoted to safe RL with instantaneous constraints is projection-based Safe RL [18–20], where at each step the agent selects one action from a pre-computed safe action set. However, one potential drawback of this approach is that the pre-computed safe action set could be conservative, leading to a suboptimal policy [21, 22].

In this paper, we consider an infinite-horizon optimal control problem with instantaneous safety constraints. We adapt the classical Augmented Lagrangian method [23] to obtain a safe policy satisfying instantaneous safety constraints. Our work is closely related to [24], where an interior-point method is adapted to solve the safe RL problem. One major

Jingqi Li, Somayeh Sojoudi and Claire J. Tomlin are with the University of California, Berkeley. David Fridovich-Keil is with Stanford University. Correspondence to jingqili@berkeley.edu.

This research is supported by an NSF CAREER award, the Air Force Office of Scientific Research (AFOSR), NSF's CPS FORCES and VeHICaL projects, the UC-Philippine-California Advanced Research Institute, the ONR BRC grant for Multibody Systems Analysis, a DARPA Assured Autonomy grant, and the SRC CONIX Center.

difference is that we relax the assumption of an initial safe policy, which is required in [24].

We first extend the strong duality results in [13] to instantaneously constrained RL, and propose a sufficient condition for the strong duality of the instantaneously constrained RL problem. Inspired by the Augmented Lagrangian method, we then design a surrogate objective function, and we show that under certain conditions on the feasible policy set, the policy returned by optimizing the surrogate function converges to an optimal policy for the original problem. We propose a primal-dual algorithm for optimizing this surrogate function and our empirical results show that the proposed method is more data-efficient than the existing Lagrangian dual method. Our empirical results also suggest that this method reduces the total constraint violation, highlighting the potential of our method for promoting safety throughout learning.

The rest of the paper is organized as follows. In Section II, we formulate the Instantaneously Constrained RL problem. We present our main theoretical results in Sections III and IV, with proofs provided in the Appendix. In Section V, we present three illustrative examples: a tabular learning example, and an Open-AI constrained pendulum and half-cheetah example. Finally, we conclude and discuss future directions in Section VI.

II. From Cumulative to Instantaneous Constraints

In this section, we first review Constrained Markov Decision Processes, and then motivate and introduce our instantaneously constrained RL problem formulation.

A Markov Decision Process (MDP) is a tuple $(\mathcal{Z}, \mathcal{A}, \gamma, r, p_z, p_0)$, where $\mathcal{Z}$ and $\mathcal{A}$ are compact state and action spaces, $\gamma \in [0,1)$ is a discounting factor, $r(z,a) : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$ is the immediate cost function, $p_z(\cdot|z,a)$ is the transition probability distribution density, and p_0 is the initial state distribution density. In addition, let $g(z,a) : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$ be the constraint function. A function $f : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$ is bounded if there exists a constant $c \in \mathbb{R}$ such that $f(z,a) \leq c$, for $\forall (z,a) \in \mathcal{Z} \times \mathcal{A}$. The agent chooses actions sequentially based on a policy $\pi \in \mathcal{P}(\mathcal{Z})$, where $\mathcal{P}(\mathcal{Z})$ is the space of probability measures on $(\mathcal{A}, \mathcal{B}(\mathcal{A}))$ parametrized by elements of $\mathcal{Z}$, where $\mathcal{B}(\mathcal{A})$ are the Borel sets of $\mathcal{A}$.

A Constrained Markov Decision Process was introduced in [25] by incorporating an additional inequality constraint:

$$\begin{aligned} \max_{\pi} \ & \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(z_t, a_t) \middle| \pi\right] \\ \text{s.t.} \ & z_{t+1} \sim p_z(\cdot|z_t, a_t), a_t \sim \pi(z_t), \quad \forall t \in \{0,1,\dots\}, \\ & z_0 \sim p_0, \\ & \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t g(z_t, a_t) \middle| \pi\right] \leq 0. \end{aligned} \tag{1}$$

where $\mathbb{E}[\cdot]$ is the expectation operator. The z_t and a_t are state and action at time $t \in \{0,1,\dots\}$, respectively.

In what follows, we will illustrate with a simple 2D example that a cumulative constraint does not generally provide any guarantees for the associated instantaneous constraints, i.e., solving CMDPs may not be sufficient to ensure the satisfaction of instantaneous constraints.

Example 1. *Consider a linear dynamical system $z_{t+1} = Az_t + Ba_t$, where $A \in \mathbb{R}^{2\times 2}$ and $B \in \mathbb{R}^{2\times 1}$ are specified as*

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{2}$$

Let $\mathcal{K} \subseteq \mathbb{R}^2$ be the feasible policy class. Given an initial point $\check{z}_0 \in \mathbb{R}^2$, we consider the infinite-horizon constrained optimal control problem

$$\begin{aligned} K^* := \arg\max_{K \in \mathcal{K}} \Bigg[& -\sum_{t=0}^{\infty} \left(z_t^\top Q z_t + a_t^\top R a_t \right) \Bigg] \\ & \text{s.t. } z_{t+1} = Az_t + Ba_t, \quad \forall t \in \{0,1,\dots\}, \\ & \quad a_t = Kz_t, \quad \forall t \in \{0,1,\dots\}, \\ & \quad z_0 = \check{z}_0, \\ & \quad \sum_{t=0}^{\infty} z_t \leq 0 \end{aligned} \tag{3}$$

and the instantaneously constrained RL problem

$$\begin{aligned} \tilde{K}^* := \arg\max_{K \in \mathcal{K}} \Bigg[& -\sum_{t=0}^{\infty} \left(z_t^\top Q z_t + a_t^\top R a_t \right) \Bigg] \\ & \text{s.t. } z_{t+1} = Az_t + Ba_t, \quad \forall t \in \{0,1,\dots\}, \\ & \quad a_t = Kz_t, \quad \forall t \in \{0,1,\dots\}, \\ & \quad z_0 = \check{z}_0, \\ & \quad z_t \leq 0, \quad \forall t \in \{0,1,\dots\}. \end{aligned} \tag{4}$$

with the parameters $Q = I_2$ and $R = 1$. If we assume the policy class to be $\mathcal{K} = \mathbb{R}^2$, then the optimal feedback matrix K^* may be found by solving the well-known LQR Riccati equation and recognizing that the constraint $\sum_{t=0}^{\infty} z_t \leq 0$ in (3) *is inactive for K^*. Pick $\tilde{K} \in \mathbb{R}^2$ such that $(A + BK)$ has real positive eigenvalues with magnitude strictly smaller than 1 and $(A + BK)$ has two eigenvectors $v_1 \leq 0$ and $v_2 \leq 0$ whose convex hull contains z_0. By Proposition 4 in the Appendix, we can show that $\tilde{K}$ is a feasible solution for* (4)*. We plot the state trajectories under the two feedback controllers $a_t = K^* z_t$ and $\tilde{a}_t = \tilde{K}\tilde{z}_t$. In Figure 1, the state trajectory under the controller $a_t = K^* z_t$ violates the constraint $z_t \leq 0$ at time $t = 2$ while the trajectory under $\tilde{a}_t = \tilde{K}\tilde{z}_t$ does not.* ■

As illustrated in Example 1, enforcing a constraint cumulatively does not imply that it holds at each time. We emphasize that constraints may be arbitrary functions of state. In this way, an instantaneous constraint may be understood to encode desired safety configurations, such as in collision avoidance [26], human-robot interaction [27], and aerospace control [28]. Motivated by the above discussion, we formulate the Instantaneously Constrained RL problem as follows.

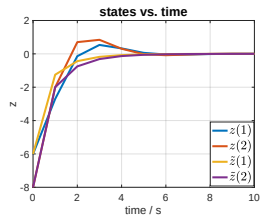


Fig. 1: State trajectories comparison between the two controllers $a_t = K^* z_t$ and $\tilde{a}_t = \tilde{K}\tilde{z}_t$.

Problem 1 (Instantaneously Constrained RL Problem)**.** *Consider an MDP with transition dynamics $z_{t+1} \sim p_z(\cdot|z_t, a_t)$ and initial state distribution p_0, along with a bounded reward function $r(z, a)$ and a bounded constraint function $g(z, a)$. The objective is to find a policy π^* that solves the following constrained optimization problem over the infinite horizon:*

$$\begin{aligned} \max_{\pi} \ & \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(z_t, a_t) \middle| \pi\right] \\ \text{s.t. } & z_{t+1} \sim p_z(\cdot|z_t, a_t), \quad \forall t \in \{0, 1, \dots\}, \\ & a_t \sim \pi(z_t), \quad \forall t \in \{0, 1, \dots\}, \\ & z_0 \sim p_0, \\ & \mathbb{E}[g(z_t, a_t)|\pi] \leq 0, \quad \forall t \in \{0, 1, \dots\}. \end{aligned} \tag{5}$$

We remark here that an optimal policy feasible for (5) could be a conservative but feasible solution for (1). Therefore, a policy learned from (5) is also safe with respect to the constraint in (1). In addition, although we only consider one set of instantaneous constraints in (5), the results of this paper could be extended to the general case with multiple sets of instantaneous constraints, by associating each constraint with a Lagrange multiplier and carrying out an analysis similar to the single constraint case (5).

III. Augmented Lagrangian Surrogate Function

In this section, we introduce our Augmented Lagrangian Surrogate Function. We first propose a sufficient condition under which strong duality holds for (5), and then design a new surrogate function which could promote safety during the learning phase.

Since most of the existing results on RL deal with unconstrained problems, it is beneficial to work with the unconstrained Lagrangian dual of the primal problem (5) given below,

$$\begin{aligned} \min_{\substack{\{\lambda_t\}_{t=0}^{\infty} \\ \lambda_t \leq 0}} \ \max_{\pi} \ & \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left(r(z_t, a_t) + \lambda_t g(z_t, a_t)\right) \middle| \pi\right] \\ \text{s.t. } & z_{t+1} \sim p_z(\cdot|z_t, a_t), \quad \forall t \in \{0, 1, \dots\}, \\ & a_t \sim \pi(z_t), \quad \forall t \in \{0, 1, \dots\}, \\ & z_0 \sim p_0. \end{aligned} \tag{6}$$

where λ_t is the Lagrange multiplier associated with the scalar constraint $\mathbb{E}[g(z_t, a_t)|\pi] \leq 0$.

It is known that strong duality holds in the case of cumulative constraints [13]. For completeness, we first introduce Assumption 1, and then build on the result of [13].

Assumption 1. *Suppose that the feasible policy set for* (5) *has a non-empty relative interior. Furthermore, suppose that for any π satisfying $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t g(z_t, a_t)|\pi] \leq 0$, π also satisfies $\mathbb{E}[g(z_t, a_t)|\pi] \leq 0$, $\forall t \in \{0, 1, \dots\}$.*

Proposition 1. *Under Assumption 1, strong duality holds for* (5)*.*

A natural question that arises is whether Assumption 1 is stringent. To ensure that $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t g(z_t, a_t)|\pi] \leq 0$ implies $\mathbb{E}[g(z_t, a_t)|\pi] \leq 0$, for all $t \in \{0, 1, \dots\}$, we propose two approaches. First, we propose a "clipping" method whereby constraint values at safe states are set to zero. For example, suppose that we have an instantaneous constraint $\mathbb{E}[h(z_t, a_t)|\pi] \leq 0$ with a bounded function $h(z_t, a_t) : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$, computing only the positive part, i.e., $\mathbb{E}[\text{Relu}(h(z_t, a_t))|\pi] \leq 0$, where $\text{Relu}(\cdot) : \mathbb{R} \to \mathbb{R}$ is defined as $\text{Relu}(x) = x$ if $x \geq 0$ and 0 if $x < 0$. The choice of the Relu function is not strictly necessary, i.e., it could be replaced by other non-negative activation functions such as Softplus or Sigmoid [29]. The other approach is to restrict the feasible policy class, as highlighted in the following 2D example:

Example 1 (Continued)**.** *Suppose that the policy class $\mathcal{K} \subseteq \mathbb{R}^2$ is such that for any $K \in \mathcal{K}$, the closed loop dynamics $(A + BK) \in \mathbb{R}^{2\times 2}$ has two real positive eigenvalues, and it has two eigenvectors $v_1 \leq 0$ and $v_2 \leq 0$ whose convex hull contains the point z_0. We will show in Proposition 4 in the Appendix that under any policy $K \in \mathcal{K}$, the constraint $\sum_{t=0}^{\infty} z_t \leq 0$ and the condition $z_t \leq 0$, $\forall t \in \{0, 1, \dots\}$, are always satisfied. In this simple instance where $g(z_t, a_t) \equiv z_t$, the constraint $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t g(z_t, a_t)|\pi] \leq 0$ implies the constraint $\mathbb{E}[g(z_t, a_t)|\pi] \leq 0$, for all $t \in \{0, 1, \dots\}$.*

Remark 1. *However, as indicated in [13], strong duality is only proved for CMDPs with arbitrary stochastic policies. Characterizing strong duality in parametric, restricted policy classes such as that of Problem 1 is an important direction for future research.*

Building upon the Lagrangian dual (6), by the linearity of the expectation operator, at each time $t \in \{0, 1, \dots\}$, (6) suggests an instantaneous reward function $r_t(z_t, a_t) = r(z_t, a_t) + \lambda_t g(z_t, a_t)$. This function depends upon the Lagrange multiplier λ_t and hence is time-varying. However, infinite-horizon RL algorithms typically assume time-invariant reward functions. We next show that, under Assumption 1, a set of optimal Lagrange multipliers $\{\lambda_t^*\}$ could share the same value. That is, we may presume that all $\{\lambda_t\}_{t=0}^{\infty}$ are equal to some constant λ, and therefore obtain a time-invariant reward function. This time-invariance permits us to apply existing RL algorithms to find the best policy maximizing the time-invariant instantaneous reward function.

Proposition 2. *Let $(\{\lambda_t\}_{t=0}^{\infty}, \pi^*)$ be an optimal solution of* (6)*. Let $(\lambda^*, \tilde{\pi}^*)$ be a pair of optimal solutions to*

$$\min_{\lambda\le 0}\max_{\pi}\ \mathbb{E}\Bigg[\sum_{t=0}^{\infty}\gamma^t\Bigg(r(z_t,a_t)+ \lambda^\top\Bigg(\sum_{t=0}^{\infty}\gamma^t g(z_t,a_t)\Bigg)\Bigg)\Bigg|\pi\Bigg] \quad (7)$$
$$\text{s.t. } z_{t+1}\sim p_z(\cdot|z_t,a_t),\quad \forall t\in\{0,1,\dots\},$$
$$a_t\sim\pi(z_t),\quad \forall t\in\{0,1,\dots\},$$
$$z_0\sim p_0.$$

Let $\tilde{\lambda}_t=\lambda^*$, for all $t\in\{0,1,\dots\}$. Under Assumption 1, we have that $(\{\tilde{\lambda}_t\}_{t=0}^{\infty},\tilde{\pi}^*)$ is also a pair of optimal solution to (6)*.*

Remark 2. *Proposition 2 does not preclude the existence of optimal Lagrange multipliers $\{\lambda_t^*\}_{t=0}^{\infty}$ of* (6) *which are time-varying.*

Building upon the above results and by assuming $\lambda_t=\lambda$, for all $t\ge 0$, we design the time-invariant instantaneous reward

$$\tilde{r}(z_t,a_t):=r(z_t,a_t)+\lambda\cdot\text{Relu}(g(z_t,a_t))-\frac{\rho}{2}\cdot\text{Relu}(g(z_t,a_t))^2, \quad (8)$$

and subsequently we obtain the infinite-horizon objective function

$$R(\pi,\lambda,\rho):=\mathbb{E}\Bigg[\sum_{t=0}^{\infty}\gamma^t\big(r(z_t,a_t)+\lambda\cdot\text{Relu}(g(z_t,a_t)) -\frac{\rho}{2}\cdot\text{Relu}(g(z_t,a_t))^2\big)\Bigg|\pi\Bigg]. \quad (9)$$

Remark 3. *We note that, under Assumption 1, Problem* (5) *could be equivalently considered as a special subclass of CMDPs in which the cumulative constraints could approximate instantaneous constraints. Thus,* (9) *can be interpreted as a new surrogate function for this class of CMDPs. We will show in Section IV that by optimizing* (9)*, we can find a high-quality policy within fewer iterations and smaller constraint violation throughout learning than a currently-used primal-dual method. That is, for this special subclass of CMDPs,* (9) *serves as an alternative surrogate function with a superior empirical performance than the existing primal-dual method.*

Notice that $R(\pi,\lambda,0)$ is not equivalent to the objective function in (7), $\mathbb{E}[\sum_{t=0}^{\infty}\gamma^t(r(z_t,a_t)+\lambda g(z_t,a_t))|\pi]$, because the constraint $\mathbb{E}[\text{Relu}(g(z_t,a_t))|\pi]\le 0$, is a sufficient but not necessary condition for the constraint $\mathbb{E}[g(z_t,a_t)|\pi]\le 0$.

We will show in Section IV that under certain conditions, as $\rho\to\infty$, any infeasible policy would become sub-optimal when we maximize the function $R(\pi,\lambda,\rho)$, with λ fixed. Thus, an optimal policy returned by optimizing (9) for both π and λ would eventually become safe and optimal as we increase ρ. In addition, we remark here that the introduction of the Relu function or other non-negative activation functions in (9) is necessary because otherwise, it is not generally true that an optimal policy for (9) is also optimal for (5), due to the fact that under an optimal policy π^* of problem (5), $\mathbb{E}\big[g(z_t,a_t)^2\big|\pi^*\big]$ may be nonzero and therefore $R(\pi,\lambda,\rho)\to-\infty$, as $\rho\to\infty$.

Algorithm 1: Augmented Lagrangian RL

Pick $c_\rho\in[1,\infty)$, and dual ascent stepsize $\ell\in\mathbb{R}_+$;
Initialize $\rho^{(0)}\in\mathbb{R}_+$, $\lambda^{(0)}=0$;
Randomly initialize the policy π_0;
for $k=0,1,2,\dots K$ **do**
  $\pi_k=\arg\max_{\pi\in\mathcal{P}}R(\lambda^{(k)},\rho^{(k)},\pi)$;
  $\lambda^{(k+1)}\leftarrow \Big\lfloor\lambda^{(k)}-\ell\Big(\mathbb{E}\big[\sum_{t=0}^{\infty}\gamma^t\text{Relu}(g(z_t,a_t))|\pi_k\big]\Big)\Big\rfloor_-$
  $\rho^{(k+1)}\leftarrow c_\rho\rho^{(k)}$;
end

Following the same spirit of the primal-dual algorithm in constrained optimization [13, 15, 23], we propose Algorithm 1.

In Algorithm 1, we initialize $\lambda^{(0)}=0$ and $\rho^{(0)}\in\mathbb{R}_+$, where $\mathbb{R}_+$ denotes the set of non-negative real numbers. At the k-th iteration, we first find a policy $\pi_k\in\arg\max_\pi R(\lambda^{(k)},\rho^{(k)},\pi)$, which could be done by any unconstrained RL algorithm in the literature (e.g., SAC [30], DDPG [31], TRPO [32]). Then, we update the Lagrange multiplier by dual ascent $\lambda^{(k+1)}=\lfloor\lambda^{(k)}-\ell(\mathbb{E}[\sum_{t=0}^{\infty}\gamma^t\text{Relu}(g(z_t,a_t))|\pi_k])\rfloor_-$, where the function $\lfloor\cdot\rfloor_-:\mathbb{R}\to\mathbb{R}_-$ is defined as follows:

$$\lfloor x\rfloor_-=\begin{cases}0 & \text{if } x>0,\\ x & \text{otherwise.}\end{cases} \quad (10)$$

We also update $\rho^{(k+1)}=c_\rho\rho^{(k)}$, where $c_\rho\in[1,\infty)$ is the increasing rate of the quadratic penalty coefficient $\rho^{(k)}$ as the iteration index k grows.

IV. Convergence analysis

In this section, we show that under certain conditions on the feasible policy set, by optimizing the surrogate function (9) we recover an optimal policy for (5).

Proposition 3. *Under Assumption 1, consider the primal maximization of* (7)*, denoted by $d_\rho(\lambda):\mathbb{R}\to\mathbb{R}$ and defined as*

$$d_\rho(\lambda):=\max_{\pi}\mathbb{E}\Bigg[\sum_{t=0}^{\infty}\gamma^t r(z_t,a_t)+\lambda\cdot \textit{Relu}(g(z_t,a_t)) -\frac{\rho}{2}\cdot \textit{Relu}(g(z_t,a_t))^2\Big|\pi\Bigg] \quad (11)$$
$$\text{s.t. } z_{t+1}\sim p_z(\cdot|z_t,a_t),\quad \forall t\in\{0,1,\dots\},$$
$$a_t\sim\pi(z_t),\quad \forall t\in\{0,1,\dots\},$$
$$z_0\sim p_0.$$

Let $\lambda_\rho^*:=\arg\min_{\lambda\le 0}d_\rho(\lambda)$. Suppose that under an optimal policy π^* of problem (5)*, $g(z_t,a_t)\le 0$, $\forall t\in\{0,1,\dots\}$. We*

define a policy $\pi^*_\rho(\lambda)$ *as*

$$\begin{aligned}\pi^*_\rho(\lambda) := \arg\max_{\pi}\ & \mathbb{E}\Bigg[\sum_{t=0}^{\infty}\gamma^t(r(z_t,a_t)+\lambda\cdot \textit{Relu}(g(z_t,a_t)) \\ & -\frac{\rho}{2}\cdot \textit{Relu}(g(z_t,a_t))^2\Big)\Big|\pi\Bigg] \\ \textit{s.t. } & z_{t+1}\sim p_z(\cdot|z_t,a_t),\quad \forall t\in\{0,1,\dots\}, \\ & a_t\sim\pi(z_t),\quad \forall t\in\{0,1,\dots\}, \\ & z_0\sim p_0.\end{aligned} \tag{12}$$

Then, as $\rho\to\infty$*, we have,*

$$\begin{aligned}\Bigg\|\mathbb{E}\Bigg[\sum_{t=0}^{\infty}\gamma^t r(z_t,a_t)\Big|\pi^*\Bigg] \\ -\mathbb{E}\Bigg[\sum_{t=0}^{\infty}\gamma^t r(z_t,a_t)\Big|\pi^*_\rho(\lambda^*_\rho)\Bigg]\Bigg\|_2\to 0.\end{aligned} \tag{13}$$

The condition that under an optimal policy π^* in the original problem (5), $g(z_t,a_t)\le 0, \forall t\in\{0,1,\dots\}$, is equivalent to the condition $\mathbb{E}[g(z_t,a_t)|\pi^*]\le 0, \forall t\in\{0,1,\dots\}$, if we have a deterministic dynamical system. In addition, this condition could be easy to meet for safety-critical systems, due to the fact in many control applications we have a safe but sub-optimal base controller, e.g., Autopilot [33], safe robot-human interaction [34], autonomous driving [35].

We remark here that the analysis in Proposition 3 is conservative. However, in Section V we consider instantaneous constraints g which we do not know a priori are deterministically satisfiable for each t. That is, we consider g for which there may not be a policy π for which $g(z_t,a_t)\le 0,\forall t$. Still, our empirical results suggest that when the parameter ρ is sufficiently large, Algorithm 1 returns a high-quality safe policy.

V. Experiments

In this section, we validate Algorithm 1 in experiments with different initial values of ρ in the settings of a tabular MDP [36], inverted pendulum [37], and half-cheetah [37]. In all experiments, we assume that the constraints are of the form $\mathbb{E}[\text{Relu}(h(z_t,a_t))|\pi]\le 0$ for some function $h:\mathcal{Z}\times\mathcal{A}\to\mathbb{R}$, and therefore when $\rho_0=0$, $R(\pi,\lambda,0)$ recovers the classical Lagrangian dual method adopted in [12–16].

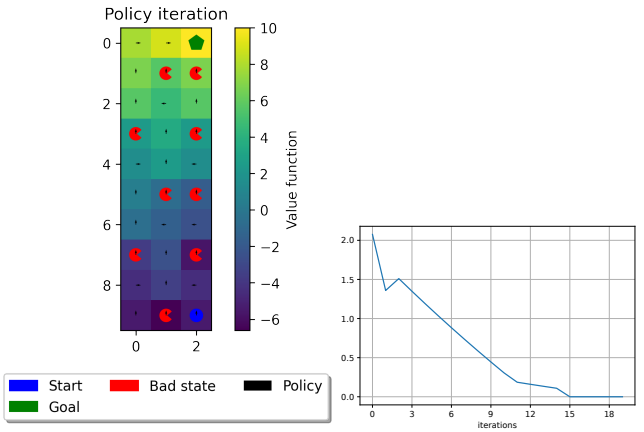
We first consider a constrained tabular MDP in Figure 2a, where we have 10×3 states, each corresponding to a grid cell of a table. The agent starts from an initial state and tries to reach the goal state. At each grid cell, the agent can stay at the same cell or move up, down, left, or right. For those grid cells on the boundary, no action moving out of the table is permitted. The constraint function $g(z_t,a_t)$ takes the value 1 if z_t is considered unsafe and 0 otherwise. The agent receives a reward $r(s,a)=10$ for reaching the goal state (which is terminal) and a reward $r(s,a)=-1$ otherwise. In this experiment, we keep the quadratic penalty coefficient fixed at each iteration in Algorithm 1, and therefore we pick the parameter $c_\rho=1$. At the k-th iteration of Algorithm 1, we apply the classical tabular Policy Iteration [36] to find the policy π_k.

In Figure 2b, we show that the duality gap eventually goes to zero as we update the Lagrange multiplier at each iteration, which empirically validates Proposition 1. In Figure 3a, we observe that as ρ_0 grows, the speed at which the policy returned by Algorithm 1 converges to the optimal policy increases. In Figure 3b, we measure the accumulated constraint $\sum_{t=0}^{\infty}\mathbb{E}[g(z_t,a_t)|\pi]$, and we observe that it decreases as we increase ρ_0. This implies that the surrogate function (9) could promote safety during learning, compared with the case $\rho_0=0$, i.e., the Lagrangian dual approach in [12–16].

Subsequently, we consider a constrained pendulum example, where we add an additional constraint corresponding to avoiding collision with an obstacle near the pendulum, i.e., $\theta_t\notin[\frac{\pi}{2},\pi]$, to the OpenAI Gym "Pendulum-v0" environment [37]. To satisfy Assumption 1, we reformulate this constraint as $\mathbb{E}[g(\theta_t)|\pi]\le 0$, where $g(\theta)=1$ if $\theta\in[\frac{\pi}{2},\pi]$ and $g(\theta)=0$ otherwise. Unlike the previous tabular MDP example where we can find a globally optimal policy, we may only obtain a locally optimal policy due to the non-convexity of RL problems. In line 5 of Algorithm 1, we find a locally optimal policy by running a fixed number of steps of Deterministic Deep Policy Gradient (DDPG) [31]. By picking $c_\rho=1.15$, we slowly increase the quadratic penalty coefficient ρ as the iteration number grows. We update the parameters λ and ρ almost after 6×10^4 steps of DDPG, as indicated by the vertical dashed lines in Figure 4.

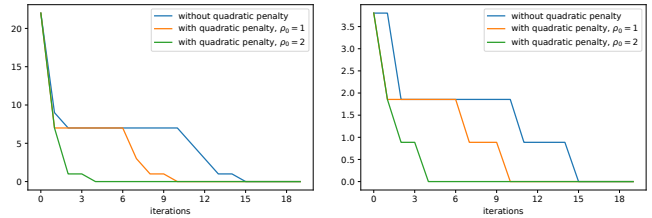
We run experiments with different random seeds and show the average performance and the standard deviation in Figure 4. We see that as ρ_0 increases, the rate at which the policy converges increases and the constraint violations decreases in Figure 4. In particular, we observe that the standard deviation of the constraint violations dramatically decreases as ρ_0 grows in Figure 4b, which suggests that optimizing the surrogate function (9) promotes both safety and stability during learning.

Finally, we consider the constrained half-cheetah example [24], which is adapted from the OpenAI gym "Half-cheetah-v0" environment [37] by adding an additional constraint $|v_x(t)|\le 1$ on the horizontal velocity of the cheetah. We reformulate the constraint as $\mathbb{E}[\text{Relu}(|v_x(t)|-1)|\pi]\le 0$. In line 5 of Algorithm 1, we find a locally optimal policy by running a fixed number of steps of Soft Actor Critic (SAC) [30]. By picking $c_\rho=1.5$, we rapidly increase the quadratic penalty coefficient ρ as the iteration number grows. Similar to the constrained Pendulum experiments, we update λ and ρ every 1.6×10^5 steps of SAC, as indicated by the vertical dashed lines in Figure 5. As we increase ρ_0, we see in Figure 5 that the speed at which the policy converges increases and the constraints violations decreases. However, when ρ_0 is too large, e.g., $\rho_0=2.0$ in Figure (4a), we observe that the exploration is inhibited, which suggests that by tuning ρ_0 we could control the trade-off between exploration and safety. We also plot the state trajectory under polices learned from Algorithm 1 with different values of ρ_0 in Figure 6. We observe that as long as ρ_0 is sufficiently large, the learned



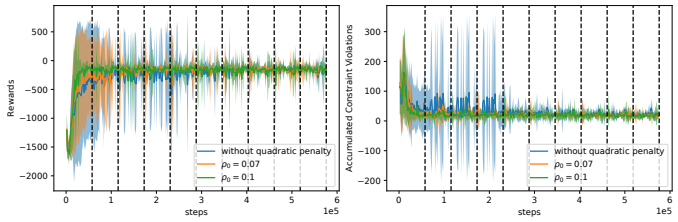
(a) tabular MDP Experiment (b) duality gap

Fig. 2: Tabular MDP Results



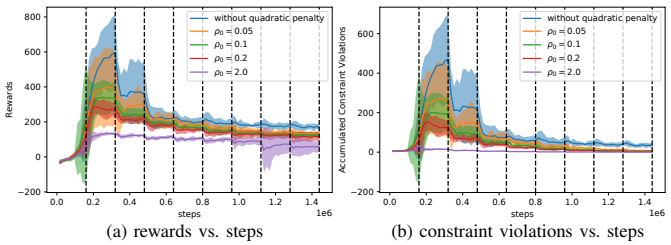
(a) distance between the learned policy and the optimal policy vs. iterations (b) constraint violations vs. iterations

Fig. 3: Tabular MDP Results



(a) rewards vs. steps (b) constraint violations vs. steps

Fig. 4: Constrained pendulum results



(a) rewards vs. steps (b) constraint violations vs. steps

Fig. 5: Constrained half-cheetah results

policies perform safely with a similar performance quality. In addition, when ρ_0 is large, the learned policy becomes conservative possibly due to poor exploration.

VI. Conclusion

In this paper, we considered Instantaneously Constrained RL problems. We first extended a recent result on Cumulatively Constrained RL problems to characterize the strong duality of Instantaneously Constrained RL problems. Inspired by the Augmented Lagrangian method, we proposed a new surrogate function that can promote safety for instantaneous constraints, i.e., reducing the constraint violations during learning. Our surrogate function can be optimized using common unconstrained RL algorithms. We provided theoretical results to justify the use of unconstrained algorithms, which requires stationary Lagrange multipliers to yield time-invariant rewards in the (augmented) Lagrangian. Theoretical results also show that under certain conditions we can recover an optimal policy. Finally, our empirical results suggested that our surrogate function could promote safety during learning. Additionally, we observed that our surrogate function reliably yielded a faster convergence relative to a standard Lagrangian dual approach.

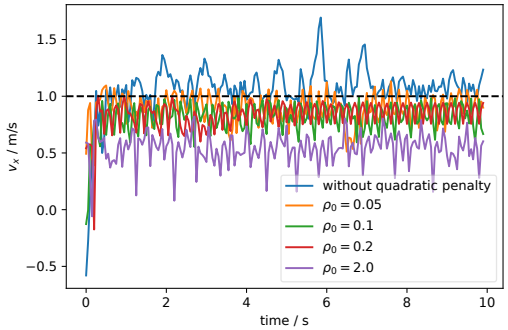


Fig. 6: The state trajectory $v_x(t)$ of constrained half-cheetah under learned policies corresponding to different values of ρ_0. The horizontal black dashed line indicates the constraint $|v_x| \leq 1$.

Appendix

Proposition 4. *Consider a linear system $z_{t+1} = Az_t$, where $A \in \mathbb{R}^{2\times 2}$. Consider a fixed initial condition $z_0 \leq 0$. Suppose that the eigenvalues of A are real and positive, and there exist two eigenvectors $v_1 \leq 0$ and $v_2 \leq 0$ whose convex hull contains z_0. Then, we have $z_t \leq 0$, $\forall t \in \{0, 1, \dots\}$ and $\sum_{t=0}^{\infty} z_t \leq 0$.*

Proof. Denote by λ_i the i-th eigenvalue of A. Let $v_i = [v_{i1}, v_{i2}]$ be an eigenvector associated with λ_i. Consider initial condition $z_0 = [z_{01}, z_{02}] \leq 0$. Construct

$$c_1 = \frac{v_{22}z_{01} - v_{12}z_{02}}{v_{22}v_{11} - v_{12}v_{21}}, c_2 \quad = \frac{v_{11}z_{02} - v_{21}z_{01}}{v_{22}v_{11} - v_{12}v_{21}}. \tag{14}$$

We can verify that $z_0 = c_1v_1 + c_2v_2$. Suppose that there exists a $t' \in \{0, 1, \dots\}$ such that $z_{t'} \leq 0$ is not true. Then, it follows that $z_{t'} = c_1\lambda_1^{t'}v_1 + c_2\lambda_2^{t'}v_2 \leq 0$ is not true, i.e., either $c_1 < 0$ or $c_2 < 0$. However, since z_0 is in the convex hull of v_1 and v_2, we have $\frac{v_{21}}{v_{22}} \leq \frac{z_{01}}{z_{02}} \leq \frac{v_{11}}{v_{12}}$, which yields $v_{22}z_{01} - v_{12}z_{02} \geq 0$ and $v_{11}z_{02} - v_{21}z_{01} \geq 0$. Recall that v_1 and v_2 are in the third quadrant. We have $v_{22}v_{11} - v_{12}v_{21} \geq 0$. This suggests that $c_1 \geq 0$ and $c_2 \geq 0$, which presents a contradiction. Thus, we have $z_t \leq 0$, for all $t \in \{0, 1, \dots\}$, and it also yields $\sum_{t=0}^{\infty} z_t \leq 0$. □

Proof of Proposition 1. By definition, the condition $\mathbb{E}[g(z_t, a_t)|\pi] \leq 0$, $\forall t \in \{0, 1, \dots\}$, implies the condition $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t g(z_t, a_t)|\pi] \leq 0$. By combining the condition in Proposition 1, we have that $\mathbb{E}[g(z_t, a_t)|\pi] \leq 0$, $\forall t \in \{0, 1, \dots\}$, if and only if $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t g(z_t, a_t)|\pi] \leq 0$.

Let π^* be an optimal policy for (5). Define $p^* := \mathbb{E}[\sum_{t=0}^{\infty}\gamma^t r(z_t,a_t)|\pi^*]$. From Theorem 1 in [13], we have

$$\begin{aligned} p^* = \min_{\lambda}\max_{\pi}&\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t(r(z_t,a_t)+\lambda g(z_t,a_t))\middle|\pi\right] \\ \text{s.t. } & z_{t+1}\sim p_z(\cdot|z_t,a_t), \quad \forall t\in\{0,1,\dots\}, \\ & a_t\sim\pi(z_t), \quad t\in\{0,1,\dots\}, \\ & z_0\sim p_0. \end{aligned} \tag{15}$$

By construction, for any non-positive Lagrange Multipliers $\{\lambda_t\}_{t=0}^{\infty}$, we have

$$\begin{aligned} &\max_{\pi}\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t(r(z_t,a_t)+\lambda_t g(z_t,a_t))\middle|\pi\right] \\ \geq&\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t(r(z_t,a_t)+\lambda_t g(z_t,a_t))\middle|\pi^*\right] \\ \geq&\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(z_t,a_t)\middle|\pi^*,z_0\right]+\sum_{t=0}^{\infty}\gamma^t\lambda_t\mathbb{E}[g(z_t,a_t)|\pi^*] \\ \geq&p^* \end{aligned} \tag{16}$$

which also implies

$$\min_{\{\lambda_t\}_{t=0}^{\infty}}\max_{\pi}\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(z_t,a_t)+\lambda_t g(z_t,a_t)\middle|\pi\right]\geq p^*. \tag{17}$$

Let λ^* be an optimal solution of (15). Subsequently, suppose $\tilde{\lambda}_t := \lambda^*$, $\forall t\in\{0,1,\dots\}$. Then, we have

$$\max_{\pi}\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t(r(z_t,a_t)+\tilde{\lambda}_t g(z_t,a_t))\middle|\pi\right]=p^*, \tag{18}$$

which implies that

$$\min_{\{\lambda_t\}_{t=0}^{\infty}}\max_{\pi}\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t(r(z_t,a_t)+\lambda_t g(z_t,a_t))\middle|\pi\right]\leq p^* \tag{19}$$

It follows from (17) and (19) that strong duality holds for (5). □

Proof of Proposition 2. Observe that the problem (7) is the Lagrangian relaxation of

$$\begin{aligned} \max_{\pi}\ &\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(z_t,a_t)\middle|\pi\right] \\ \text{s.t. } & z_{t+1}\sim p_z(\cdot|z_t,a_t), \quad \forall t\in\{0,1,\dots\}, \\ & a_t\sim\pi(z_t), \quad \forall t\in\{0,1,\dots\}, \\ & z_0\sim p_0, \\ & \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t g(z_t,a_t)\middle|\pi\right]\leq 0. \end{aligned} \tag{20}$$

Recall that the condition $\mathbb{E}[g(z_t,a_t)|\pi]\leq 0, \forall t\in\{0,1,\dots\}$, yields $\mathbb{E}[\sum_{t=0}^{\infty}\gamma^t g(z_t,a_t)|\pi]\leq 0$. Suppose that under any closed-loop dynamics $z_{t+1}\sim\mathbb{P}_z(\cdot|z_t,a_t)$ with $a_t\sim\pi(z_t)$, the constraint $\mathbb{E}[\sum_{t=0}^{\infty}\gamma^t g(z_t,a_t)|\pi]\leq 0$ implies that $\mathbb{E}[g(z_t,a_t)|\pi]\leq 0$, $\forall t\in\{0,1,\dots\}$. Then, the problem (20) shares the same feasible domain with problem (6). From Theorem 1 in [13], strong duality holds for (7). By Proposition 1, strong duality also holds for (6). Therefore, a pair of optimal solutions to problem (7) implies that $(\{\tilde{\lambda}_t\}_{t=0}^{\infty},\tilde{\pi}^*)$ is a pair of optimal solutions to (6). □

Before we present the proof of Proposition 3, we first introduce the following Lemma, which builds the foundation for the proof of Proposition 3.

Lemma 1. *Under Assumption 1, consider the function $d(\lambda):\mathbb{R}_-\to\mathbb{R}$ defined as*

$$\begin{aligned} d(\lambda) := \max_{\pi}&\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t(r(z_t,a_t)+\lambda(g(z_t,a_t))\middle|\pi\right] \\ \textit{s.t. } & z_{t+1}\sim p_z(\cdot|z_t,a_t), \quad \forall t\in\{0,1,\dots\}, \\ & a_t\sim\pi(z_t), \quad \forall t\in\{0,1,\dots\}, \\ & z_0\sim p_0, \end{aligned} \tag{21}$$

and the function $d_\rho(\lambda):\mathbb{R}_-\to\mathbb{R}$ defined as

$$\begin{aligned} d_\rho(\lambda) := \max_{\pi}&\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(z_t,a_t)+\lambda\cdot\textit{Relu}(g(z_t,a_t))\right. \\ &\left.-\frac{\rho}{2}\cdot\textit{Relu}(g(z_t,a_t))^2\middle|\pi\right] \\ \textit{s.t. } & z_{t+1}\sim p_z(\cdot|z_t,a_t), \quad \forall t\in\{0,1,\dots\}, \\ & a_t\sim\pi(z_t), \quad \forall t\in\{0,1,\dots\}, \\ & z_0\sim p_0. \end{aligned} \tag{22}$$

Let $\lambda^* := \arg\min_{\lambda\leq 0} d(\lambda)$ and $\lambda_\rho^* := \arg\min_{\lambda\leq 0} d_\rho(\lambda)$. Suppose that under an optimal policy π^*, $g(z_t,a_t)\leq 0$ for all $t\geq 0$ under an optimal policy π^*. Then, we have

$$d(\lambda^*)\leq d_\rho(\lambda_\rho^*)\leq d_\rho(\lambda^*)\leq d(\lambda^*). \tag{23}$$

Proof. On one hand, we observe that for any $\rho\geq 0$, $d_\rho(\lambda)\leq d_{\rho=0}(\lambda)=d(\lambda)$, and therefore, $d_\rho(\lambda^*)\leq d(\lambda^*)$. By definition, $d_\rho(\lambda_\rho^*)\leq d_\rho(\lambda^*)$. Moreover, observe that π^* is a feasible solution to problem (22), and under the policy π^*,

$$\begin{aligned} d_\rho(\lambda_\rho^*)\geq&\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t(r(z_t,a_t)+\lambda^*\cdot\text{Relu}(g(z_t,a_t))\right. \\ &\left.-\frac{\rho}{2}\cdot\text{Relu}(g(z_t,a_t))^2\right)\middle|\pi^*\Big] \\ =&\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(z_t,a_t)\middle|\pi^*\right] \\ =&d(\lambda). \end{aligned} \tag{24}$$

Thus, $d(\lambda^*)\leq d_\rho(\lambda_\rho^*)\leq d_\rho(\lambda^*)\leq d(\lambda^*)$. □

Proof of Proposition 3. We aim to show that as $\rho\to\infty$, any infeasible policy π' will become suboptimal for problem (12). Given $\rho\geq 0$, suppose that $\pi_\rho^*(\lambda_\rho^*)$ is infeasible. Then,

$$\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t g(z_t,a_t)\middle|\pi_\rho^*(\lambda_\rho^*)\right]>0, \tag{25}$$

because otherwise $\pi_\rho^*(\lambda_\rho^*)$ would be feasible.

Define the function

$$J(\pi):=\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(z_t,a_t)\middle|\pi\right]. \tag{26}$$

There are only two cases: either $\lambda_\rho^* < \lambda^*$ or $\lambda_\rho^* \geq \lambda^*$.

For the first case that $\lambda_\rho^* < \lambda^*$, since $\pi_\rho^*(\lambda_\rho^*)$ is an optimal policy in problem (12), we have

$$\begin{aligned} d(\lambda^*) \geq & \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t(r(z_t,a_t)+\lambda^* g(z_t,a_t))\middle|\pi_\rho^*(\lambda_\rho^*)\right] \\ = & J(\pi_\rho^*(\lambda_\rho^*)) + \lambda^*\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t g(z_t,a_t)\middle|\pi_\rho^*(\lambda_\rho^*)\right] \\ > & J(\pi_\rho^*(\lambda_\rho^*)) + \lambda_\rho^*\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t g(z_t,a_t)\middle|\pi_\rho^*(\lambda_\rho^*)\right] \\ > & d_\rho(\lambda_\rho^*) \end{aligned} \tag{27}$$

which contradicts that $d(\lambda^*) = d_\rho(\lambda_\rho^*)$, as shown in Lemma 1.

For the second case that $\lambda_\rho^* \geq \lambda^*$, we can pick $\rho' \geq 0$ such that

$$J(\pi_\rho^*(\lambda_\rho^*)) - \frac{\rho'}{2}\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t \text{Relu}(g(z_t,a_t))^2\middle|\pi_\rho^*(\lambda_\rho^*)\right] < J(\pi^*). \tag{28}$$

Subsequently, we have

$$\begin{aligned} d_{\rho'}(\lambda_{\rho'}^*) = & J(\pi^*) \\ > & J(\pi_\rho^*(\lambda_\rho^*)) - \frac{\rho'}{2}\mathbb{E}\big[\text{Relu}(g(z_t,a_t))^2\big|\pi_\rho^*(\lambda_\rho^*)\big] \\ \geq & J(\pi_\rho^*(\lambda_\rho^*)) - \lambda_{\rho'}^*\mathbb{E}\left[\sum_{t=0}^{\infty}\text{Relu}(\gamma^t g(z_t,a_t))\middle|\pi_\rho^*(\lambda_\rho^*)\right] \\ & - \frac{\rho'}{2}\mathbb{E}\big[\text{Relu}(g(z_t,a_t))^2\big|\pi_\rho^*(\lambda_\rho^*)\big] \end{aligned} \tag{29}$$

which implies that $\pi_\rho^*(\lambda_\rho^*)$ becomes a sub-optimal solution in problem (12), as ρ increases to ρ'.

As $\rho \to \infty$, any infeasible policy π' will become sub-optimal for problem (12). Recall that π^* is an optimal policy. Therefore, as $\rho \to \infty$, it holds that

$$\left\|\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(z_t,a_t)\middle|\pi^*\right] - \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(z_t,a_t)\middle|\pi_\rho^*(\lambda_\rho^*)\right]\right\|_2 \to 0. \tag{30}$$

□

REFERENCES

[1] V. Mnih et al. "Human-level control through deep reinforcement learning". *nature* 518.7540 (2015).

[2] S. Levine et al. "End-to-end training of deep visuomotor policies". *The Journal of Machine Learning Research* 17.1 (2016).

[3] S. Gu et al. "Continuous deep Q-learning with model-based acceleration". *International Conference on Machine Learning*. 2016.

[4] D. Silver, R. S. Sutton, and M. Müller. "Reinforcement learning of local shape in the game of Go." *IJCAI*. Vol. 7. 2007.

[5] Y. Duan et al. "Benchmarking deep reinforcement learning for continuous control". *International conference on machine learning*. PMLR. 2016.

[6] C. Wu et al. "Flow: Architecture and benchmarking for reinforcement learning in traffic control". *arXiv preprint arXiv:1710.05465* (2017).

[7] P. Auer. "Using confidence bounds for exploitation-exploration tradeoffs". *Journal of Machine Learning Research* 3.Nov (2002).

[8] E. Hazan et al. "Provably efficient maximum entropy exploration". *International Conference on Machine Learning*. PMLR. 2019.

[9] P.-A. Andersen, M. Goodwin, and O.-C. Granmo. "The dreaming variational autoencoder for reinforcement learning environments". *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer. 2018.

[10] S. Shalev-Shwartz, S. Shammah, and A. Shashua. "Safe, multi-agent, reinforcement learning for autonomous driving". *arXiv preprint arXiv:1610.03295* (2016).

[11] A. Esteva et al. "A guide to deep learning in healthcare". *Nature medicine* 25.1 (2019).

[12] E. Altman. *Constrained Markov decision processes*. Vol. 7. CRC Press, 1999.

[13] S. Paternain et al. "Constrained reinforcement learning has zero duality gap". *Advances in Neural Information Processing Systems*. 2019.

[14] J. Achiam et al. "Constrained policy optimization". *International Conference on Machine Learning*. 2017.

[15] Y. Chow et al. "Risk-constrained reinforcement learning with percentile risk criteria". *The Journal of Machine Learning Research* 18.1 (2017).

[16] Y. Chow et al. "A Lyapunov-based approach to safe reinforcement learning". *Advances in neural information processing systems*. 2018.

[17] J. F. Fisac et al. "Bridging Hamilton-Jacobi safety analysis and reinforcement learning". *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019.

[18] S. Gu et al. "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates". *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2017.

[19] T.-H. Pham, G. De Magistris, and R. Tachibana. "Optlayer-practical constrained optimization for deep reinforcement learning in the real world". *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018.

[20] S. Gros, M. Zanon, and A. Bemporad. "Safe reinforcement learning via projection on a safe set: How to achieve optimality?" *arXiv preprint arXiv:2004.00915* (2020).

[21] G. Kalweit et al. *Deep constrained Q-learning*. 2020.

[22] O. Bastani, Y. Pu, and A. Solar-Lezama. "Verifiable reinforcement learning via policy extraction". *Advances in neural information processing systems*. 2018.

[23] J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[24] Y. Liu, J. Ding, and X. Liu. "IPO: Interior-point policy optimization under constraints". *arXiv preprint arXiv:1910.09615* (2019).

[25] E. Altman. "Constrained Markov decision processes with total cost criteria: Lagrangian approach and dual linear program". *Mathematical methods of operations research* 48.3 (1998).

[26] S. Karaman and E. Frazzoli. "Sampling-based algorithms for optimal motion planning". *The international journal of robotics research* 30.7 (2011).

[27] C. Liu and M. Tomizuka. "Designing the robot behavior for safe human–robot interactions". *Trends in Control and Decision-Making for Human–Robot Collaboration Systems* (2017).

[28] B. Acikmese and S. R. Ploen. "Convex programming approach to powered descent guidance for mars landing". *Journal of Guidance, Control, and Dynamics* 30.5 (2007).

[29] P. Ramachandran, B. Zoph, and Q. V. Le. "Searching for activation functions". *arXiv preprint arXiv:1710.05941* (2017).

[30] T. Haarnoja et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor". *International Conference on Machine Learning*. PMLR. 2018.

[31] T. P. Lillicrap et al. "Continuous control with deep reinforcement learning". *arXiv preprint arXiv:1509.02971* (2015).

[32] J. Schulman et al. "Trust region policy optimization". *International Conference on Machine Learning*. 2015.

[33] P. A. Scholten et al. "Variable stability in-flight simulation system based on existing autopilot hardware". *Journal of Guidance, Control, and Dynamics* 43.12 (2020).

[34] B. Navarro et al. "An ISO10218-compliant adaptive damping controller for safe physical human-robot interaction". *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2016.

[35] D. Graves, K. Rezaee, and S. Scheideman. "Perception as prediction using general value functions in autonomous driving applications". *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2019.

[36] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. 2018.

[37] G. Brockman et al. "Openai gym". *arXiv preprint arXiv:1606.01540* (2016).