2022 American Control Conference (ACC)
Atlanta, USA, June 8-10, 2022

Convergence and optimality of policy gradient primal-dual method for constrained Markov decision processes

Dongsheng Ding, Kaiqing Zhang, Tamer Başar, and Mihailo R. Jovanović

Abstract— **We study constrained Markov decision processes with finite state and action spaces. The optimal solution of a discounted infinite-horizon optimal control problem is obtained using a Policy Gradient Primal-Dual (PG-PD) method without any policy parametrization. This method updates the primal variable via projected policy gradient ascent and the dual variable via projected sub-gradient descent. Despite the lack of concavity of the constrained maximization problem in policy space, we exploit the underlying structure to provide non-asymptotic global convergence guarantees with sublinear rates in terms of both the optimality gap and the constraint violation. Furthermore, for a sample-based PG-PD algorithm, we quantify sample complexity and offer computational experiments to demonstrate the effectiveness of our results.**

I. Introduction

Constrained Markov decision processes (CMDPs) [1] have been widely used in the reinforcement learning (RL) literature [2], [3] as a class of real-world environment models since they are suitable for many constraint-rich domains, e.g., autonomous driving [4], medicine test [5], and financial management [6]. Policy gradient (PG) methods that directly search policies via gradient ascent descent or primal dual methods lie at the heart of recent model-free algorithms for solving CMDPs [7]–[12]. Reasonably more appealing are their inherited versatility from PG methods [13] and universality in adopting Lagrange method to deal with constraints [7], [8], making them powerful in real-world RL.

As an important and useful extension of PG methods to incorporating constraints into policy search, policy gradient primal-dual (PG-PD) methods [8], [14]–[17] offer to use policy gradient and constraint violation as ascent descent directions to solve Lagrangian-based max-min problems [1]. PG-PD methods seek to find a saddle point of the Lagrangian for CMDPs and fall under the class of non-convex minimax optimization methods that are challenging to show global convergence [18], [19]. PG-PD methods also connect to recent RL algorithms for CMDPs, e.g., constrained policy optimization [9] and optimistic primal-dual methods [20], [21]. However, the finite-time or non-asymptotic global convergence of PG-PD methods is unknown for the tabular CMDPs due to lack of structural properties from policy parameterizations.

Our contribution: In this work, we focus on the first-order policy gradient method for obtaining the optimal solution to the discounted infinite-horizon constrained Markov decision processes with finite state and action spaces. We study a Policy Gradient Primal-Dual (PG-PD) method without any policy parametrization; the primal variable is updated via projected policy gradient ascent and the dual variable via projected sub-gradient descent. Despite non-concavity of the constrained maximization problem in policy space, we exploit problem structure to prove that PG-PD achieves non-asymptotic global convergence with sublinear rates in terms of both the optimality gap and the constraint violation. Furthermore, we present a sample-based PG-PD algorithm and quantify its sample complexity. We provide computational experiments to demonstrate the effectiveness of these developments. To the best of our knowledge, our work is the first to provide non-asymptotic convergence guarantees for policy gradient primal-dual methods in the context of solving discounted infinite-horizon CMDPs without any policy parameterizations or preconditioning.

We remark that the PG-PD method shares similarities with the natural policy gradient primal-dual method (NPG-PD) [11], but they are different in method and analysis. PG-PD directly works with policy probability simplex instead of softmax policy in [11] and it uses policy gradient without Fisher preconditioning as a search direction. Hence, structural results established in [11] for NPG-PD do not hold here. In contrast, our convergence analysis exploits the underlying convexity of RL that departs from the structure used in [11].

Related work: Our work is pertinent to the Lagrangian method for discounted infinite-horizon CMDPs [1]. This method casts the CMDP problem into a max-min saddle-point problem and searches for optimal policies in the primal-dual domain. Similar algorithms include projected PG [17], trajectory-based PG [7], RCPO [8], APDO [22], and advanced AC algorithms [7], [12], [16], [23]. One exception is dualDescent [24] that works in the dual domain. Typically, these algorithms converge to a local stationary point asymptotically. Despite the non-convexity of the max-min problem [11], [13], several recent references show that a more favorable global convergence is achievable. In [11], we incorporated natural PG into a primal-dual algorithm and demonstrated global convergence under strong duality [24].

Financial support from the National Science Foundation under awards ECCS-1708906 and ECCS-1809833, and from the Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196 is gratefully acknowledged.

D. Ding and M. R. Jovanović are with the Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089. K. Zhang is with LIDS and CSAIL, Massachusetts Institute of Technology, Cambridge, MA 02139. T. Başar is with the Department of Electrical and Computer Engineering and Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801. E-mails: dongshed@usc.edu, kaiqing@mit.edu, basar1@illinois.edu, mihailo@usc.edu.

978-1-6654-5196-3/$31.00 ©2022 AACC

In [25], the authors prove that the primal-dual algorithm is globally convergent when KL-regularized policy iteration is utilized as the primal update. In [26], the authors propose a globally convergent primal method with a feasibility correction. All of these global convergence results utilize either a particular softmax policy parametrization [11], [25], [26] or preconditioning [11]. However, a more basic theoretical question is whether and how fast the first-order primal-dual methods converge to a globally optimal solution. In this work, we take an initial step towards answering this question.

Paper organization: The rest of the paper is organized as follows. In Section II, we provide background material on constrained Markov decision processes. In Section III, we propose a policy gradient primal-dual method and establish non-asymptotic convergence guarantees. In Section IV, we present a sample-based policy gradient primal-dual algorithm, quantify its sample complexity, and use computational experiments to illustrate our results. We close the paper with concluding remarks in Section V and relegate technical details to appendices.

II. Constrained Markov Decision Processes

A discounted constrained Markov decision process is specified by CMDP$(S, A, P, r, g, b, \gamma, \rho)$, where S and A are finite state and action spaces, P is a transition probability measure which specifies the transition probability $P(s' \,|\, s, a)$ from state s to the next state s' under action $a \in A$, r: $S \times A \to [0,1]$ is a reward function, g: $S \times A \to [0,1]$ is a utility function, b is a constraint offset, $\gamma \in [0,1)$ is a discount factor, and ρ is an initial state distribution over S.

A stochastic policy of an agent is a function, π: $S \to \Delta_A$, which determines a probability simplex Δ_A over action space A chosen by the agent based on the current state, e.g., $a_t \sim \pi(\cdot \,|\, s_t)$ at time t. Let the set of all possible policies be Π. A policy $\pi \in \Pi$, together with initial state distribution ρ, induces a distribution over trajectories $\tau := \{(s_t, a_t, r_t, g_t)\}_{t=0}^{\infty}$, where $s_0 \sim \rho$, $a_t \sim \pi(\cdot \,|\, s_t)$ and $s_{t+1} \sim P(\cdot \,|\, s_t, a_t)$ for all $t \geq 0$.

Given a policy π, the value functions V_r^π, V_g^π: $S \to \mathbb{R}$, associated with reward r or utility g are expected values of total discounted rewards or utilities received under policy π,

$$\begin{aligned} V_r^\pi(s) &:= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\Big|\, \pi, s_0 = s\right] \\ V_g^\pi(s) &:= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t g(s_t, a_t) \,\Big|\, \pi, s_0 = s\right] \end{aligned}$$

where $\mathbb{E}$ is taken over trajectory τ induced by π. We further introduce the state-action value functions $Q_r^\pi(s,a)$, $Q_g^\pi(s,a)$: $S \times A \to \mathbb{R}$ when agent starts from an arbitrary state-action pair (s,a) and follows policy π, together with their advantage functions A_r^π, A_g^π: $S \times A \to \mathbb{R}$,

$$\begin{aligned} Q_\diamond^\pi(s,a) &:= \mathbb{E}\left[\sum_{t=0}^{\infty} \diamond(s_t, a_t) \,\Big|\, \pi, s_0 = s, a_0 = a\right] \\ A_\diamond^\pi &:= Q_\diamond^\pi(s,a) - V_\diamond^\pi(s), \end{aligned}$$

where symbol $\diamond$ stands for r or g. Since $r, g \in [0,1]$, it is easy to see that $0 \leq V_\diamond^\pi(s) \leq 1/(1-\gamma)$. Their expected values under ρ are: $V_\diamond^\pi(\rho) := \mathbb{E}_{s_0 \sim \rho}[V_\diamond^\pi(s_0)]$. The agent's goal is to find a policy $\pi^\star$ that maximizes the expected reward value subject to a constraint on the expected utility value,

$$\underset{\pi \in \Pi}{\text{maximize}} \; V_r^\pi(\rho) \;\; \text{subject to} \;\; V_g^\pi(\rho) \geq b. \tag{1}$$

Maximization is over all policies and, to avoid trivial cases, the constraint offset is set to $b \in (0, 1/(1-\gamma)]$. We note that maximization problem (1) is not concave [11].

Let $V_L^{\pi,\lambda}(\rho) := V_r^\pi(\rho) + \lambda(V_g^\pi(\rho) - b)$ be the Lagrangian. We can cast problem (1) as a max-min problem,

$$\underset{\pi \in \Pi}{\text{maximize}} \;\; \underset{\lambda \geq 0}{\text{minimize}} \;\; V_L^{\pi,\lambda}(\rho) \tag{2}$$

where π is the primal variable and λ is the nonnegative Lagrange multiplier or dual variable. The associated dual function is defined as $V_D^\lambda(\rho) := \text{maximize}_\pi \, V_L^{\pi,\lambda}(\rho)$. We assume that problem (1) is strictly feasible.

Assumption 1 (Strict Feasibility): There exists $\xi > 0$ and $\bar{\pi}$ such that $V_g^{\bar{\pi}}(\rho) - b \geq \xi$.

Let the optimal dual variable be $\lambda^\star = \arg\min_{\lambda \geq 0} V_D^\lambda(\rho)$. We use shorthand $V_r^{\pi^\star}(\rho) = V_r^\star(\rho)$ and $V_D^{\lambda^\star}(\rho) = V_D^\star(\rho)$.

Despite lack of convexity in (2), the problem structure can be utilized to establish nice properties, e.g., [11, Lemma 2].

III. Algorithm and Convergence

We introduce a policy gradient primal-dual method under the direct policy parametrization in Section III-A and establish its convergence in Sections III-B and III-C.

A. Policy Gradient Primal-Dual (PG-PD) Method

Let $\Theta = \Delta_A^{|S|}$ be the set of all policies and let $\{\pi_\theta = \theta \,|\, \theta \in \Theta\}$ be the policy class without parametrization. We utilize a Policy Gradient Primal-Dual (PG-PD) method to update primal and dual variables θ and λ,

$$\begin{aligned} \theta^{(t+1)} &= \mathcal{P}_\Theta\big(\theta^{(t)} + \eta_1 \nabla_\theta V_L^{\theta^{(t)},\lambda^{(t)}}(\rho)\big) \\ \lambda^{(t+1)} &= \mathcal{P}_\Lambda\big(\lambda^{(t)} - \eta_2 \big(V_g^{\theta^{(t)}}(\rho) - b\big)\big) \end{aligned} \tag{3}$$

where $\nabla_\theta V_L^{\theta^{(t)},\lambda^{(t)}}(\rho) := \nabla_\theta V_r^{\theta^{(t)}}(\rho) + \lambda^{(t)} \nabla_\theta V_g^{\theta^{(t)}}(\rho)$, $\mathcal{P}_\Theta$ is a projection onto the probability simplex, $\mathcal{P}_\Lambda$ is a projection onto the interval Λ, and $\eta_1, \eta_2 > 0$ are stepsizes.

PG-PD utilizes a simple primal-ascent dual-descent structure: the primal update $\theta^{(t+1)}$ performs projected gradient ascent using the policy gradient $\nabla_\theta V_L^{(t)}(\rho)$ and the dual update $\lambda^{(t+1)}$ runs projected sub-gradient descent by collecting constraint violation $b - V_g^{(t)}(\rho)$. We use shorthand $V_g^{(t)}(\rho)$ for $V_g^{\theta^{(t)}}(\rho)$; similar for other values.

PG-PD is a basic policy search method [8], [14], [15] that differs from the natural policy gradient primal-dual method (NPG-PD) [11]. PG-PD directly works with the policy, and the policy gradient is absent of Fisher preconditioning. Thus, policy-induced structural result established in [11] for NPG-PD does not hold here: multiplicative weight update used in [11] cannot be employed to update the policy in PG-PD.

In what follows, we exploit the underlying convexity to prove the convergence for PG-PD.

B. Convergence and Optimality

We establish convergence of PG-PD for particular projection interval Λ, initial policy, and stepsizes, in Theorem 1.

Theorem 1 (Sublinear Rate): Let Assumption 1 hold and let $\Lambda = [0, 2/((1-\gamma)\xi)]$, $\rho > 0$, $\lambda^{(0)} = 0$, and $\theta^{(0)} \in \Theta$ be such that $V_r^{\theta^{(0)}}(\rho) \geq V_r^\star(\rho)$. For stepsizes

$$\begin{aligned} \eta_1 &= \frac{(1-\gamma)^4}{2|A|(1+2/\xi)} \\ \eta_2 &= \frac{8|A||S|(1+2/\xi)}{(1-\gamma)^4\sqrt{T}} \left\| \frac{d_\rho^{\pi^\star}}{\rho} \right\|_\infty^2 \end{aligned}$$

the iterates $\theta^{(t)}$ generated by PG-PD satisfy

$$\frac{1}{T}\sum_{t=0}^{T-1} \big(V_r^\star(\rho) - V_r^{(t)}(\rho)\big) \leq C_1 \frac{|A||S|\|d_\rho^{\pi^\star}/\rho\|_\infty^2}{(1-\gamma)^6 T^{1/4}} \quad (4a)$$

$$\left[\frac{1}{T}\sum_{t=0}^{T-1} \big(b - V_g^{(t)}(\rho)\big)\right]_+ \leq C_2 \frac{|A||S|\|d_\rho^{\pi^\star}/\rho\|_\infty^2}{(1-\gamma)^6 T^{1/4}} \quad (4b)$$

where C_1 and C_2 are absolute constants.

In Theorem 1, for tabular CMDPs, we show that the average reward value function converges to the optimal one and that the constraint violation decays to zero, both in sublinear rate $T^{1/4}$. Despite lack of policy's convexity, we provide much stronger convergence guarantees than deterministic nonconvex minimax optimization [18]. Even though PG-PD does not have policy-induced structural properties, we show that the proof strategy employed in [11, Theorem 1] works once we exploit the underlying convexity of the problem discussed in Section III-C. Alternatively, with iteration complexity $O(1/\epsilon^4)$, PG-PD yields an ϵ-optimal policy by selecting π^{out} uniformly over iterates from $\pi^{(1)}$ to $\pi^{(T)}$,

$$\mathbb{E}\big[V_r^\star(\rho) - V_r^{\pi^{\text{out}}}(\rho)\big] \leq \epsilon \quad \text{and} \quad \mathbb{E}\big[b - V_g^{\pi^{\text{out}}}(\rho)\big] \leq \epsilon.$$

Compared with the softmax policy result [11], the sublinear rate for optimality gap/constraint violation in Theorem 1 is slightly worse. This is because PG-PD uses gradient ascent/descent without resorting to any policy-induced structural properties and the rate depends on problem dimensions $|S|$ and $|A|$ as well as distribution shift $\|d_\rho^{\pi^\star}/\rho\|_\infty$ that specifies the exploration factor [13].

C. Convexity in Occupancy Measure

The occupancy measure is a useful analytical tool for analyzing CMDPs [1]. An occupancy measure q^π defines a set of probability distributions generated by taking a policy π,

$$q_{s,a}^\pi = \sum_{t=0}^{\infty} \gamma^t P(s_t = s, a_t = a \,|\, \pi, s_0 \sim \rho) \quad (5)$$

for all $s \in S$, $a \in A$. For brevity, we put all $q_{s,a}^\pi$ together as $q^\pi \in \mathbb{R}^{|S||A|}$ and $q_a^\pi = [\, q_{1,a}^\pi, \cdots, q_{|S|,a}^\pi \,]^\top$. For an action a, we collect transition probabilities $P(s' \,|\, s, a)$ for all $s', s \in S$ to denote $P_a \in \mathbb{R}^{|S|\times|S|}$. The occupancy measure q^π is defined in the domain $\mathcal{Q} := \{q^\pi \in \mathbb{R}^{|S||A|} \,|\, \sum_{a\in A}(I - \gamma P_a^\top) q_a^\pi = \rho \text{ and } q^\pi \geq 0\}$, a set of linear constraints.

With a slight abuse of notation, we write $r \in [0,1]^{|S||A|}$ and $g \in [0,1]^{|S||A|}$. Thus, the value functions V_r^π, V_g^π: $S \to \mathbb{R}$ under the initial state distribution ρ are linear functions [1]:

$$\begin{aligned} V_r^\pi(\rho) &= \langle q^\pi, r\rangle := F_r(q^\pi) \\ V_g^\pi(\rho) &= \langle q^\pi, g\rangle := F_g(q^\pi). \end{aligned}$$

We are now in a position to re-write (1) as a linear program,

$$\underset{q^\pi \in \mathcal{Q}}{\text{maximize}} \; F_r(q^\pi) \;\; \text{subject to} \;\; F_g(q^\pi) \geq b \quad (6)$$

where $\mathcal{Q}$ is the domain. Since the transition P is unknown, we can not solve the linear program (6) directly. For any $q^\pi \in \mathcal{Q}$, the associated policy π is given by

$$\pi(a \,|\, s) = \frac{q_{s,a}^\pi}{\sum_{a\in A} q_{s,a}^\pi} \;\; \text{for all } s \in S, a \in A. \quad (7)$$

Abstractly, we denote by π^q: $\mathcal{Q} \to \Delta_A^{|S|}$ a mapping from an occupancy measure q^π to a policy π. Similarly, as defined by (5) we denote by q^π: $\Delta_A^{|S|} \to \mathcal{Q}$ a mapping from a policy π to an occupancy measure q^π. Clearly, $q^\pi = (\pi^q)^{-1}$.

Despite the non-convexity of (1) in policy, its reformulation (6) reveals underlying convexity in occupancy measure q^π. We first exploit this convexity to show the average policy improvement over T steps in Lemma 2; see Appendix A for proof. We use shorthand $q^{(t)}$, $q^\star$ for $q^{\theta^{(t)}}$, $q^{\theta^\star}$, respectively.

Lemma 2 (Bounded Average Performance): Let assumptions in Theorem 1 hold. For stepsizes $\eta_1 = 1/L$ and $\eta_2 = (1-\gamma)^2 DL/(2\sqrt{T})$, the iterates $\pi^{(t)} = \theta^{(t)}$ generated by PG-PD satisfy

$$\frac{1}{T}\sum_{t=0}^{T-1} Z^{(t)} \leq \frac{DL}{T^{1/4}} \quad (8)$$

where $Z^{(t)} = F_r(q^\star) - F_r(q^{(t)}) + \lambda^{(t)}\big(F_g(q^\star) - F_g(q^{(t)})\big)$,

$$D := \frac{8|S|}{(1-\gamma)^2} \left\| \frac{d_\rho^{\pi^\star}}{\rho} \right\|_\infty^2 \;\; \text{and} \;\; L := \frac{2|A|(1+2/\xi)}{(1-\gamma)^4}.$$

The proof of Lemma 2 differs from that of proving [11, Lemma 7]. Since PG-PD's policy update is not in multiplicative form, it is not possible to apply performance difference lemma as in [11]. Instead, we next begin with standard descent lemma and exploit the convexity in occupancy measure.

Proof of Theorem 1: We first bound the optimality gap. By the equality $(\lambda^{(T)})^2 = \sum_{t=0}^{T-1}\big((\lambda^{(t+1)})^2 - (\lambda^{(t)})^2\big)$ and the dual update (3), we can bound $(\lambda^{(T)})^2$ by

$$\begin{aligned} &2\eta_2 \sum_{t=0}^{T-1} \lambda^{(t)}\big(b - F_g(q^{(t)})\big) + \eta_2^2 \sum_{t=0}^{T-1} \big(F_g(q^{(t)}) - b\big)^2 \\ &\leq 2\eta_2 \sum_{t=0}^{T-1} \lambda^{(t)}\big(F_g(q^\star) - F_g(q^{(t)})\big) + \frac{\eta_2^2 T}{(1-\gamma)^2} \end{aligned}$$

where the inequality is due to the feasibility of $q^\star$: $F_g(q^\star) \geq b$, and $|F_g(q^{(t)}) - b| \leq \frac{1}{1-\gamma}$. Hence,

$$-\frac{1}{T}\sum_{t=0}^{T-1} \lambda^{(t)}\big(F_g(q^\star) - F_g(q^{(t)})\big) \;\leq\; \frac{\eta_2}{2(1-\gamma)^2}. \qquad (9)$$

By Lemma 2, substituting (9) into (8) leads to the first bound, where we take $\eta_2 = \frac{(1-\gamma)^2 DL}{2\sqrt{T}}$.

We next bound the constraint violation. By the dual update in (3), for any $\lambda \in [0, 2/((1-\gamma)\xi)]$,

$$\begin{aligned}&|\lambda^{(t+1)} - \lambda|^2\\ &\leq\; \big|\lambda^{(t)} - \lambda\big|^2 - 2\eta_2\big(F_g(q^{(t)}) - b\big)\big(\lambda^{(t)} - \lambda\big) + \frac{\eta_2^2}{(1-\gamma)^2}\end{aligned}$$

where the inequality is due to the non-expansiveness of projection operator $\mathcal{P}_\Lambda$ and $(F_g(q^{(t)}) - b)^2 \leq \frac{1}{(1-\gamma)^2}$. Summing it up from $t = 0$ to $t = T-1$, and dividing by T, yield

$$\begin{aligned}&\frac{1}{T}|\lambda^{(T)} - \lambda|^2 - \frac{1}{T}|\lambda^{(0)} - \lambda|^2\\ &\leq\; -\frac{2\eta_2}{T}\sum_{t=0}^{T-1}\big(F_g(q^{(t)}) - b\big)\big(\lambda^{(t)} - \lambda\big) + \frac{\eta_2^2}{(1-\gamma)^2}\end{aligned}$$

which by some rearrangement further implies that

$$\frac{1}{T}\sum_{t=0}^{T-1}\big(F_g(q^{(t)}) - b\big)\big(\lambda^{(t)} - \lambda\big) \;\leq\; \frac{|\lambda^{(0)} - \lambda|^2}{2\eta_2 T} + \frac{\eta_2}{2(1-\gamma)^2}.$$

We note that $F_g(q^{\theta^\star}) \geq b$. By adding the inequality above to (8) in Lemma 2 from both sides,

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\big(F_r(q^{\theta^\star}) - F_r(q^{(t)})\big) \;+\; \frac{\lambda}{T}\sum_{t=0}^{T-1}\big(b - F_g(q^{(t)})\big)\\ &\leq\; \frac{DL}{T^{1/4}} \;+\; \frac{|\lambda^{(0)} - \lambda|^2}{2\eta_2 T} \;+\; \frac{\eta_2}{2(1-\gamma)^2}.\end{aligned} \qquad (10)$$

We choose $\lambda = \frac{2}{(1-\gamma)\xi}$ in (10) if $\sum_{t=0}^{T-1}\big(b - F_g(q^{(t)})\big) \geq 0$; otherwise $\lambda = 0$. Thus,

$$\begin{aligned}&F_r(q^{\theta^\star}) - F_r(q') \;+\; \frac{2}{(1-\gamma)\xi}\,[b - F_g(q')]_+\\ &\leq\; \frac{DL}{T^{1/4}} \;+\; \frac{1}{2\eta_2(1-\gamma)^2\xi^2 T} \;+\; \frac{\eta_2}{2(1-\gamma)^2}\end{aligned}$$

where $F_r(q') := \frac{1}{T}\sum_{t=0}^{T-1} F_r(q^{(t)})$ and $F_g(q') := \frac{1}{T}\sum_{t=0}^{T-1} F_g(q^{(t)})$ for some occupancy measure q'.

Notice that $\frac{2}{(1-\gamma)\xi} \geq 2\lambda^\star$. By [11, Lemma 2],

$$[b - F_g(q')]_+ \;\leq\; \frac{(1-\gamma)\xi DL}{T^{1/4}} + \frac{1}{2\eta_2(1-\gamma)\xi T} + \frac{\eta_2\xi}{2(1-\gamma)}$$

which readily leads to the desired constraint violation bound by noting that $\frac{1}{T}\sum_{t=0}^{T-1}\big(b - F_g(q^{(t)})\big) = b - F_g(q')$, $\eta_2 = \frac{8|A||S|(1+2/\xi)}{(1-\gamma)^4\sqrt{T}}\|d_\rho^{\pi^\star}/\rho\|_\infty^2$, and $\|d_\rho^{\pi^\star}/\rho\|_\infty^2 \geq (1-\gamma)^2$.

IV. Sample-Based Implementation

We present a sample-based implementation of PG-PD in Section IV-A and show its sample complexity. We provide computational experiments in Section IV-B.

A. Sample-Based PG-PD

We introduce a sample-based implementation of PG-PD by assuming access to policy simulators or generative models. We estimate policy gradient via

$$\frac{\partial V_L^\theta(\rho)}{\partial \theta_{s,a}} \;=\; \frac{1}{1-\gamma}\, d_\rho^\pi(s)\, Q_L^\pi(s,a) \qquad (11)$$

where $d_\rho^\pi(s) = (1-\gamma)\sum_{t=0}^\infty \gamma^t P^\pi(s_t = s \,|\, s_0 \sim \rho)$ and $Q_L^\pi(s,a) = Q_r^\pi(s,a) + \lambda Q_g^\pi(s,a)$. It suffices to estimate $d_\rho^\pi(s)$ and $Q_L^\pi(s,a)$, independently. At each time t, we can apply the random horizon rollouts [27] to obtain unbiased estimates of $Q_L^{(t)}(s,a)$ and $d_\rho^{(t)}(s)$,

$$\mathbb{E}\big[\hat{Q}_L^{(t)}(s,a)\big] \;=\; Q_L^{(t)}(s,a) \;\text{ and }\; \mathbb{E}\big[\hat{d}_{s_0}^{(t)}(s)\big] \;=\; d_{s_0}^{(t)}(s).$$

Estimating $V_g^{(t)}(\rho)$ is similar. Hence, (11) can be estimated by $\frac{1}{1-\gamma}\hat{d}_\rho^{(t)}(s)\hat{Q}_L^{(t)}(s,a)$ for all $(s,a) \in S \times A$. For each estimate, we collect K trajectories with random horizon and average K estimates to reduce the variance. We show sample complexity guarantee in Theorem 3.

Theorem 3 (Sample Complexity): Let Assumption 1 hold. Fix $\Lambda = [0, 2/((1-\gamma)\xi)]$ and $\rho > 0$. Fix $T > 0$, $K = \Theta(T)$, $\lambda^{(0)} = 0$, and $\theta^{(0)}$ such that $\mathbb{E}[V_r^{\theta^{(0)}}(\rho)] \geq V_r^\star(\rho)$. Suppose the iterates $\pi^{(t)}$ and $\lambda^{(t)}$ are generated by sample-based PG-PD with $\eta_1 = O(1)$ and $\eta_2 = O(1/\sqrt{T})$, in which K rounds of trajectory samples are used at each time t. Then,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\big(V_r^\star(\rho) - V_r^{(t)}(\rho)\big)\right] \;\leq\; C_3\frac{|A||S|\|d_\rho^{\pi^\star}/\rho\|_\infty^2}{(1-\gamma)^6 T^{1/4}}\left(1 + \frac{C_3'}{K}\right)$$

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\big(b - V_g^{(t)}(\rho)\big)\right]_+ \;\leq\; C_4\frac{|A||S|\|d_\rho^{\pi^\star}/\rho\|_\infty^2}{(1-\gamma)^6 T^{1/4}}\left(1 + \frac{C_4'}{K}\right)$$

where C_3, C_4, C_3', C_4' are absolute constants.

Proof: It is similar to proving Theorem 1, except that we use estimated values and their variances as in [11]. ∎

Theorem 3 shows the sampling effect via the sample size K. If K is large enough, the rate matches Theorem 1. Also, Theorem 3 matches the rate for stochastic minimax optimization [18] in the number of trajectories KT.

B. Computational Experiments

In this experiment, we randomly generate a CMDP with $|S| = 10$, $|A| = 5$, $\gamma = 0.8$, and $b = 3$. We first simulate PG-PD (3), where we choose algorithm parameters $\eta_1 = \eta_2 = 1$ and initialize policy by a policy generated by the policy iteration. We compute the policy gradient (11) exactly via the Bellman equations [28]. As shown in Fig. 1, the optimality gap converges to zero sublinearly and the constraint violation approaches a non-positive constant, yielding zero violation. Secondly, we test sample-based PG-PD with the same CMDP setting and algorithm stepsizes. In Fig. 2, we show two random instances with different total numbers of samples K. For fixed T, as we increase K, our algorithm approaches stationary point with better reward objective and constraint violation.

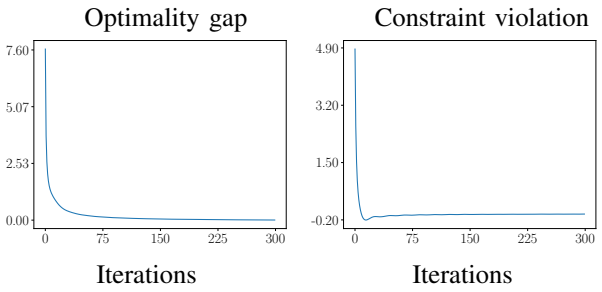

Fig. 1: Performance of PG-PD.

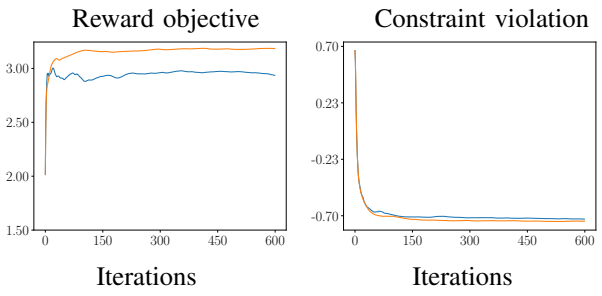

Fig. 2: Performance of Sample-Based PG-PD: $K = 300$ (—) and $K = 900$ (—).

V. Conclusion

We have utilized the policy gradient primal-dual (PG-PD) method for solving tabular CMDPs. In addition to establishing non-asymptotic global convergence guarantees, we have also proved convergence and quantified sample complexity for an associated sample-based algorithm. Our ongoing work focuses on a unified framework for policy gradient primal-dual methods with or without Fisher preconditioning.

References

[1] E. Altman, *Constrained Markov Decision Processes*. CRC Press, 1999, vol. 7.

[2] G. Dulac-Arnold, D. Mankowitz, and T. Hester, "Challenges of real-world reinforcement learning," *arXiv preprint arXiv:1904.12901*, 2019.

[3] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[4] J. F. Fisac, A. K. Akametalu, M. N. Zeilinger, S. Kaynama, J. Gillula, and C. J. Tomlin, "A general safety framework for learning-based control in uncertain robotic systems," *IEEE Transactions on Automatic Control*, vol. 64, no. 7, pp. 2737–2752, 2018.

[5] C. J. Girard, "Structural results for constrained Markov decision processes," Ph.D. dissertation, Cornell University, 2018.

[6] N. Abe, P. Melville, C. Pendus, C. K. Reddy, D. L. Jensen, V. P. Thomas, J. J. Bennett, G. F. Anderson, B. R. Cooley, M. Kowalczyk *et al.*, "Optimizing debt collections using constrained reinforcement learning," in *International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 75–84.

[7] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, "Risk-constrained reinforcement learning with percentile risk criteria," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6070–6120, 2017.

[8] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," in *International Conference on Learning Representations*, 2019.

[9] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *International Conference on Machine Learning*, vol. 70, 2017, pp. 22–31.

[10] S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro, "Constrained reinforcement learning has zero duality gap," in *Advances in Neural Information Processing Systems*, 2019, pp. 7553–7563.

[11] D. Ding, K. Zhang, T. Başar, and M. Jovanović, "Natural policy gradient primal-dual method for constrained Markov decision processes," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[12] A. Stooke, J. Achiam, and P. Abbeel, "Responsive safety in reinforcement learning by PID Lagrangian methods," in *International Conference on Machine Learning*, 2020, pp. 9133–9143.

[13] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "On the theory of policy gradient methods: Optimality, approximation, and distribution shift," *arXiv preprint arXiv:1908.00261*, 2019.

[14] F. V. Abad, V. Krishnamurthy, K. Martin, and I. Baltcheva, "Self learning control of constrained markov chains-a gradient approach," in *Proceedings of the 41st IEEE Conference on Decision and Control, 2002.*, vol. 2. IEEE, 2002, pp. 1940–1945.

[15] F. V. Abad and V. Krishnamurthy, "Policy gradient stochastic approximation algorithms for adaptive control of constrained time varying markov decision processes," in *42nd IEEE International Conference on Decision and Control (IEEE Cat. No. 03CH37475)*, vol. 3. IEEE, 2003, pp. 2823–2828.

[16] V. S. Borkar, "An actor-critic algorithm for constrained Markov decision processes," *Systems & control letters*, vol. 54, no. 3, pp. 207–213, 2005.

[17] E. Uchibe and K. Doya, "Constrained reinforcement learning from intrinsic and extrinsic rewards," in *International Conference on Development and Learning*, 2007, pp. 163–168.

[18] T. Lin, C. Jin, and M. I. Jordan, "On gradient descent ascent for nonconvex-concave minimax problems," in *International Conference on Machine Learning*, 2019.

[19] ——, "Near-optimal algorithms for minimax optimization," in *Conference on Learning Theory*, 2020.

[20] D. Ding, X. Wei, Z. Yang, Z. Wang, and M. R. Jovanović, "Provably efficient safe exploration via primal-dual policy optimization," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, vol. 130, 2021, pp. 3304–3312.

[21] Y. Efroni, S. Mannor, and M. Pirotta, "Exploration-exploitation in constrained MDPs," *arXiv preprint arXiv:2003.02189*, 2020.

[22] Q. Liang, F. Que, and E. Modiano, "Accelerated primal-dual policy optimization for safe reinforcement learning," *arXiv preprint arXiv:1802.06480*, 2018.

[23] M. Yu, Z. Yang, M. Kolar, and Z. Wang, "Convergent policy optimization for safe reinforcement learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 3121–3133.

[24] S. Paternain, M. Calvo-Fullana, L. F. Chamon, and A. Ribeiro, "Safe policies for reinforcement learning via primal-dual methods," *arXiv preprint arXiv:1911.09101*, 2019.

[25] Y. Chen, J. Dong, and Z. Wang, "A primal-dual approach to constrained Markov decision processes," *arXiv preprint arXiv:2101.10895*, 2021.

[26] T. Xu, Y. Liang, and G. Lan, "A primal approach to constrained policy optimization: Global optimality and finite-time analysis," *arXiv preprint arXiv:2011.05869*, 2020.

[27] S. Paternain, "Stochastic control foundations of autonomous behavior," Ph.D. dissertation, University of Pennsylvania, 2018.

[28] J. Bhandari and D. Russo, "Global optimality guarantees for policy gradient methods," *arXiv preprint arXiv:1906.01786*, 2019.

Appendix

A. Proof of Lemma 2

By the smoothness of value functions [13, Lemma D.3],

$$|F_r(q^\theta) - F_r(q^{(t)}) - \langle \nabla_\theta F_r(q^{(t)}), \theta - \theta^{(t)} \rangle| \leq \frac{\gamma|A|}{(1-\gamma)^3} \|\theta - \theta^{(t)}\|^2.$$

If we fix $\lambda^{(t)} \in [0, 2/((1-\gamma)\xi)]$, then

$$\left|(F_r + \lambda^{(t)} F_g)(q^\theta) - (F_r + \lambda^{(t)} F_g)(q^{(t)}) - \left\langle \nabla_\theta F_r(q^{(t)}) + \lambda^{(t)} \nabla_\theta F_g(q^{(t)}), \theta - \theta^{(t)} \right\rangle\right| \leq \frac{L}{2} \|\theta - \theta^{(t)}\|^2.$$

Application of absolute value inequalities above twice yields,

$$\begin{aligned}(F_r+\lambda^{(t)}F_g)(q^\theta) &\geq (F_r+\lambda^{(t)}F_g)(q^{(t)})\\ +\big\langle\nabla_\theta F_r(q^{(t)})+\lambda^{(t)}\nabla_\theta F_g(q^{(t)}),\theta-\theta^{(t)}\big\rangle &- \tfrac{L}{2}\|\theta-\theta^{(t)}\|^2\\ &\geq (F_r+\lambda^{(t)}F_g)(q^\theta)-L\|\theta-\theta^{(t)}\|^2.\end{aligned} \tag{12}$$

The primal update in (3) is equivalent to

$$\begin{aligned}\theta^{(t+1)} &= \operatorname*{arg\,max}_{\theta\in\Theta}\Big\{V_r^{(t)}(\rho)+\lambda^{(t)}V_g^{(t)}(\rho)\\ +\big\langle\nabla_\theta V_r^{(t)}(\rho)+\lambda^{(t)}\nabla_\theta V_g^{(t)}(\rho),\theta-\theta^{(t)}\big\rangle &- \tfrac{1}{2\eta_1}\|\theta-\theta^{(t)}\|^2\Big\}.\end{aligned}$$

Take $\eta_1=\frac{1}{L}$ and let $\theta=\theta^{(t+1)}$ in (12). Hence,

$$\begin{aligned}&(F_r+\lambda^{(t)}F_g)(q^{(t+1)})\\ &\geq \operatorname*{maximize}_{\theta\in\Theta}\Big\{(F_r+\lambda^{(t)}F_g)(q^{(t)})\\ &+\big\langle\nabla_\theta F_r(q^{(t)})+\lambda^{(t)}\nabla_\theta F_g(q^{(t)}),\theta-\theta^{(t)}\big\rangle-\tfrac{L}{2}\|\theta-\theta^{(t)}\|^2\Big\}\\ &\geq \operatorname*{maximize}_{\theta\in\Theta}\Big\{(F_r+\lambda^{(t)}F_g)(q^\theta)-L\|\theta-\theta^{(t)}\|^2\Big\}\\ &\geq \operatorname*{maximize}_{\alpha\in[0,1]}\Big\{(F_r+\lambda^{(t)}F_g)(q^{\theta_\alpha})-L\|\theta_\alpha-\theta^{(t)}\|^2\Big\}\end{aligned} \tag{13}$$

where $\theta_\alpha:=\pi^q(\alpha q^\star+(1-\alpha)q^{(t)})$, we use (12) in the second inequality, the last inequality is due to $\pi^q\circ q^\pi=\mathrm{id}_{SA}$ and linearity of q^θ in θ. By the linearity of F_r, F_g in q^θ,

$$\begin{aligned}&(F_r+\lambda^{(t)}F_g)(q^{\theta_\alpha})\\ &= \alpha(F_r+\lambda^{(t)}F_g)(q^\star)+(1-\alpha)(F_r+\lambda^{(t)}F_g)(q^{(t)}).\end{aligned} \tag{14}$$

By the definition of π^q, $(\pi^q(q)-\pi^q(q'))_{sa}$ equals to

$$\frac{1}{\sum_{a\in A}q_{sa}}(q_{sa}-q'_{sa})+\frac{\sum_{a\in A}q'_{sa}-\sum_{a\in A}q_{sa}}{\sum_{a\in A}q_{sa}\sum_{a\in A}q_{sa}}q'_{sa}$$

which yields an upper bound on $\|\pi^q(q)-\pi^q(q')\|^2$,

$$\begin{aligned}&\|\pi^q(q)-\pi^q(q')\|^2\\ &= \textstyle\sum_{s\in S}\sum_{a\in A}\left((\pi^q(q)-\pi^q(q'))_{sa}\right)^2\\ &\leq 2\textstyle\sum_{s\in S}\sum_{a\in A}\frac{(q_{sa}-q'_{sa})^2}{(\sum_{a\in A}q_{sa})^2}\\ &\quad+2\textstyle\sum_{s\in S}\sum_{a\in A}\left(\frac{\sum_{a\in A}q'_{sa}-\sum_{a\in A}q_{sa}}{\sum_{a\in A}q_{sa}\sum_{a\in A}q_{sa}}\right)^2(q'_{sa})^2\\ &\leq 2\textstyle\sum_{s\in S}\frac{1}{(\sum_{a\in A}q_{sa})^2}\Big(\sum_{a\in A}(q_{sa}-q'_{sa})^2\\ &\qquad+\left(\textstyle\sum_{a\in A}q'_{sa}-\sum_{a\in A}q_{sa}\right)^2\Big).\end{aligned}$$

where we also use $\|x+y\|^2\leq 2\|x\|^2+2\|y\|^2$. Let $q^\star=q^{\theta^\star}$. Then, we bound $\|\theta_\alpha-\theta^{(t)}\|^2$ by

$$\begin{aligned}&\|\theta_\alpha-\theta^{(t)}\|^2\\ &= \left\|\pi^q\left(\alpha q^\star+(1-\alpha)q^{(t)}\right)-\pi^q\left(q^{(t)}\right)\right\|^2\\ &\leq \textstyle\sum_{s\in S}\frac{2\alpha^2}{\left(\sum_{a\in A}q_{sa}^{(t)}\right)^2}\Big(\sum_{a\in A}\left(q_{sa}^\star-q_{sa}^{(t)}\right)^2\\ &\qquad+\left(\textstyle\sum_{a\in A}q_{sa}^{(t)}-\sum_{a\in A}q_{sa}^\star\right)^2\Big)\end{aligned}$$

in which the upper bound further can be expressed as

$$\begin{aligned}&\textstyle\sum_{s\in S}\frac{4\alpha^2}{\left(\sum_{a\in A}q_{sa}^{(t)}\right)^2}\left(\left(\sum_{a\in A}q_{sa}^\star\right)^2+\left(\sum_{a\in A}q_{sa}^{(t)}\right)^2\right)\\ &= 4\alpha^2\textstyle\sum_{s\in S}\frac{\left(d_\rho^{\pi^\star}(s)\right)^2+\left(d_\rho^{\pi^{(t)}}(s)\right)^2}{\left(d_\rho^{\pi^{(t)}}(s)\right)^2}\\ &\leq 4\alpha^2|S|\left\|\tfrac{d_\rho^{\pi^\star}}{d_\rho^{\pi^{(t)}}}\right\|_\infty^2+4\alpha^2|S|\\ &\leq \alpha^2 D\end{aligned} \tag{15}$$

where we apply $d_\rho^{\pi^{(t)}}\geq(1-\gamma)\rho$ componentwise in the second inequality. We now apply (14) and (15) to (13),

$$\begin{aligned}&(F_r+\lambda^{(t)}F_g)(q^\star)-(F_r+\lambda^{(t)}F_g)(q^{(t+1)})\\ &\leq \operatorname*{minimize}_{\alpha\in[0,1]}\Big\{L\|\theta_\alpha-\theta^{(t)}\|^2\\ &\qquad+(F_r+\lambda^{(t)}F_g)(q^\star)-(F_r+\lambda^{(t)}F_g)(q^{\theta_\alpha})\Big\}\\ &\leq \operatorname*{minimize}_{\alpha\in[0,1]}\Big\{\alpha^2DL\\ &\qquad+(1-\alpha)\big((F_r+\lambda^{(t)}F_g)(q^\star)-(F_r+\lambda^{(t)}F_g)(q^{(t)})\big)\Big\}\end{aligned}$$

which further implies

$$\begin{aligned}&(F_r+\lambda^{(t+1)}F_g)(q^\star)-(F_r+\lambda^{(t+1)}F_g)(q^{(t+1)})\\ &\leq \operatorname*{minimize}_{\alpha\in[0,1]}\Big\{\alpha^2DL\\ &\qquad+(1-\alpha)\big((F_r+\lambda^{(t)}F_g)(q^\star)-(F_r+\lambda^{(t)}F_g)(q^{(t)})\big)\Big\}\\ &\qquad-(\lambda^{(t)}-\lambda^{(t+1)})\big(F_g(q^\star)-F_g(q^{(t+1)})\big).\end{aligned} \tag{16}$$

We now check the right-hand side of (16). By the dual update in (3), it is easy to see that $-(\lambda^{(t)}-\lambda^{(t+1)})(F_g(q^\star)-F_g(q^{(t+1)}))\leq|\lambda^{(t)}-\lambda^{(t+1)}|\frac{1}{1-\gamma}\leq\frac{\eta_2}{(1-\gamma)^2}$. We can solve the minimization problem in (16) by setting $\alpha=0$ if $\alpha^{(t)}<0$; $\alpha=1$ if $\alpha^{(t)}>1$; $\alpha=\alpha^{(t)}$ if $\alpha^{(t)}\in[0,1]$,

$$\alpha^{(t)} := \tfrac{(F_r+\lambda^{(t)}F_g)(q^\star)-(F_r+\lambda^{(t)}F_g)(q^{(t)})}{2DL}.$$

By setting $\eta_2=\frac{(1-\gamma)^2DL}{2\sqrt{T}}$, we consider three cases: (i) when $\alpha^{(t)}<0$, we set $\alpha=0$ for (16),

$$(F_r+\lambda^{(t+1)}F_g)(q^\star)-(F_r+\lambda^{(t+1)}F_g)(q^{(t+1)})\leq\tfrac{DL}{2\sqrt{T}}; \tag{17}$$

(ii) when $\alpha^{(t)}>1$, we set $\alpha=1$ that leads to $(F_r+\lambda^{(t+1)}F_g)(q^\star)-(F_r+\lambda^{(t+1)}F_g)(q^{(t+1)})\leq\frac{3}{2}DL$, i.e., $\alpha^{(t+1)}\leq\frac{3}{4}$. Thus, this case reduces to the next case (iii): $0\leq\alpha^{(t)}\leq1$ in which we can express (16) as

$$\alpha^{(t+1)}\leq\left(1-\tfrac{\alpha^{(t)}}{2}\right)\alpha^{(t)}+\tfrac{1}{4\sqrt{T}}. \tag{18}$$

By choosing $\lambda^{(0)}=0$ and $\theta^{(0)}$ such that $V_r^{\theta^{(0)}}(\rho)\geq V_r^{\theta^\star}(\rho)$, we know that $\alpha^{(0)}\leq0$. Thus, $\alpha^{(1)}\leq\frac{1}{4\sqrt{T}}$. By (17), the case $\alpha^{(1)}\leq0$ is trivial.

Without loss of generality, we assume that $0\leq\alpha^{(t)}\leq\frac{1}{T^{1/4}}\leq1$. By induction of (18) over t, $\alpha^{(t+1)}\leq\frac{1}{T^{1/4}}$. Combining this with (17), and averaging over $t=0,1,\cdots,T-1$, we obtain the desired bound.