Natural Policy Gradient Primal-Dual Method for Constrained Markov Decision Processes

Dongsheng Ding
ECE
University of Southern California
dongshed@usc.edu

Kaiqing Zhang
ECE and CSL
University of Illinois at Urbana-Champaign
kzhang66@illinois.edu

Tamer Başar
ECE and CSL
University of Illinois at Urbana-Champaign
basar1@illinois.edu

Mihailo R. Jovanović
ECE
University of Southern California
mihailo@usc.edu

Abstract

We study sequential decision-making problems in which each agent aims to maximize the expected total reward while satisfying a constraint on the expected total utility. We employ the natural policy gradient method to solve the discounted infinite-horizon Constrained Markov Decision Processes (CMDPs) problem. Specifically, we propose a new Natural Policy Gradient Primal-Dual (NPG-PD) method for CMDPs which updates the primal variable via natural policy gradient ascent and the dual variable via projected sub-gradient descent. Even though the underlying maximization involves a nonconcave objective function and a nonconvex constraint set under the softmax policy parametrization, we prove that our method achieves global convergence with sublinear rates regarding both the optimality gap and the constraint violation. Such a convergence is independent of the size of the state-action space, i.e., it is dimension-free. Furthermore, for the general smooth policy class, we establish sublinear rates of convergence regarding both the optimality gap and the constraint violation, up to a function approximation error caused by restricted policy parametrization. Finally, we show that two sample-based NPG-PD algorithms inherit such non-asymptotic convergence properties and provide finite-sample complexity guarantees. To the best of our knowledge, our work is the first to establish non-asymptotic convergence guarantees of policy-based primal-dual methods for solving infinite-horizon discounted CMDPs. We also provide computational results to demonstrate merits of our approach.

1 Introduction

Reinforcement learning (RL) studies sequential decision-making problems where the agent aims to maximize its expected total reward by interacting with an unknown environment over time [44]. The model of Markov Decision Processes (MDPs) is usually used to represent the environment dynamics. However, in many safety-critical applications, e.g., in autonomous driving [19], robotics [35], cyber-security [58], and financial management [1], the agent is also subject to constraints on its utilities/costs. This naturally leads to a generalization of the environment dynamics to constrained MDPs (CMDPs) [4]. Besides maximizing the expected total reward, the agent also has to take into account the constraint on the expected total utility/cost as an additional learning objective.

Policy gradient (PG) methods [45], including the natural policy gradient (NPG) [21], have enjoyed substantial empirical success in solving MDPs [39, 24, 31, 40, 44]. PG methods, or more generally *direct policy search* methods, have also been used to solve CMDPs [47, 12, 11, 15, 46, 23, 36, 2, 43].

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

However, most existing theoretical guarantees are asymptotic in nature and/or only provide local convergence guarantees to stationary-point policies. Theoretical non-asymptotic global convergence guarantees are largely absent: for arbitrary initial points, algorithms with a finite number of iterations and samples converge to an ϵ-optimal solution that enjoys ϵ-optimality gap and ϵ-constraint violation. It is thus imperative to establish theoretical guarantees for PG methods in solving CMDPs. Our motivation also comes from recent advances on the global convergence properties of PG methods [18, 56, 9, 48, 3, 57].

In this work, we provide a theoretical foundation for the non-asymptotic global convergence of the NPG method in solving CMDPs and answer the following questions: (i) can we employ NPG methods for solving CMDPs?; (ii) if and how fast do these methods converge to the globally optimal value within the underlying constraints?; (iii) what is the effect of the function approximation error caused by a restricted policy parametrization?; and (iv) what is the sample complexity of NPG methods?

Contribution. Our contribution is four-fold: (i) We propose a simple but effective primal-dual algorithm – Natural Policy Gradient Primal-Dual (NPG-PD) method – for solving discounted infinite-horizon CMDPs. We employ natural policy gradient ascent to update the primal variable and projected sub-gradient descent to update the dual variable; (ii) Even though we show that the maximization problem has a nonconcave objective function and nonconvex constraint set under the softmax policy parametrization, we prove that our NPG-PD method achieves global convergence with rate $O(1/\sqrt{T})$ regarding both the optimality gap and the constraint violation, where T is the total number of iterations. Our convergence guarantees are dimension-free, i.e., the rate is independent of the size of the state-action space; (iii) For the general smooth policy class, we establish convergence with rate $O(1/\sqrt{T})$ for the optimality gap and $O(1/T^{1/4})$ for the constraint violation, up to a function approximation error caused by restricted policy parametrization; and (iv) We show that two sample-based NPG-PD algorithms that we propose inherit such non-asymptotic convergence properties and provide the finite-sample complexity guarantees. To the best of our knowledge, our work is the first to provide non-asymptotic convergence guarantees for solving infinite-horizon discounted CMDPs in the primal-dual framework.

Related Work. Our work is related to Lagrangian-based CMDP algorithms [4, 12, 11, 15, 46, 23, 37, 36, 53]. However, convergence guarantees of these algorithms are either local (to stationary-point or locally optimal policies) [11, 15, 46] or asymptotic [12]. When function approximation is used for policy parametrization, [53] recognized the lack of convexity and showed asymptotic convergence (to a stationary point) of a method based on successive convex relaxations. In contrast, we establish global convergence in spite of the lack of convexity. References [36, 37] are closely related to our work. In [37], the authors provide duality analysis for CMDPs in the policy space and propose a provably convergent dual descent algorithm by assuming access to a nonconvex optimization oracle. However, how to obtain the solution to this nonconvex optimization was not analyzed/understood, and the global convergence of their algorithm was not established. In [36], the authors provide a primal-dual algorithm but do not offer any theoretical justification. In spite of the lack of convexity, our work provides global convergence guarantees for a new primal-dual algorithm without using any optimization oracles. Other related policy optimization methods include CPG [47], CPO [2, 51], and IPPO [27]. However, theoretical guarantees for these algorithms are still lacking. Recently, optimism principles have been used for efficient exploration in CMDPs [42, 59, 16, 38, 17, 6]. In comparison, our work focuses on the optimization landscape within a primal-dual framework.

Our work is also pertinent to recent global convergence results for PG methods. References [18, 32, 33] provide global convergence guarantees for (natural) PG methods for nonconvex linear quadratic regulator problems in both discrete- and continuous-time. For general MDPs, [56] shows that locally optimal policies are achievable using PG methods with a simple reward-reshaping. Reference [48] shows that (natural) PG methods converge to the globally optimal value when overparametrized neural networks are used. As a variant of natural PG, trust-region policy optimization (TRPO) [39] has also been shown to converge to the globally optimal policy with overparametrized neural networks [26] and, in general, with regularized MDPs [41]. References [9, 10] study global optimality and convergence of PG methods from a policy iteration perspective. Reference [3] provides characterizations of the global convergence properties of (natural) PG methods regarding computational, approximation, and sample size issues. Recent advances along this line include references [30, 55, 13, 28]. While all these references handle a lack of convexity in the objective function, additional effort is required to deal with nonconvex constraint sets that arise in CMDPs, and our paper addresses this challenge.

2 Constrained Markov Decision Processes

Consider a discounted Constrained Markov Decision Process [4] – CMDP$(S, A, P, r, g, b, \gamma, \rho)$ – where S is a finite state space, A is a finite action space, 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: $S \to \Delta_A$, determining 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 Π be a set of all possible policies. 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 the reward r or the utility g are the following expected values of total rewards or utilities received under policy π, respectively,

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

where the expectation $\mathbb{E}$ is taken over the randomness of the 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 the 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: $S \times A \to \mathbb{R}$,

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

where symbol $\diamond$ is r or g. Since $r, g \in [0, 1]$, it is easy to see that $V_r^\pi(s), V_g^\pi(s) \in [0, 1/(1-\gamma)]$. Their expected values under ρ are: $V_r^\pi(\rho) := \mathbb{E}_{s_0 \sim \rho}[V_r^\pi(s_0)]$ and $V_g^\pi(\rho) := \mathbb{E}_{s_0 \sim \rho}[V_g^\pi(s_0)]$.

Having defined policy, value/action-value functions for the discounted CMDP, the agent's goal is to find a policy that maximizes the expected reward value subject to a constraint on the expected utility value,

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

in which we maximize over all policies and we set constraint offset $b \in (0, 1/(1-\gamma)]$ to avoid triviality. For multiple constraints, our formulation (1) and convergence results are readily generalizable.

Via the method of Lagrange multipliers [8], we formulate the problem (1) into the following max-min problem for the associated Lagrangian $V_L^{\pi,\lambda}(\rho)$,

$$\underset{\pi \,\in\, \Pi}{\text{maximize}} \;\; \underset{\lambda \,\geq\, 0}{\text{minimize}} \;\; V_L^{\pi,\lambda}(\rho) \;:=\; V_r^\pi(\rho) \,+\, \lambda\,(V_g^\pi(\rho) - b) \tag{2}$$

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

Instead of the linear program method [4], this work focuses on *direct policy search* methods for solving the problem (2). Direct methods are attractive since they can deal with large state-action spaces via policy parameterization, e.g., neural nets, and they allow us to directly optimize/monitor the value functions that we are interested in. They are useful especially when we can utilize policy gradient estimates via simulations of the policy. It is worth mentioning that (1) is a nonconcave optimization problem as we prove in Lemma 3, and the decision is on an infinite-dimensional policy space Π. These reasons make the problem (1) challenging.

Nevertheless, the problem (1) has nice properties in the policy space once it is strictly feasible. We adapt the Slater condition in the constrained optimization [8] to assume strict feasibility of (1).

Assumption 1 (Slater Condition)**.** *There exists* $\xi > 0$ *and* $\bar{\pi} \in \Pi$ *such that* $V_g^{\bar{\pi}}(\rho) - b \geq \xi$.

The Slater condition is mild in practice since we usually have *a priori* knowledge on a strictly feasible policy, e.g., the minimal utility is achievable by a particular policy so that the constraint becomes loose. We will assume this throughout the paper.

Let an optimal primal variable be $\pi^\star$, i.e., an optimal solution to (1). Let an optimal dual variable be $\lambda^\star \in \text{argmin}_{\lambda \,\geq\, 0} V_D^\lambda(\rho)$. Use the shorthand notation $V_r^{\pi^\star}(\rho) = V_r^\star(\rho)$ and $V_D^{\lambda^\star}(\rho) = V_D^\star(\rho)$

whenever it is clear from the context. We now recall the strong duality for CMDPs [4, 36] and we further prove boundedness of optimal dual variable $\lambda^\star$ in Corollary 1 in Appendix C.

Lemma 1 (Strong Duality and Boundedness of $\lambda^\star$)**.** *Let Assumption 1 hold. Then, (i) $V_r^\star(\rho) = V_D^\star(\rho)$; (ii) $0 \leq \lambda^\star \leq (V_r^\star(\rho) - V_r^{\bar\pi}(\rho))/\xi$.*

Let $v(\tau) = \text{maximize}_{\pi\in\Pi}\{V_r^\pi(\rho)\,|\,V_g^\pi(\rho) \geq b+\tau\}$ be the value function associated with problem (1). Using concavity of $v(\tau)$ given by [36, Proposition 1], we provide a useful bound on the constraint violation in Lemma 2 and it is proved in Theorem 6 in Appendix C. Take $[x]_+ = \max(x,0)$.

Lemma 2 (Constraint Violation)**.** *Let Assumption 1 hold. For any $C \geq 2\lambda^\star$, if there exists a policy $\pi\in\Pi$ and $\delta > 0$ such that $V_r^\star(\rho) - V_r^\pi(\rho) + C[b - V_g^\pi(\rho)]_+ \leq \delta$, then $[b - V_g^\pi(\rho)]_+ \leq 2\delta/C$.*

Aided by the above properties implied by the Slater condition, this work directly studies the max-min problem (2) in the primal-dual domain.

3 Policy Parametrization and Natural Policy Gradient Primal-Dual Method

To make problem (1) tractable, we introduce a set of parametrized policies $\{\pi_\theta\,|\,\theta\in\Theta\}$ where Θ is a finite-dimensional parameter space. We reduce the problem (1) into a parametric optimization problem,

$$\underset{\theta\,\in\,\Theta}{\text{maximize}}\;\; V_r^{\pi_\theta}(\rho) \;\;\text{subject to}\;\; V_g^{\pi_\theta}(\rho) \;\geq\; b \tag{3}$$

together with a parametric max-min problem (2) for the Lagrangian $V_L^{\pi_\theta,\lambda}(\rho)$,

$$\underset{\theta\,\in\,\Theta}{\text{maximize}}\;\;\underset{\lambda\,\geq\,0}{\text{minimize}}\;\; V_L^{\pi_\theta,\lambda}(\rho) \;:=\; V_r^{\pi_\theta}(\rho) \,+\, \lambda\,(V_g^{\pi_\theta}(\rho) - b). \tag{4}$$

The dual function reads $V_D^\lambda(\rho) := \text{maximize}_\theta\, V_L^{\pi_\theta,\lambda}(\rho)$. The problem (3) is finite-dimensional, but still nonconcave even if the constraint is absent [3]. We state it formally as follows and prove it in Appendix A via an easily-constructed CMDP example.

Lemma 3. *There is a CMDP such that for the problem (3), $V_r^{\pi_\theta}(s)$ is not concave and the constraint set $\{\theta\in\Theta\,|\,V_g^{\pi_\theta}(s)\geq b\}$ is not convex.*

The associated Lagrangian $V_L^{\pi_\theta,\lambda}(\rho)$ is thus nonconcave in θ and convex in λ and the problem (4) is a nonconcave-convex max-min problem. Many algorithms, e.g., [25, 34, 50], for solving max-min optimization problems though, strong assumptions on the max-min structure or only stationary-point convergence guarantees make them not suitable here. In this work, we will exploit our problem geometry to propose a new method to study the max-min problem (4). Before doing that, we first introduce two classes of policies that we are interested in.

Softmax Parametrization. A natural class of policies is parametrized by the softmax function,

$$\pi_\theta(a\,|\,s) \;=\; \frac{\exp(\theta_{s,a})}{\sum_{a'\in A}\exp(\theta_{s,a'})} \;\;\text{for all}\;\; \theta\in\mathbb{R}^{|S||A|}. \tag{5}$$

Nice analytical properties of the softmax policy include completeness and differentiability. It can represent any stochastic policy, and its closure contains all stationary policies. Other reasons for us to begin with this policy class are: (i) it equips the policy with a rich structure so that the natural PG update works like the classical multiplicative weights update in the online learning literature, e.g., [14]; (ii) it has served as lens to interpreting the function approximation error [3]. It is a warm-up for studying convergence properties of many RL algorithms [9, 3, 30, 13].

General Parametrization. A general class of stochastic policies is given by $\{\pi_\theta\,|\,\theta\in\Theta\}$ in which we assume $\Theta\subset\mathbb{R}^d$ without providing the structure of π_θ. The parameter space has dimension d. This policy class covers a more practical setting using function approximation, e.g., (deep) neural networks [26, 48]. However, when we choose $d \ll |S||A|$, the policy class has a limited expressiveness, and it may not contain all stochastic policies, e.g., being *restricted*. With this in mind, it is reasonable for our theory to define global convergence up to some error caused by the restricted policy class.

Natural Policy Gradient Primal-Dual (NPG-PD) Method. To introduce our method, we first introduce some useful definitions. The discounted visitation distribution $d_{s_0}^\pi$ of a policy π and its expectation over initial distribution ρ are given by,

$$d_{s_0}^{\pi}(s) \;=\; (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P^{\pi}(s_t = s \,|\, s_0) \;\text{ and }\; d_{\rho}^{\pi}(s) \;=\; \mathbb{E}_{s_0\sim\rho}\left[\, d_{s_0}^{\pi}(s)\,\right] \tag{6}$$

where $P^{\pi}(s_t = s \,|\, s_0)$ is the probability of visiting state s at time t when the agent follows the policy π with initial state s_0. When the parametrized policy π_θ is clear from the context, we use $V_r^{\theta}(\rho)$ instead of $V_r^{\pi_\theta}(\rho)$, and similarly for others. When $\pi_\theta(\cdot\,|\,s)$ is differentiable and it is in the probability simplex, i.e., $\pi_\theta \in \Delta_A^{|S|}$ for all θ, the policy gradient (PG) of the Lagrangian (4) reads,

$$\nabla_\theta V_L^{\theta,\lambda}(s_0) \;=\; \frac{1}{1-\gamma}\,\mathbb{E}_{s_0\,\sim\, d_{s_0}^{\pi_\theta}}\,\mathbb{E}_{a\,\sim\,\pi_\theta(\cdot\,|\,s)}\left[\nabla_\theta \log \pi_\theta(a\,|\,s)\cdot A_L^{\theta,\lambda}(s,a)\right]$$

which equals $\nabla_\theta V_r^{\theta}(s_0) + \lambda \nabla_\theta V_g^{\theta}(s_0)$ where $A_L^{\theta,\lambda}(s,a) := A_r^{\theta}(s,a) + \lambda A_g^{\theta}(s,a)$. The Fisher information matrix induced by π_θ is $F_\rho(\theta) := \mathbb{E}_{s\,\sim\, d_\rho^{\pi_\theta}}\mathbb{E}_{a\,\sim\,\pi_\theta(\cdot\,|\,s)}\left[\nabla_\theta \log\pi_\theta(a\,|\,s)\,(\nabla_\theta\log\pi_\theta(a\,|\,s))^\top\right]$.

With this notion, we propose a new policy gradient type method – Natural Policy Gradient Primal-Dual (NPG-PD) method – for the problem (4),

$$\theta^{(t+1)} \;=\; \theta^{(t)} \,+\, \eta_1\, F_\rho(\theta^{(t)})^\dagger \cdot \nabla_\theta V_L^{\theta^{(t)},\lambda^{(t)}}(\rho) \;\text{ and }\; \lambda^{(t+1)} \;=\; \mathcal{P}_\Lambda\big(\lambda^{(t)} - \eta_2\left(V_g^{\theta^{(t)}}(\rho) - b\right)\big) \tag{7}$$

where $A^\dagger$ takes the Moore-Penrose inverse of matrix A, $\mathcal{P}_\Lambda(x)$ projects x into the interval Λ that will be specified later, and $\eta_1 > 0, \eta_2 > 0$ are constants. This method displays a first-principle design of primal-dual updates: (i) the primal update $\theta^{(t+1)}$ performs gradient ascent using the natural policy gradient: $F_\rho(\theta^{(t)})^\dagger \cdot \nabla_\theta V_L^{(t)}(\rho)$, which is the policy gradient of $V_L^{(t)}(\rho)$ in the geometry induced by Fisher information $F_\rho(\theta^{(t)})$; (ii) the dual update $\lambda^{(t+1)}$ works as projected sub-gradient descent by adding up constraint violation $b - V_g^{(t)}(\rho)$. We use the shorthand $V_g^{(t)}(\rho)$ instead of $V_g^{\theta^{(t)}}(\rho)$, and similarly for others.

In what follows, we first establish global convergence of the NPG-PD method (7) under the softmax parametrization in Section 4. We move to the general parametrization in Section 5 and show convergence of a generalized version of (7). In the end, we propose two model-free sample-based algorithms for implementing (7) and analyze their sample complexities. Before our analysis, it is useful to recall: $V_\diamond^{\pi}(s_0) - V_\diamond^{\pi'}(s_0) = \frac{1}{1-\gamma}\,\mathbb{E}_{s\,\sim\, d_{s_0}^{\pi},\, a\,\sim\,\pi(\cdot\,|\,s)}[A_\diamond^{\pi'}(s,a)]$ for any two policies π, π', and any state s_0, where the symbol $\diamond$ is r or g. This is from the performance difference lemma [20, 3].

4 Softmax Parametrization: Dimension-free Global Convergence

We now study the NPG-PD method (7) under the softmax parametrization (5). Thanks to the completeness of the softmax policy class, the strong duality in Lemma 1 holds on the closure of the softmax policy class. We establish the global convergence with dimension-independent convergence rates, i.e., *dimension-free*, while the maximization problem (3) is nonconcave.

We first exploit the softmax policy structure to show that the NPG-PD update (7) enjoys a concise primal update given as follows. We provide a proof in Appendix B.

Lemma 4. *Let $\Lambda = [0, 2/((1-\gamma)\xi)]$. Further let $A_L^{(t)}(s,a) := A_r^{(t)}(s,a) + \lambda^{(t)} A_g^{(t)}(s,a)$ and $Z^{(t)}(s) := \sum_{a\in A}\pi^{(t)}(a\,|\,s)\exp\left(\frac{\eta_1}{1-\gamma}A_L^{(t)}(s,a)\right)$. Using parametrized policy (5), (7) equals to*

$$\begin{aligned}\theta_{s,a}^{(t+1)} \;=\; \theta_{s,a}^{(t)} \,+\, \tfrac{\eta_1}{1-\gamma}\,A_L^{(t)}(s,a) \quad \textit{or} \quad \pi^{(t+1)}(a\,|\,s) \;=\; \pi^{(t)}(a\,|\,s)\,\frac{\exp\left(\frac{\eta_1}{1-\gamma}A_L^{(t)}(s,a)\right)}{Z^{(t)}(s)} \\ \textit{and}\;\; \lambda^{(t+1)} \;=\; \mathcal{P}_\Lambda\big(\lambda^{(t)} - \eta_2\left(V_g^{(t)}(\rho) - b\right)\big).\end{aligned} \tag{8}$$

Due to the Moore-Penrose inverse of Fisher information, the updates (8) are free of the state distribution $d_\rho^{\pi^{(t)}}$ that appears in (7) through the policy gradient. The policy update imitates the multiplicative weights update that is recognized in the online linear optimization [14]. Here, the linear function is translated as an advantage function of the current policy at each iteration.

Next, we show the global convergence of the algorithm (8) regarding the optimality gap: $V_r^{\star}(\rho) - V_r^{(t)}(\rho)$ and the constraint violation: $b - V_g^{(t)}(\rho)$. We prove it in Appendix D.

Theorem 1 (Global Convergence: Softmax Parametrization)**.** *Let Assumption 1 hold for $\xi > 0$. Fix $T > 0$, $\rho \in \Delta_S$, $\theta^{(0)} = 0$, and $\lambda^{(0)} = 0$. If we choose $\eta_1 = 2\log|A|$ and $\eta_2 = (1-\gamma)/\sqrt{T}$, then* *the iterates $\pi^{(t)}$ generated by the algorithm (8) satisfy,*

$$\textit{(Optimality gap)}\quad \frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho)\,-\,V_r^{(t)}(\rho)\right)\;\leq\;\frac{4}{(1-\gamma)^2}\frac{1}{\sqrt{T}} \tag{9a}$$

$$\textit{(Constraint violation)}\quad \left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b\,-\,V_g^{(t)}(\rho)\right)\right]_+\;\leq\;\frac{1/\xi+4\xi}{(1-\gamma)^2}\frac{1}{\sqrt{T}}. \tag{9b}$$

What we capture in Theorem 1 is that on the average the reward value function converges to the global optimal one and the constraint violation decays to zero. Putting it differently, to find an ϵ-near-optimal value, e.g., ϵ-optimality gap and ϵ-constraint violation, the number of steps is $O(1/\epsilon^2)$ which is *independent of the sizes of the state space or the action space*. Although the maximization problem (3) is nonconcave, our bound $(\sqrt{T},\sqrt{T})$ for the accumulative optimality gap/constraint violation is better than the classical one $(\sqrt{T},T^{3/4})$ [29] and matches the rate [52] for solving online *convex* optimization with *convex* constraint sets.

Comparing to the unconstrained setting [3, Section 5.3], our proof needs additional efforts. As shown in Lemma 6 in Appendix D, the reward value function is coupled with the utility value function and neither of them enjoy monotonic improvement for the vanilla natural policy gradient method. Therefore, we must introduce a new line of analysis. We first establish the bounded average performance in Lemma 7 in Appendix D. It enables us to bound the optimality gap via drift analysis of the dual update. To deduce the constraint violation, it is tempting to use methods from the constrained convex optimization, e.g., [29, 52, 49, 54]. However, they are not satisfactory due to slow rate or the needed extra assumption. Instead, we establish that the constraint violation enjoys the same rate as the optimality gap under the strong duality in Lemma 2, although our problem (3) is nonconcave. To the best of our knowledge, this appears to be the first such result for a class of "nonconvex" problems.

Regarding the global convergence, our proof of Theorem 1 holds for arbitrary initializations and we use $\theta^{(0)}=0$, $\lambda^{(0)}=0$ just to ease the exposition. We use this simplification in the sequel.

5 General Parametrization: Convergence Rate and Optimality

In this section, we consider a general policy class $\{\pi_\theta\,|\,\theta\in\Theta\}$ in which $\Theta\subset\mathbb{R}^d$ is the parameter space. Let us consider a more general form for the update (7) with $\Lambda=[0,\infty)$,

$$\theta^{(t+1)}\;=\;\theta^{(t)}\,+\,\frac{\eta_1}{1-\gamma}\,w^{(t)}\;\;\text{and}\;\;\lambda^{(t+1)}\;=\;\mathcal{P}_\Lambda\big(\lambda^{(t)}-\eta_2\left(V_g^{(t)}(\rho)-b\right)\big) \tag{10}$$

where $w^{(t)}/(1-\gamma)$ is either the exact natural policy gradient (NPG) or some (sample-based) approximation of it. Note that in this general parametrization setting, the strong duality in Lemma 1 does not necessarily hold [36]. Thus, our analysis in Section 4 does not apply. Let us first generalize NPG. For a distribution over state-action pair: $\nu\in\Delta_{S\times A}$, the compatible function approximation error [21] is given by $E^\nu(w;\theta,\lambda):=\mathbb{E}_{s,a\sim\nu}[(A_L^{\theta,\lambda}(s,a)-w\cdot\nabla_\theta\log\pi_\theta(a\,|\,s))^2]$ where $A_L^{\theta,\lambda}(s,a):=A_r^\theta(s,a)+\lambda A_g^\theta(s,a)$, and the minimal error is $E_\star^\nu(\theta,\lambda):=\min_w E^\nu(w;\theta,\lambda)$. We view the natural PG in (7) as a minimizer of $E^\nu(w;\theta,\lambda)$ once we take $\nu(s,a)=d_\rho^{\pi_\theta}(s)\pi_\theta(a\,|\,s)$,

$$(1-\gamma)F_\rho(\theta)^\dagger\cdot\nabla_\theta V_L^{\theta,\lambda}(\mu)\;\in\;\underset{w}{\operatorname{argmin}}\;E^\nu(w;\theta,\lambda)$$

which follows from the first-order optimality condition. If the minimizer has zero compatible function approximation error, we have already established the global convergence in Theorem 1 for the softmax parametrization. However, this is not the case for a general policy class, since it may not include all possible policies, e.g., if we take $d\ll|S||A|$ for the tabular CMDP. The intuition behind *compatibility* is that we can use any one of the minimizers of $E^\nu(w;\theta,\lambda)$ and it does not affect the convergence properties of algorithm; see discussions in [21, 45, 3]. Note that our compatible function approximation error defines over the advantage function for the Lagrangian, which is natural and different from the unconstrained case.

However, ν is an unknown state-action measure of a feasible comparison policy π. To relieve this issue, we introduce an exploratory initial distribution ν_0 over states and actions. We define a state-action visitation distribution $\nu_{\nu_0}^\pi$ of a policy π as $\nu_{\nu_0}^\pi(s,a)=(1-\gamma)\mathbb{E}_{(s_0,a_0)\sim\nu_0}\sum_{t=0}^\infty\gamma^tP^\pi\left(s_t=s,a_t=a\,|\,s_0,a_0\right)$ where $P^\pi\left(s_t=s,a_t=a\,|\,s_0,a_0\right)$ is the probability of visiting state-action (s,a) following policy π from initial state-action (s_0,a_0). We unload notation $\nu_{\nu_0}^{\pi^{(t)}}$ as $\nu^{(t)}$ if it is clear from the context. We now update $w^{(t)}$ in (10) in a general manner,

$$w^{(t)} \in \underset{w}{\operatorname{argmin}}\, E^{\nu^{(t)}}(w;\theta^{(t)},\lambda^{(t)}) := \mathbb{E}_{s,a\sim\nu^{(t)}}\Big[\big(A_L^{\theta^{(t)},\lambda^{(t)}}(s,a) - w\cdot\nabla_\theta \log\pi_\theta(a\,|\,s)\big)^2\Big] \quad (11)$$

in which we assume that the minimizer is computed exactly.

To establish convergence theory, we adopt the standard smoothness assumption [56, 3].

Assumption 2 (Policy Smoothness). *For all $s\in S$, $a\in A$, $\log\pi_\theta(a\,|\,s)$ is a β-smooth of θ, i.e.,*

$$\|\nabla_\theta \log\pi_\theta(a\,|\,s) \,-\, \nabla_{\theta'}\log\pi_\theta(a\,|\,s)\| \;\le\; \beta\,\|\theta\,-\,\theta'\| \;\text{ for all } \theta,\theta'\in\mathbb{R}^d.$$

One example that satisfies Assumption 2 is the linear softmax policy [3]. Thus, Assumption 2 strictly generalizes our previous result for the softmax parametrization (5).

Theorem 2 (Convergence and Optimality: General Parametrization). *Let Assumptions 1 and 2 hold with a policy class $\{\pi_\theta\,|\,\theta\in\Theta\}$. Fix a state distribution ρ, a state-action distribution ν_0, and $T>0$. Let the best feasible policy be $\pi^\star = \pi_\theta^\star$. Define the induced state-action visitation measure under $\pi^\star$: $\nu^\star(s,a) = d_\rho^{\pi^\star}(s)\pi^\star(a\,|\,s)$. Suppose the iterates $\pi^{(t)}$ and $\lambda^{(t)}$ generated by the method (7) with a general primal update (11), $\theta^{(0)}=0$, $\lambda^{(0)}=0$, and $\eta_1=\eta_2=1/\sqrt{T}$ satisfy $E_\star^{\nu^{(t)}}(\theta^{(t)},\lambda^{(t)}) \le \epsilon_{approx}$ and $\|w^{(t)}\|\le W$ for all $0\le t<T$. Then,*

$$\textit{(Optimality gap)}\quad \frac{1}{T}\sum_{t=0}^{T-1}\big(V_r^\star(\rho) - V_r^{(t)}(\rho)\big) \;\le\; \frac{C_1}{(1-\gamma)^3}\frac{1}{\sqrt{T}} + \sqrt{\frac{\epsilon_{approx}}{(1-\gamma)^3}\left\|\frac{\nu^\star}{\nu_0}\right\|_\infty}$$

$$\textit{(Constaint violation)}\quad \left[\frac{1}{T}\sum_{t=0}^{T-1}\big(b - V_g^{(t)}(\rho)\big)\right]_+ \;\le\; \frac{C_2}{(1-\gamma)^2}\frac{1}{T^{1/4}} + \left(\frac{4\epsilon_{approx}}{T(1-\gamma)^3}\left\|\frac{\nu^\star}{\nu_0}\right\|_\infty\right)^{1/4}$$

where $C_1 := 1+\log|A|+\beta W^2$ and $C_2 := \sqrt{3+2\lambda^\star+2\log|A|+\beta W^2}$.

We provide a proof for Theorem 2 in Appendix E. The optimality gap and the constraint violation decays to zero up to the approximation error ϵ_{approx}, an upper bound of the compatible error. Such an error is negligible in the constraint violation due to the factor $1/T^{1/4}$. Theorem 2 generalizes (7) via compatible function approximation that raises a state-action distribution mismatch. The distribution mismatch coefficient $\|\nu^\star/\nu_0\|_\infty$ carries exploration duty in the PG methods [9, 3, 41]. This coefficient can be made finite and small, with an explorative/random enough initial distribution ν_0 [3]. When the error ϵ_{approx} is zero, the distribution mismatch effect disappears, and Theorem 2 is similar to Theorem 1. This is the case for the tabular softmax parametrization or the linear MDP case [3]. The overparametrized neural nets [48] also can lead to small ϵ_{approx}. Moreover, instead of $\|\nu^\star/\nu_0\|_\infty$, it is straightforward to interpret Theorem 2 using the concept of transfer error [3].

Since the strong duality does not necessarily hold, we cannot utilize the previous method to control the constraint violation. Consequently, there is a slightly slower rate that is similar to [29, 22]. We exploit the boundedness of value functions and return to the unparametrized problem (1) via the worst-case analysis. The worst-case gap notwithstanding, we show that the constraint violation enjoys a sublinear rate, and the function approximation error ϵ_{approx} appears as decaying sublinearly.

6 Sample-Based NPG-PD Algorithms

We have assumed access to the exact natural policy gradient in Section 4 or the ability to exactly solve the minimization of the compatible function approximation error in Section 5. In these connections, Theorem 1 and Theorem 2 have established non-asymptotic convergence results. We now leverage our theoretical results to design sample-based algorithms using only empirical estimates, i.e., *model-free*.

We build on the general version of the NPG-PD method (10) to propose a sample-based NPG-PD algorithm with function approximation and $\Lambda=[0,\infty)$,

$$\theta^{(t+1)} \;=\; \theta^{(t)} + \frac{\eta_1}{1-\gamma}\,\widehat{w}^{(t)} \;\text{ and }\; \lambda^{(t+1)} \;=\; \mathcal{P}_\Lambda\big(\lambda^{(t)} - \eta_2\,\big(\widehat{V}_g^{(t)}(\rho)-b\big)\big) \quad (12)$$

where the gradient $\widehat{w}^{(t)}$ and the value function $\widehat{V}_g^{(t)}(\rho)$ are sample-based estimates. We display our algorithm as Algorithm 1 in Appendix F. At each time t, the CMDP environment is executed for K rounds and it terminates with a probability $1-\gamma$ at each round. For the population problem (11), we run SGD K rounds: $w_{k+1} = w_k - \alpha G_k$, to estimate $\widehat{w}^{(t)} = K^{-1}\sum_{k=1}^K w_k$; see [5]. Here, we use the following G_k as an estimate of the population gradient $\nabla_w E^{\nu^{(t)}}(w;\theta^{(t)},\lambda^{(t)})$,

$$G_k \;=\; \left(w_k\cdot\nabla_\theta\log\pi^{(t)}(a\,|\,s)-\widehat{A}_L^{(t)}(s,a)\right)\cdot\nabla_\theta\log\pi^{(t)}(a\,|\,s)$$

where $\widehat{A}_L^{(t)}(s,a)=\widehat{Q}_L^{(t)}(s,a)-\widehat{V}_L^{(t)}(s)$; $\widehat{Q}_L^{(t)}(s,a)$ and $\widehat{V}_L^{(t)}(s)$ are undiscounted sums in each round. In addition, we run another K rounds with initial $s\sim\rho$ to estimate $\widehat{V}_g^{(t)}(s)$ as an undiscounted sum in each round and take the average of K rounds to obtain $\widehat{V}_g^{(t)}(\rho)$. As shown in Appendix F, $\widehat{Q}_L^{(t)}(s,a)$, $\widehat{V}_L^{(t)}(s)$, and $\widehat{V}_g^{(t)}(\rho)$ are unbiased.

To establish the convergence result, we make two assumptions that are standard in the literature [56, 3].

Assumption 3 (Lipschitz Policy). *For $0\le t<T$, the policy $\pi^{(t)}$ satisfies $\left\|\nabla_\theta\log\pi^{(t)}(a\,|\,s)\right\|\le L_\pi$.*

Assumption 4 (Bounded Error and Weight). *Take $w^{(t)}=\operatorname{argmin}_w E^{\nu^{(t)}}(w;\theta^{(t)},\lambda^{(t)})$. For $0\le t<T$, the iterates generated by Algorithm 1 satisfy,*

$$\mathbb{E}\left[E_\star^{\nu^{(t)}}(\theta^{(t)},\lambda^{(t)})\right]\;\le\;\epsilon_{approx},\;\;\mathbb{E}\left[\|\widehat{w}^{(t)}\|^2\right]\;\le\;\widehat{W}^2,\;\;\text{and}\;\;\mathbb{E}\left[\|w^{(t)}\|^2\right]\;\le\;W^2$$

where the expectation is over randomness in $\theta^{(t)}$ and $\lambda^{(t)}$ in Algorithm 1.

Theorem 3 (Sample Complexity: General Parametrization). *Let Assumptions 1, 2, 3, and 4 hold with a policy class $\{\pi_\theta\,|\,\theta\in\Theta\}$. Fix a state distribution ρ, a state-action distribution ν_0, and $T>0$. Let the best feasible policy be $\pi^\star=\pi_\theta^\star$. Define the induced state-action visitation measure under $\pi^\star$: $\nu^\star(s,a)=d_\rho^{\pi^\star}(s)\pi^\star(a\,|\,s)$. Suppose the iterates $\pi^{(t)}$ and $\lambda^{(t)}$ are generated by the sample-based NPG-PD algorithm: Algorithm 1, with $\theta^{(0)}=0$, $\lambda^{(0)}=0$, $\eta_1=\eta_2=1/\sqrt{T}$, and $\alpha=1/L_\pi$, in which K rounds of trajectory samples are used at each time t. Then,*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho)\,-\,V_r^{(t)}(\rho)\right)\right]\;\le\;\frac{C_3}{(1-\gamma)^3}\frac{1}{\sqrt{T}}\;+\;\sqrt{\frac{1}{(1-\gamma)^3}\left\|\frac{\nu^\star}{\nu_0}\right\|_\infty}\left(\sqrt{\epsilon_{approx}}+\frac{C_5}{\sqrt{K}}\right)$$

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b\,-\,V_g^{(t)}(\rho)\right)\right]_+\;\le\;\frac{C_4}{(1-\gamma)^2}\frac{1}{T^{1/4}}+\left(\frac{4}{T(1-\gamma)^3}\left\|\frac{\nu^\star}{\nu_0}\right\|_\infty\right)^{1/4}\left((\epsilon_{approx})^{1/4}+\frac{\sqrt{C_5}}{K^{1/4}}\right)$$

where $C_3:=2+\log|A|+\beta\widehat{W}^2$, $C_4:=2\sqrt{1+\lambda^\star+C_3}$, and $C_5:=2\sqrt{d}\,(WL_\pi+1/(1-\gamma))$.

In Appendix G, we provide a proof for Theorem 3. Theorem 3 describes both the role of function approximation and the sampling effect. As more samples are used, the optimality gap and the constraint violation behave similarly as Theorem 2. The constraint violation is less susceptible to both the function approximation error and the sampling estimation error than the optimality gap.

Moreover, a special case of (12) is a sample-based primal-dual algorithm with softmax parametrization if we take $\widehat{w}^{(t)}=\widehat{A}_L^{(t)}$ and $\Lambda=[0,2/((1-\gamma)\xi)]$. We describe our algorithm as Algorithm 2 in Appendix H and show its sample complexity; a proof is given in Appendix I.

Theorem 4 (Sample Complexity: Softmax Parametrization). *Let Assumption 1 hold for $\xi>0$. Fix a state distribution ρ and $T>0$. Suppose the iterates $\pi^{(t)}$ and $\lambda^{(t)}$ are generated by the sample-based NPG-PD algorithm: Algorithm 2, with $\theta^{(0)}=0$, $\lambda^{(0)}=0$, $\eta_1=2\log|A|$ and $\eta_2=(1-\gamma)/\sqrt{T}$, in which K rounds of trajectory samples are used at each time t. Then,*

$$\textit{(Optimality gap)}\;\;\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho)\,-\,V_r^{(t)}(\rho)\right)\right]\;\le\;\frac{5}{(1-\gamma)^2}\frac{1}{\sqrt{T}}\;+\;\frac{1}{(1-\gamma)K\sqrt{T}}$$

$$\textit{(Constaint violation)}\;\;\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b-V_g^{(t)}(\rho)\right)\right]_+\;\le\;\frac{1/\xi+5\xi}{(1-\gamma)^2}\frac{1}{\sqrt{T}}\;+\;\frac{\xi}{K\sqrt{T}}.$$

For the softmax parametrization, Theorem 4 shows better dependence on T and K than Theorem 3. If there is no sampling effect, the convergence rates match those in Theorem 1. It is noted that this result still has the property of being *dimension-free* for the optimality gap and the constraint violation.

To verify our convergence theory, we provide computational results by simulating the algorithm (8) and its sample-based version: Algorithm 2, for a finite CMDP with random initializations. Given $T>0$, the total number of optimization iterations, our stepsizes in theorems become constants, and multiplying them with positive constants does not affect convergence rates. We generalize the shared MDP code [9] to CMDPs. We first compare the NPG-PD method (8) with the dualDescent [37] that

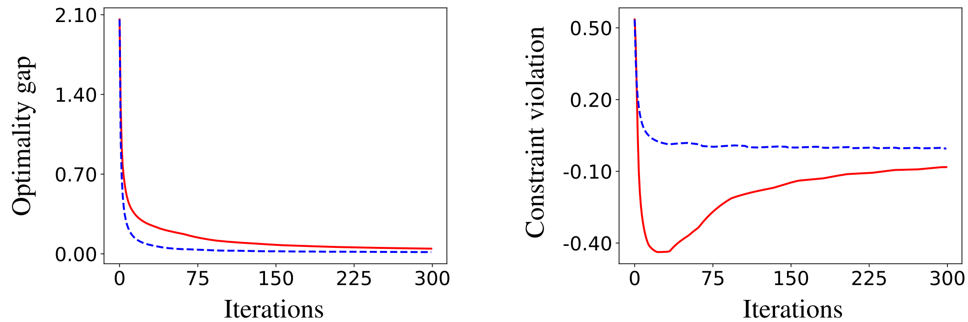

Figure 1: Comparison of the dualDescent [37] (– –) with the NPG-PD method (8) (—). In this experiment, we have randomly generated a CMDP with $|S| = 20$, $|A| = 10$, $\gamma = 0.8$, and $b = 3$, and chosen: $\eta_1 = \eta_2 = 1$.

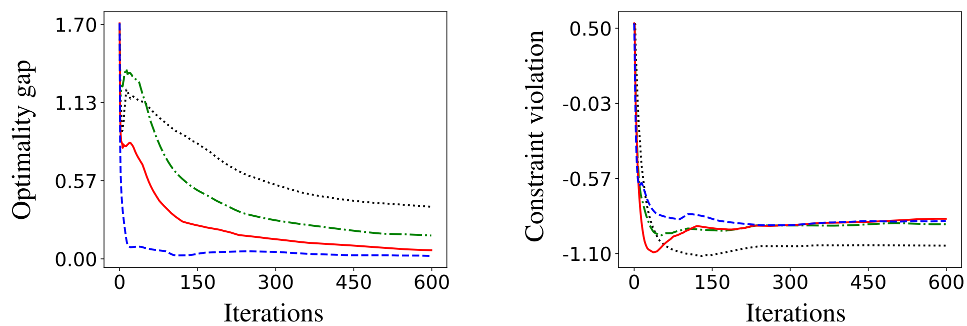

Figure 2: Comparison of the dualDescent [37] (– –) with the sample-based NPG-PD algorithm: Algorithm 2, using different sample sizes: $K = 20$ (····), $K = 50$ (–·–) and $K = 100$ (—). In this experiment, we have randomly generated a CMDP with $|S| = 20$, $|A| = 10$, $\gamma = 0.8$, and $b = 3$, and chosen: $\eta_1 = \eta_2 = 1$ for Algorithm 2, and $\eta = 1$, $K = 100$, and $L = 10$ for the dualDescent.

takes the exact PG method as an RL algorithm. In Figure 1, we see that both optimality gaps decay to zero quickly and our NPG-PD algorithm displays an outstanding constraint satisfaction. We also compare them by using only simulated policy gradients with the sample size K. In Figure 2, a key observation is that our sample-based NPG-PD algorithm performs as the dualDescent for large K. We point out that the dualDescent needs, roughly L times more computation, than our algorithm since at each iteration it takes an extra inner-loop of executing an RL algorithm for L steps. In this sense, our NPG-PD algorithm has better efficiency and is simple to apply without any inner-loop computation. See Appendix J for more experimental results.

7 Conclusion

In this paper, we have proposed an NPG-PD method for CMDPs with the primal natural policy gradient ascent and the dual projected sub-gradient descent. Even though the underlying maximization problem has a nonconcave objective function and a nonconvex constraint set, we provide a systematic study of the non-asymptotic convergence properties of this method with either the softmax parametrization or the general parametrization. We have also proposed two associated sample-based NPG-PD algorithms and established their finite-sample complexity guarantees. Our work is the first to offer non-asymptotic convergence guarantees of policy-based primal-dual methods for solving infinite-horizon discounted CMDPs.

A natural future direction is to investigate how we can achieve a fast rate, e.g., $O(1/T)$, for the NPG-PD method. The rate could be improved by utilizing the standard variance reduction technique. Another important direction is to study the generalization of our results to the vanilla primal-dual method without Fisher preconditioning. Moreover, it is relevant to exploit structure of particular CMDPs in order to provide improved convergence theory.

Broader Impact

Our development could be added to a growing literature of constrained Markov decision processes (CMDPs) in a broad area of safe reinforcement learning (safe RL). Not only aiming to maximize the total reward, but almost all real-world sequential decision-making applications must also take control of safety regarding cost, utility, error rate, or efficiency, e.g., autonomous driving, medical test, financial management, and space exploration. Handling these additional safety objectives leads to constrained decision-making problems. Our research could be used to provide an algorithmic solution for practitioners to solve such constrained problems with non-asymptotic convergence and optimality guarantees. Our methodology could be new knowledge for RL researchers on the direct policy search methods for solving infinite-horizon discounted CMDPs.

The decision-making processes that build on our research could enjoy the flexibility of adding practical constraints and this would improve a large range of uses, e.g., autonomous systems, healthcare services, and financial and legal services. We may expect a broad range of societal implications and we list some of them as follows. The autonomous robotics could be deployed to hazard environments, e.g., forest fires or earthquakes, with added safety guarantees. This could accelerate rescuing while saving robotics. The discovery of medical treatments could be less risky by restraining the side effect. Thus the bias of treatments could be minimized effectively. The policymaker in government or enterprises could encourage economic productivity as much as they can but under law/environment/public health constraints. Overall, one could expect a lot of social welfare improvements supported by the uses of our research.

However, applying any theory to practice has to care about assumption/model mismatches. For example, our theory is in favor of well-defined feasible problems. This usually requires domain knowledge to justify. We would suggest domain experts develop guidelines for assumption/model validation. We would also encourage further work to establish the generalizability to other settings. Another issue could be the bias on gender or race. Policy parametrization selected by biased policymakers may inherit those biases. We would also encourage research to understand and mitigate the biases.

Acknowledgments and Disclosure of Funding

D. Ding and M. R. Jovanović were supported by the National Science Foundation under Awards ECCS-1708906 and ECCS-1809833. Research of K. Zhang and T. Başar was supported in part by the US Army Research Laboratory (ARL) Cooperative Agreement W911NF-17-2-0196, and in part by the Office of Naval Research (ONR) MURI Grant N00014-16-1-2710.

References

[1] 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*, pages 75–84, 2010.

[2] J. Achiam, D. Held, A. Tamar, and P. Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, volume 70, pages 22–31, 2017.

[3] 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.

[4] E. Altman. *Constrained Markov Decision Processes*, volume 7. CRC Press, 1999.

[5] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in neural information processing systems*, pages 773–781, 2013.

[6] Q. Bai, A. Gattami, and V. Aggarwal. Model-free algorithm and regret analysis for MDPs with peak constraints. *arXiv preprint arXiv:2003.05555*, 2020.

[7] A. Beck. *First-order Methods in Optimization*, volume 25. SIAM, 2017.

[8] D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic press, 2014.

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

[10] J. Bhandari and D. Russo. A note on the linear convergence of policy gradient methods. *arXiv preprint arXiv:2007.11120*, 2020.

[11] S. Bhatnagar and K. Lakshmanan. An online actor–critic algorithm with function approximation for constrained Markov decision processes. *Journal of Optimization Theory and Applications*, 153(3):688–708, 2012.

[12] V. S. Borkar. An actor-critic algorithm for constrained Markov decision processes. *Systems & control letters*, 54(3):207–213, 2005.

[13] S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *arXiv preprint arXiv:2007.06558*, 2020.

[14] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[15] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

[16] D. Ding, X. Wei, Z. Yang, Z. Wang, and M. R. Jovanović. Provably efficient safe exploration via primal-dual policy optimization. *arXiv preprint arXiv:2003.00534*, 2020.

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

[18] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476, 2018.

[19] 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*, 64(7):2737–2752, 2018.

[20] S. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, volume 2, pages 267–274, 2002.

[21] S. M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, pages 1531–1538, 2002.

[22] A. Koppel, K. Zhang, H. Zhu, and T. Başar. Projected stochastic primal-dual method for constrained online learning with kernels. *IEEE Transactions on Signal Processing*, 67(10):2528–2542, 2019.

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

[24] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

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

[26] B. Liu, Q. Cai, Z. Yang, and Z. Wang. Neural trust region/proximal policy optimization attains globally optimal policy. In *Advances in Neural Information Processing Systems*, pages 10564–10575, 2019.

[27] Y. Liu, J. Ding, and X. Liu. IPO: Interior-point policy optimization under constraints. *AAAI Conference on Artificial Intelligence*, 2020.

[28] Y. Liu, K. Zhang, T. Başar, and W. Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In *Advances in Neural Information Processing Systems*, 2020.

[29] M. Mahdavi, R. Jin, and T. Yang. Trading regret for efficiency: online convex optimization with long term constraints. *Journal of Machine Learning Research*, 13(Sep):2503–2528, 2012.

[30] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, 2020.

[31] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1928–1937, 2016.

[32] H. Mohammadi, M. Soltanolkotabi, and M. R. Jovanović. On the linear convergence of random search for discrete-time LQR. *IEEE Control Systems Letters*, 5(3):989–994, 2020.

[33] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanović. Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem. *IEEE Transactions on Automatic Control*, 2019. submitted; also arXiv:1912.11899.

[34] M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, pages 14905–14916, 2019.

[35] M. Ono, M. Pavone, Y. Kuwata, and J. Balaram. Chance-constrained dynamic programming with application to risk-aware robotic space exploration. *Autonomous Robots*, 39(4):555–571, 2015.

[36] 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.

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

[38] S. Qiu, X. Wei, Z. Yang, J. Ye, and Z. Wang. Upper confidence primal-dual optimization: Stochastically constrained Markov decision processes with adversarial losses and unknown transitions. *arXiv preprint arXiv:2003.00660*, 2020.

[39] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

[40] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[41] L. Shani, Y. Efroni, and S. Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs. In *AAAI Conference on Artificial Intelligence*, 2020.

[42] R. Singh, A. Gupta, and N. B. Shroff. Learning in Markov decision processes under constraints. *arXiv preprint arXiv:2002.12435*, 2020.

[43] T. Spooner and R. Savani. A natural actor-critic algorithm with downside risk constraints. *arXiv preprint arXiv:2007.04203*, 2020.

[44] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An introduction*. MIT press, 2018.

[45] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1063, 2000.

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

[47] E. Uchibe and K. Doya. Constrained reinforcement learning from intrinsic and extrinsic rewards. In *International Conference on Development and Learning*, pages 163–168, 2007.

[48] L. Wang, Q. Cai, Z. Yang, and Z. Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2019.

[49] X. Wei, H. Yu, and M. J. Neely. Online primal-dual mirror descent under stochastic constraints. *International Conference on Measurement and Modeling of Computer Systems*, 2020.

[50] J. Yang, N. Kiyavash, and N. He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2002.09621*, 2020.

[51] T.-Y. Yang, J. Rosca, K. Narasimhan, and P. J. Ramadge. Projection-based constrained policy optimization. In *International Conference on Learning Representations*, 2020.

[52] H. Yu, M. Neely, and X. Wei. Online convex optimization with stochastic constraints. In *Advances in Neural Information Processing Systems*, pages 1428–1438, 2017.

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

[54] J. Yuan and A. Lamperski. Online convex optimization for cumulative constraints. In *Advances in Neural Information Processing Systems*, pages 6137–6146, 2018.

[55] J. Zhang, A. Koppel, A. S. Bedi, C. Szepesvari, and M. Wang. Variational policy gradient method for reinforcement learning with general utilities. *arXiv preprint arXiv:2007.02151*, 2020.

[56] K. Zhang, A. Koppel, H. Zhu, and T. Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 2019.

[57] K. Zhang, Z. Yang, and T. Başar. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems*, pages 11598–11610, 2019.

[58] X. Zhang, K. Zhang, E. Miehling, and T. Başar. Non-cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 9482–9493, 2019.

[59] L. Zheng and L. J. Ratliff. Constrained upper confidence reinforcement learning. In *Conference on Learning for Dynamics and Control*, 2020.

Supplementary Materials for "Natural Policy Gradient Primal-Dual Method for Constrained Markov Decision Processes"

A Proof of Lemma 3

We prove Lemma 3 by providing a concrete CMDP example as shown in Figure 3. States s_3, s_4, and s_5 are terminal states with zero reward and utility. We consider non-trivial state s_1 with two actions: a_1 moving 'up' and a_2 going 'right', and the associated value functions are given by

$$\begin{aligned} V_r^\pi(s_1) &= \pi(a_2 \,|\, s_1)\pi(a_1 \,|\, s_2) \\ V_g^\pi(s_1) &= \pi(a_1 \,|\, s_1) + \pi(a_2 \,|\, s_1)\pi(a_1 \,|\, s_2). \end{aligned}$$

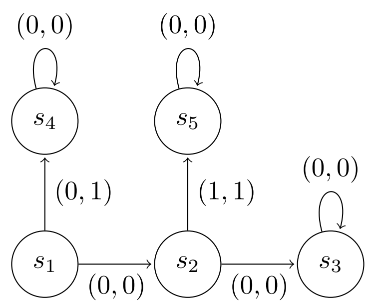

Figure 3: An example of CMDP in the proof of Lemma 3 where $V_r^{\pi_\theta}(s)$ is nonconcave and the set $\{\theta \in \Theta \,|\, V_g^{\pi_\theta}(s) \geq b\}$ is not convex. The pair (r, g) alongside the arrow depicts reward r and utility g of taking an action at certain state.

We consider the following two policies $\pi^{(1)}$ and $\pi^{(2)}$ using the softmax parametrization (5),

$$\begin{aligned} \theta^{(1)} &= (\log 1, \log x, \log x, \log 1) \\ \theta^{(2)} &= (-\log 1, -\log x, -\log x, -\log 1) \end{aligned}$$

where the parameter takes form of $(\theta_{s_1,a_1}, \theta_{s_1,a_2}, \theta_{s_2,a_1}, \theta_{s_2,a_2})$ with $x > 0$.

First, we show that V_r^π is not concave. We compute that

$$\begin{aligned} \pi^{(1)}(a_1 \,|\, s_1) &= \frac{1}{1+x}, \; \pi^{(1)}(a_2 \,|\, s_1) = \frac{x}{1+x}, \; \pi^{(1)}(a_1 \,|\, s_2) = \frac{x}{1+x} \\ V_r^{(1)}(s_1) &= \left(\frac{x}{1+x}\right)^2, \; V_g^{(1)}(s_1) = \frac{1+x+x^2}{(1+x)^2} \\ \pi^{(2)}(a_1 \,|\, s_1) &= \frac{x}{1+x}, \; \pi^{(2)}(a_2 \,|\, s_1) = \frac{1}{1+x}, \; \pi^{(2)}(a_1 \,|\, s_2) = \frac{1}{1+x} \\ V_r^{(2)}(s_1) &= \left(\frac{1}{1+x}\right)^2, \; V_g^{(2)}(s_1) = \frac{1+x+x^2}{(1+x)^2}. \end{aligned}$$

Now, we consider policy $\pi^{(\zeta)}$,

$$\zeta\,\theta^{(1)} + (1-\zeta)\,\theta^{(2)} = \left(\log 1, \log\left(x^{2\zeta-1}\right), \log\left(x^{2\zeta-1}\right), \log 1\right)$$

for some $\zeta \in [0,1]$, which is defined on the segment between $\theta^{(1)}$ and $\theta^{(2)}$. Therefore,

$$\begin{aligned} \pi^{(1)}(a_1 \,|\, s_1) &= \frac{1}{1+x^{2\zeta-1}}, \; \pi^{(1)}(a_2 \,|\, s_1) = \frac{x^{2\zeta-1}}{1+x^{2\zeta-1}}, \; \pi^{(1)}(a_1 \,|\, s_2) = \frac{x^{2\zeta-1}}{1+x^{2\zeta-1}} \\ V_r^{(\zeta)}(s_1) &= \left(\frac{x^{2\zeta-1}}{1+x^{2\zeta-1}}\right)^2, \; V_g^{(\zeta)}(s_1) = \frac{1+x^{2\zeta-1}+(x^{2\zeta-1})^2}{(1+x^{2\zeta-1})^2}. \end{aligned}$$

When $x = 3$ and $\zeta = \frac{1}{2}$,

$$\frac{1}{2}V_r^{(1)}(s_1) + \frac{1}{2}V_r^{(2)}(s_1) = \frac{5}{16} > V_r^{(\frac{1}{2})}(s_1) = \frac{4}{16}$$

which implies that V_r^π is not concave.

When $x = 10$ and $\zeta = \frac{1}{2}$,

$$V_g^{(1)}(s_1) \;=\; V_g^{(2)}(s_1) \;\geq\; 0.9 \;\text{ and }\; V_g^{(\frac{1}{2})}(s_1) \;=\; 0.75$$

which shows that if we take constraint offset $b = 0.9$, then $V_g^{(1)}(s_1) = V_g^{(2)}(s_1) \geq b$, and $V_g^{(\frac{1}{2})}(s_1) < b$ in which the policy $\pi^{(\frac{1}{2})}$ is infeasible. Therefore, the set $\{\theta \,|\, V_g^{\pi_\theta}(s) \geq b\}$ is not convex.

B Proof of Lemma 4

The dual update is based on Lemma 1. Since $\lambda^\star \leq \left(V_r^\star(\rho) - V_r^{\bar\pi}(\rho)\right)/\xi$ with $0 \leq V_r^\star, V_r^{\bar\pi} \leq \frac{1}{1-\gamma}$, we take projection interval $\Lambda = [0, \frac{2}{(1-\gamma)\xi}]$ such that upper bound $\frac{2}{(1-\gamma)\xi}$ is such that $\frac{2}{(1-\gamma)\xi} \geq 2\lambda^\star$.

We now verify the primal update. We expand the primal update in (7) into the following form,

$$\theta^{(t+1)} \;=\; \theta^{(t)} \;+\; \eta_1 F_\rho(\theta^{(t)})^\dagger \cdot \nabla_\theta V_r^{\theta^{(t)}}(\rho) \;+\; \eta_1 \lambda^{(t)} F_\rho(\theta^{(t)})^\dagger \cdot \nabla_\theta V_g^{\theta^{(t)}}(\rho). \tag{13}$$

We now deal with: $F_\rho(\theta^{(t)})^\dagger \cdot \nabla_\theta V_r^{\theta^{(t)}}(\rho)$ and $F_\rho(\theta^{(t)})^\dagger \cdot \nabla_\theta V_g^{\theta^{(t)}}(\rho)$. For the first one, the proof begins with solutions to the following approximation error minimization problem:

$$\underset{w \,\in\, \mathbb{R}^{|S||A|}}{\text{minimize}} \;\; E_r(w) \;:=\; \mathbb{E}_{s \,\sim\, d_\rho^{\pi_\theta}, a \,\sim\, \pi_\theta(a\,|\,s)} \left[\left(A_r^{\pi_\theta}(s,a) - w \cdot \nabla_\theta \log \pi_\theta(a\,|\,s)\right)^2\right].$$

Using the Moore-Penrose inverse, the optimal solution reads,

$$w_r^\star \;=\; F_\rho(\theta)^\dagger \mathbb{E}_{s \,\sim\, d_\rho^{\pi_\theta}, a \,\sim\, \pi_\theta(a\,|\,s)} \left[\nabla_\theta \log \pi_\theta(a\,|\,s)\, A_r^{\pi_\theta,\lambda}(s,a)\right] \;=\; (1-\gamma) F_\rho(\theta)^\dagger \cdot \nabla_\theta V_r^{\pi_\theta,\lambda}(\rho)$$

where $F_\rho(\theta)$ is the Fisher information matrix induced by π_θ. One key observation from this solution is that $w_r^\star$ is parallel to the natural PG direction $F_\rho(\theta)^\dagger \cdot \nabla_\theta V_r^{\pi_\theta,\lambda}(\rho)$.

On the other hand, it is easy to verify that $A_r^{\pi_\theta}$ is a minimizer of $E_r(w)$. The softmax parametrization (5) implies that

$$\frac{\partial \log \pi_\theta(a\,|\,s)}{\partial \theta_{s',a'}} \;=\; \mathbb{I}\{s = s'\}\left(\mathbb{I}\{a = a'\} - \pi_\theta(a'\,|\,s)\right) \tag{14}$$

where $\mathbb{I}\{E\}$ is the indicator function of event E being true. Thus, we have

$$w \cdot \nabla_\theta \log \pi_\theta(a\,|\,s) \;=\; w_{s,a} \;-\; \sum_{a' \,\in\, A} w_{s,a'} \pi_\theta(a'\,|\,s).$$

The above equality together with the fact: $\sum_{a \,\in\, A} \pi_\theta(a\,|\,s) A_r^{\pi_\theta,\lambda}(s,a) = 0$, shows that $E_r(A_r^{\pi_\theta}) = 0$. However, $A_r^{\pi_\theta}$ may not be the unique minimizer. We consider the following general form of possible solutions,

$$A_r^{\pi_\theta} \;+\; u, \;\; \text{where} \;\; u \;\in\; \mathbb{R}^{|S||A|}.$$

For any state s and action a such that s is reachable under ρ, using (14) yields

$$u \cdot \nabla_\theta \log \pi_\theta(a\,|\,s) \;=\; u_{s,a} \;-\; \sum_{a' \,\in\, A} u_{s,a'} \pi_\theta(a'\,|\,s).$$

Here, we make use of the following fact: π_θ is a stochastic policy with $\pi_\theta(a\,|\,s) > 0$ for all actions a in each state s, so that if a state is reachable under ρ, then it will also be reachable using π_θ. Therefore, we require zero derivative at each reachable state:

$$u \cdot \nabla_\theta \log \pi_\theta(a\,|\,s) \;=\; 0$$

for all s, a so that $u_{s,a}$ is independent of the action and becomes a constant c_s for each s. Therefore, the minimizer of $E_r(w)$ is given up to some state-dependent offset,

$$F_\rho(\theta)^\dagger \cdot \nabla_\theta V_r^{\pi_\theta}(\rho) \;=\; \frac{A_r^{\pi_\theta}}{1-\gamma} \;+\; u \tag{15}$$

where $u_{s,a} = c_s$ for some $c_s \in \mathbb{R}$ for each state s and action a.

We can repeat the above procedure for $F_\rho(\theta^{(t)})^\dagger \nabla_\theta V_g^{\theta^{(t)}}(\rho)$ and show,

$$F_\rho(\theta)^\dagger \cdot \nabla_\theta V_g^{\pi_\theta}(\rho) \;=\; \frac{A_g^{\pi_\theta}}{1-\gamma} \;+\; v \tag{16}$$

where $v_{s,a} = d_s$ for some $d_s \in \mathbb{R}$ for each state s and action a.

Substituting (15) and (16) into the primal update (13) yields,

$$\theta^{(t+1)} \;=\; \theta^{(t)} \;+\; \frac{\eta_1}{1-\gamma}\,\left(A_r^{(t)} + \lambda^{(t)} A_g^{(t)}\right) \;+\; \eta_1\left(u + \lambda^{(t)} v\right)$$

$$\pi^{(t+1)}(a\,|\,s) \;=\; \pi^{(t)}(a\,|\,s)\,\frac{\exp\left(\frac{\eta_1}{1-\gamma}\left(A_r^{(t)}(s,a) + \lambda^{(t)} A_g^{(t)}(s,a)\right) + \eta_1\left(c_s + \lambda^{(t)} d_s\right)\right)}{Z^{(t)}(s)}$$

where the second equality also utilizes the normalization term $Z^{(t)}(s)$. Finally, we complete the proof by setting $c_s = d_s = 0$.

C Supporting Results from Optimization

We collect some optimization results from the literature for readers' convenience.

It is noted that all these results hold for the parametric setting of (3) and (4) if the parametrized policy class is *complete*, e.g., the closure of the softmax policy class (5). To rephrase them for our general purpose, we recall the maximization problem (1),

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

in which we maximize over all policies and $b \in (0, 1/(1-\gamma))$ with $\gamma \in [0,1)$. Let the optimal solution be $\pi^\star$ such that

$$V_r^{\pi^\star}(\rho) \;=\; \underset{\pi\,\in\,\Pi}{\text{maximize}}\;\{\, V_r^\pi(\rho)\,|\, V_g^\pi(\rho) \;\geq\; b\,\}.$$

Let the Lagrangian be $V_L^{\pi,\lambda}(\rho) := V_r^\pi(\rho) + \lambda(V_g^\pi(\rho) - b)$, where $\lambda \geq 0$ is the Lagrange multiplier or dual variable. The associated dual function is defined as

$$V_D^\lambda(\rho) \;:=\; \underset{\pi\,\in\,\Pi}{\text{maximize}}\;\; V_L^{\pi,\lambda}(\rho) := V_r^\pi(\rho) + \lambda\,(V_g^\pi(\rho) - b)$$

and the optimal dual is $\lambda^\star = \text{argmin}_{\lambda\,\geq\,0}\, V_D^\lambda(\rho)$,

$$V_D^{\lambda^\star}(\rho) \;:=\; \underset{\lambda\,\geq\,0}{\text{minimize}}\;\; V_D^\lambda(\rho)$$

We recall that the problem (1) enjoys strong duality under the Slater condition [36, Proposition 1].

Assumption 5 (Slater condition)**.** *There exists $\xi > 0$ and $\bar\pi$ such that $V_g^{\bar\pi}(\rho) - b \geq \xi$.*

We use the shorthand notation $V_r^{\pi^\star}(\rho) = V_r^\star(\rho)$ and $V_D^{\lambda^\star}(\rho) = V_D^\star(\rho)$ whenever it is clear from the context.

Lemma 5 (Strong duality)**.** *[36, Proposition 1] If the Slater condition holds, then the strong duality holds,*

$$V_r^\star(\rho) \;=\; V_D^\star(\rho).$$

It is implied by the strong duality that the optimal solution to the dual problem: $\text{minimize}_{\lambda\,\geq\,0}\;\; V_D^\lambda(\rho)$ is obtained at $\lambda^\star$. Denote the set of all optimal dual variables as $\Lambda^\star$.

Under the Slater condition, a useful property of the dual variable is that the sublevel sets are bounded [7, Section 8.5]. Although our problem is nonconcave, we customize it as follows.

Theorem 5 (Boundedness of sublevel sets of the dual function)**.** *Let the Slater condition hold. Fix $C_\lambda \in \mathbb{R}$. For any $\lambda \in \{\lambda \geq 0\,|\, V_D^\lambda(\rho) \leq C_\lambda\}$, it holds that*

$$\lambda \;\leq\; \frac{1}{\xi}\left(C_\lambda - V_r^{\bar\pi}(\rho)\right).$$

Proof. If $\lambda \in \{\lambda \geq 0\,|\, V_D^\lambda(\rho) \leq C_\lambda\}$, then,

$$C_\lambda \;\geq\; V_D^\lambda(\rho) \;\geq\; V_r^{\bar\pi}(\rho) + \lambda\,(V_g^{\bar\pi}(\rho) - b) \;\geq\; V_r^{\bar\pi}(\rho) + \lambda\,\xi$$

where we utilize the Slater point $\bar\pi$ in the last inequality. We complete the proof by noting $\xi > 0$. □

Corollary 1. *If we take $C_\lambda = V_r^\star(\rho) = V_D^\star$, then $\Lambda^\star = \{\lambda \geq 0\,|\, V_D^\lambda(\rho) \leq C_\lambda\}$. Thus, for any $\lambda \in \Lambda^\star$,*

$$\lambda \;\leq\; \frac{1}{\xi}\left(V_r^{\pi^\star}(\rho) - V_r^{\bar\pi}(\rho)\right).$$

Another useful theorem from the convex optimization [7, Section 3.5] is given as follows. It describes that the constraint violation $b - V_g^{\pi}(\rho)$ can be bounded similarly even if we have some weak bound. We next state and prove it for our problem, which is used in our constraint violation analysis.

Theorem 6. *Let the Slater condition hold and $\lambda^\star \in \Lambda^\star$. Let $C_{\lambda^\star} \geq 2\lambda^\star$. Assume that $\widetilde{\pi} \in \Pi$ satisfies*

$$V_r^\star(\rho) \,-\, V_r^{\widetilde{\pi}}(\rho) \,+\, C_{\lambda^\star} \left[b - V_g^{\widetilde{\pi}}(\rho) \right]_+ \;\leq\; \delta.$$

Then,

$$\left[b - V_g^{\widetilde{\pi}}(\rho) \right]_+ \;\leq\; \frac{2\delta}{C_{\lambda^\star}}$$

where $[x]_+ = \max(x, 0)$.

Proof. Let

$$v(\tau) \;=\; \underset{\pi \,\in\, \Pi}{\text{maximize}} \,\{\, V_r^{\pi}(\rho) \,|\, V_g^{\pi}(\rho) \geq b + \tau \,\}.$$

By the definition of $v(\tau)$, we have $v(0) = V_r^\star(\rho)$. We note the proof of [36, Proposition 1] that $v(\tau)$ is concave. First, we show that $-\lambda^\star \in \partial v(0)$. By the definition of Lagrangian $V_L^{\pi,\lambda}(\rho)$ and the strong duality,

$$V_L^{\pi,\lambda^\star}(\rho) \;\leq\; \underset{\pi \,\in\, \Pi}{\text{maximize}}\; V_L^{\pi,\lambda^\star}(\rho) \;=\; V_D^\star(\rho) \;=\; V_r^\star(\rho) \;=\; v(0), \;\; \text{for all } \pi \in \Pi.$$

Hence, for any $\pi \in \{\pi \in \Pi \,|\, V_g^{\pi}(\rho) \geq b + \tau\}$, we have

$$\begin{aligned} v(0) - \tau\lambda^\star \;&\geq\; V_L^{\pi,\lambda^\star}(\rho) - \tau\lambda^\star \\ &=\; V_r^{\pi}(\rho) + \lambda^\star(V_g^{\pi}(\rho) - b) - \tau\lambda^\star \\ &=\; V_r^{\pi}(\rho) + \lambda^\star(V_g^{\pi}(\rho) - b - \tau) \\ &\geq\; V_r^{\pi}(\rho). \end{aligned}$$

If we maximize the right-hand side of above inequality over $\pi \in \{\pi \in \Pi \,|\, V_g^{\pi}(\rho) \geq b + \tau\}$, then

$$v(0) - \tau\lambda^\star \;\geq\; v(\tau) \tag{17}$$

which show that $-\lambda^\star \in \partial v(0)$.

On the other hand, if we take $\tau = \widetilde{\tau} := -(b - V_g^{\widetilde{\pi}}(\rho))_+$, then

$$V_r^{\widetilde{\pi}}(\rho) \;\leq\; V_r^\star(\rho) \;=\; v(0) \;\leq\; v(\widetilde{\tau}). \tag{18}$$

Combing (17) and (18) yields

$$V_r^{\widetilde{\pi}}(\rho) - V_r^\star(\rho) \leq -\widetilde{\tau}\lambda^\star.$$

Thus,

$$\begin{aligned} (C_{\lambda^\star} - \lambda^\star)\,|\widetilde{\tau}| \;&=\; -\lambda^\star\,|\widetilde{\tau}| + C_{\lambda^\star}\,|\widetilde{\tau}| \\ &=\; \widetilde{\tau}\lambda^\star + C_{\lambda^\star}\,|\widetilde{\tau}| \\ &\leq\; V_r^\star(\rho) - V_r^{\widetilde{\pi}}(\rho) + C_{\lambda^\star}\,|\widetilde{\tau}|\,. \end{aligned}$$

By our assumption and $\widetilde{\tau} = \left[b - V_g^{\widetilde{\pi}}(\rho) \right]_+$,

$$\left[b - V_g^{\widetilde{\pi}}(\rho) \right]_+ \;\leq\; \frac{\delta}{C_{\lambda^\star} - \lambda^\star} \;\leq\; \frac{2\delta}{C_{\lambda^\star}}.$$

□

D Proof of Theorem 1

We warm-up with an improvement lemma, stating a difference for two consecutive policies.

Lemma 6 (Non-monotonic Improvement)**.** *The iterates $\pi^{(t)}$ generated by algorithm (8) satisfy*

$$V_r^{(t+1)}(\mu) \,-\, V_r^{(t)}(\mu) \,+\, \lambda^{(t)}\big(V_g^{(t+1)}(\mu) \,-\, V_g^{(t)}(\mu)\big) \;\geq\; \frac{1-\gamma}{\eta_1}\, \mathbb{E}_{s\,\sim\,\mu} \log Z^{(t)}(s) \tag{19}$$

and $\mathbb{E}_{s\,\sim\,\mu} \log Z^{(t)}(s) \geq 0$ for any initial state distributions μ, where notation $d_\mu^{(t+1)}$ means $d_\mu^{\pi^{(t+1)}}$.

Proof. To prove our main inequality, we first apply the performance difference lemma as follows:

$$\begin{aligned}
V_r^{(t+1)}(\mu) \,-\, V_r^{(t)}(\mu) &= \frac{1}{1-\gamma}\,\mathbb{E}_{s\,\sim\, d_\mu^{(t+1)},a\,\sim\,\pi^{(t+1)}(\cdot\,|\,s)}\left[A_r^{(t)}(s,a)\right] \\
&= \frac{1}{1-\gamma}\,\mathbb{E}_{s\,\sim\, d_\mu^{(t+1)}}\left[\sum_{a\,\in\, A}\pi^{(t+1)}(a\,|\,s)A_r^{(t)}(s,a)\right] \\
&= \frac{1}{\eta_1}\,\mathbb{E}_{s\,\sim\, d_\mu^{(t+1)}}\left[\sum_{a\,\in\, A}\pi^{(t+1)}(a\,|\,s)\log\left(\frac{\pi^{(t+1)}(a\,|\,s)}{\pi^{(t)}(a\,|\,s)}Z^{(t)}(s)\right)\right] \\
&\quad -\frac{\lambda^{(t)}}{1-\gamma}\,\mathbb{E}_{s\,\sim\, d_\mu^{(t+1)}}\left[\sum_{a\,\in\, A}\pi^{(t+1)}(a\,|\,s)A_g^{(t)}(s,a)\right] \\
&= \frac{1}{\eta_1}\,\mathbb{E}_{s\,\sim\, d_\mu^{(t+1)}}\left[D_{\rm KL}\left(\pi^{(t+1)}(a\,|\,s)\,\|\,\pi^{(t)}(a\,|\,s)\right)\right] \\
&\quad +\frac{1}{\eta_1}\,\mathbb{E}_{s\,\sim\, d_\mu^{(t+1)}}\log Z^{(t)}(s) \\
&\quad -\frac{\lambda^{(t)}}{1-\gamma}\,\mathbb{E}_{s\,\sim\, d_\mu^{(t+1)}}\left[\sum_{a\,\in\, A}\pi^{(t+1)}(a\,|\,s)A_g^{(t)}(s,a)\right] \\
&\geq \frac{1}{\eta_1}\,\mathbb{E}_{s\,\sim\, d_\mu^{(t+1)}}\log Z^{(t)}(s) \\
&\quad -\frac{\lambda^{(t)}}{1-\gamma}\,\mathbb{E}_{s\,\sim\, d_\mu^{(t+1)}}\left[\sum_{a\,\in\, A}\pi^{(t+1)}(a\,|\,s)A_g^{(t)}(s,a)\right] \\
&= \frac{1}{\eta_1}\,\mathbb{E}_{s\,\sim\, d_\mu^{(t+1)}}\log Z^{(t)}(s) \,-\, \lambda^{(t)}\left(V_g^{(t+1)}(\mu)\,-\,V_g^{(t)}(\mu)\right)
\end{aligned}$$

where the first two equalities are clear from definitions, the third equality is due to the multiplicative weights update in (8), the fourth equality utilizes the Kullback–Leibler divergence or relative entropy between two distributions p, q: $D_{\rm KL}(p\,\|\,q) := \mathbb{E}_{x\,\sim\,p}\log\frac{p(x)}{q(x)}$, we drop a nonnegative term in the inequality, and the last equality is due to the performance difference lemma again. Finally, we obtain the desired inequality by noting $d_\mu^{(t+1)} \geq (1-\gamma)\mu$ componentwise from (6).

It is easy to show that $\log Z^{(t)}(s) \geq 0$.

$$\begin{aligned}
\log Z^{(t)}(s) &= \log\left(\sum_{a\in A}\pi^{(t)}(a\,|\,s)\exp\left(\frac{\eta_1}{1-\gamma}\left(A_r^{(t)}(s,a)+\lambda^{(t)}A_g^{(t)}(s,a)\right)\right)\right) \\
&\geq \sum_{a\in A}\pi^{(t)}(a\,|\,s)\log\left(\exp\left(\frac{\eta_1}{1-\gamma}\left(A_r^{(t)}(s,a)+\lambda^{(t)}A_g^{(t)}(s,a)\right)\right)\right) \\
&= \frac{\eta_1}{1-\gamma}\sum_{a\in A}\pi^{(t)}(a\,|\,s)\left(A_r^{(t)}(s,a)+\lambda^{(t)}A_g^{(t)}(s,a)\right) \\
&= \frac{\eta_1}{1-\gamma}\sum_{a\in A}\pi^{(t)}(a\,|\,s)A_r^{(t)}(s,a)+\frac{\eta_1}{1-\gamma}\lambda^{(t)}\sum_{a\in A}\pi^{(t)}(a\,|\,s)A_g^{(t)}(s,a) \\
&= 0
\end{aligned}$$

In the above inequality, we apply the Jensen's inequality to the concave function $\log(x)$. The last equality is due to

$$\sum_{a\in A}\pi^{(t)}(a\,|\,s)A_r^{(t)}(s,a) \;=\; \sum_{a\in A}\pi^{(t)}(a\,|\,s)A_g^{(t)}(s,a) \;=\; 0.$$

□

Next, we prove the average difference to the optimal policy.

Lemma 7 (Bounded Average Performance)**.** *Let Assumption 1 hold. Fix $T>0$, $\rho\in\Delta_S$, $\theta^{(0)}=0$, and $\lambda^{(0)}=0$. Then the iterates $\pi^{(t)}$ and $\lambda^{(t)}$ generated by algorithm (8) satisfy*

$$\frac{1}{T}\sum_{t\,=\,0}^{T-1}\left(V_r^\star(\rho)-V_r^{(t)}(\rho)\right)+\frac{1}{T}\sum_{t\,=\,0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho)-V_g^{(t)}(\rho)\right) \;\leq\; \frac{\log|A|}{\eta_1 T}+\frac{1}{(1-\gamma)^2T}+\frac{2\eta_2}{(1-\gamma)^3}.$$

Proof. Since ρ is fixed, we unload notation $d_\rho^{\pi^\star}$ as $d^\star$. We first apply the performance difference lemma as follows:

$$
\begin{aligned}
V_r^\star(\rho) \,-\, V_r^{(t)}(\rho) \;&=\; \frac{1}{1-\gamma}\,\mathbb{E}_{s\,\sim\,d^\star}\left[\sum_{a\,\in\,A}\pi^\star(a\,|\,s)A_r^{(t)}(s,a)\right]\\
&=\; \frac{1}{\eta_1}\,\mathbb{E}_{s\,\sim\,d^\star}\left[\sum_{a\,\in\,A}\pi^\star(a\,|\,s)\log\left(\frac{\pi^{(t+1)}(a\,|\,s)}{\pi^{(t)}(a\,|\,s)}Z^{(t)}(s)\right)\right]\\
&\quad\; -\frac{\lambda^{(t)}}{1-\gamma}\,\mathbb{E}_{s\,\sim\,d^\star}\left[\sum_{a\,\in\,A}\pi^\star(a\,|\,s)A_g^{(t)}(s,a)\right]\\
&=\; \frac{1}{\eta_1}\,\mathbb{E}_{s\,\sim\,d^\star}\left[D_{\mathrm{KL}}\left(\pi^\star(a\,|\,s)\,\|\,\pi^{(t)}(a\,|\,s)\right)-D_{\mathrm{KL}}\left(\pi^\star(a\,|\,s)\,\|\,\pi^{(t+1)}(a\,|\,s)\right)\right]\\
&\quad\; +\frac{1}{\eta_1}\,\mathbb{E}_{s\,\sim\,d^\star}\log Z^{(t)}(s) \,-\, \frac{\lambda^{(t)}}{1-\gamma}\,\mathbb{E}_{s\,\sim\,d^\star}\left[\sum_{a\,\in\,A}\pi^\star(a\,|\,s)A_g^{(t)}(s,a)\right]\\
&=\; \frac{1}{\eta_1}\,\mathbb{E}_{s\,\sim\,d^\star}\left[D_{\mathrm{KL}}\left(\pi^\star(a\,|\,s)\,\|\,\pi^{(t)}(a\,|\,s)\right)-D_{\mathrm{KL}}\left(\pi^\star(a\,|\,s)\,\|\,\pi^{(t+1)}(a\,|\,s)\right)\right]\\
&\quad\; +\frac{1}{\eta_1}\,\mathbb{E}_{s\,\sim\,d^\star}\log Z^{(t)}(s) \,-\, \lambda^{(t)}\left(V_g^\star(\rho)\,-\,V_g^{(t)}(\rho)\right)
\end{aligned} \tag{20}
$$

where the second equality is due to the multiplicative weights update in (8), the third equality utilizes the Kullback–Leibler divergence or relative entropy between two distributions p, q: $D_{\mathrm{KL}}(p\,\|\,q) := \mathbb{E}_{x\,\sim\,p}\log\frac{p(x)}{q(x)}$, and the last equality is due to the performance difference lemma again.

According to Lemma 6, if we choose $\mu = d^\star$, then,

$$
V_r^{(t+1)}(d^\star)\,-\,V_r^{(t)}(d^\star)\,+\,\lambda^{(t)}\left(V_g^{(t+1)}(d^\star)\,-\,V_g^{(t)}(d^\star)\right)\;\geq\;\frac{1-\gamma}{\eta_1}\,\mathbb{E}_{s\,\sim\,d^\star}\log Z^{(t)}(s). \tag{21}
$$

Therefore, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t\,=\,0}^{T-1}\left(V_r^\star(\rho)\,-\,V_r^{(t)}(\rho)\right)\\
&=\; \frac{1}{\eta_1 T}\sum_{t\,=\,0}^{T-1}\mathbb{E}_{s\,\sim\,d^\star}\left[D_{\mathrm{KL}}\left(\pi^\star(a\,|\,s)\,\|\,\pi^{(t)}(a\,|\,s)\right)-D_{\mathrm{KL}}\left(\pi^\star(a\,|\,s)\,\|\,\pi^{(t+1)}(a\,|\,s)\right)\right]\\
&\quad\; +\frac{1}{\eta_1 T}\sum_{t\,=\,0}^{T-1}\mathbb{E}_{s\,\sim\,d^\star}\log Z^{(t)}(s)\,-\,\frac{1}{T}\sum_{t\,=\,0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho)\,-\,V_g^{(t)}(\rho)\right)\\
&\leq\; \frac{1}{\eta_1 T}\sum_{t\,=\,0}^{T-1}\mathbb{E}_{s\,\sim\,d^\star}\left[D_{\mathrm{KL}}\left(\pi^\star(a\,|\,s)\,\|\,\pi^{(t)}(a\,|\,s)\right)-D_{\mathrm{KL}}\left(\pi^\star(a\,|\,s)\,\|\,\pi^{(t+1)}(a\,|\,s)\right)\right]\\
&\quad\; +\frac{1}{(1-\gamma)T}\sum_{t\,=\,0}^{T-1}\left(V_r^{(t+1)}(d^\star)\,-\,V_r^{(t)}(d^\star)\right)\\
&\quad\; +\frac{1}{(1-\gamma)T}\sum_{t\,=\,0}^{T-1}\lambda^{(t)}\left(V_g^{(t+1)}(d^\star)\,-\,V_g^{(t)}(d^\star)\right)\,-\,\frac{1}{T}\sum_{t\,=\,0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho)\,-\,V_g^{(t)}(\rho)\right)\\
&\leq\; \frac{1}{\eta_1 T}\,\mathbb{E}_{s\,\sim\,d^\star}D_{\mathrm{KL}}\left(\pi^\star(a\,|\,s)\,\|\,\pi^{(0)}(a\,|\,s)\right)\,+\,\frac{1}{(1-\gamma)T}V_r^{(T)}(d^\star)\,+\,\frac{2\eta_2}{(1-\gamma)^3}\\
&\quad\; -\frac{1}{T}\sum_{t\,=\,0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho)\,-\,V_g^{(t)}(\rho)\right)
\end{aligned} \tag{22}
$$

where in the second inequality we take telescoping sums for the first two sums and drop all non-positive terms: $D_{\mathrm{KL}}\left(\pi^\star(a\,|\,s)\,\|\,\pi^{(T)}(a\,|\,s)\right)$, $V_r^{(0)}(d^\star)$; we utilize the following result with $\mu = d^\star$

to the third sum,

$$
\begin{aligned}
& \frac{1}{T}\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^{(t+1)}(\mu)\,-\,V_g^{(t)}(\mu)\right) \\
= \;& \frac{1}{T}\sum_{t=0}^{T-1}\left(\lambda^{(t+1)}V_g^{(t+1)}(\mu)\,-\,\lambda^{(t)}V_g^{(t)}(\mu)\right)\,+\,\frac{1}{T}\sum_{t=0}^{T-1}\left(\lambda^{(t)}\,-\,\lambda^{(t+1)}\right)V_g^{(t+1)}(\mu) \\
\leq \;& \frac{1}{T}\lambda^{(T)}V_g^{(T)}(\mu)\,+\,\frac{1}{T}\sum_{t=0}^{T-1}\left|\lambda^{(t)}\,-\,\lambda^{(t+1)}\right|V_g^{(t+1)}(\mu) \\
\leq \;& \frac{2\eta_2}{(1-\gamma)^2}
\end{aligned}
$$

where in the first inequality we take telescoping sums for the first sum and drop a non-positive term, and in the second inequality we utilize the fact: $\left|\lambda^{(T)}\right| \leq \frac{\eta_2 T}{1-\gamma}$ and $\left|\lambda^{(t)} - \lambda^{(t+1)}\right| \leq \frac{\eta_2}{1-\gamma}$ due to the dual update in (8) and the non-expansiveness of projection, and the inequality $V_g^{(t)}(\mu) \leq \frac{1}{1-\gamma}$ due to the bounded value.

Finally, we use the fact that: $D_{\text{KL}}\left(p \,\|\, q\right) \leq \log|A|$, where $p, q \in \Delta_A$ and q is the unifom distribution, $V_r^{(T)}(d^\star) \leq \frac{1}{1-\gamma}$, and $V_g^\star(\rho) \geq b$ to complete the proof. □

We now prove our main statement in Theorem 1. We recall a key inequality from Lemma 7,

$$
\begin{aligned}
& \frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho)\,-\,V_r^{(t)}(\rho)\right)\,+\,\frac{1}{T}\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho)\,-\,V_g^{(t)}(\rho)\right) \\
\leq \;& \frac{\log|A|}{\eta_1 T}\,+\,\frac{1}{(1-\gamma)^2 T}\,+\,\frac{2\eta_2}{(1-\gamma)^3}.
\end{aligned}
\tag{23}
$$

Bounding the optimality gap. By the dual update in (8),

$$
\begin{aligned}
0 \;\leq\; \left(\lambda^{(T)}\right)^2 \;&=\; \sum_{t=0}^{T-1}\left((\lambda^{(t+1)})^2\,-\,(\lambda^{(t)})^2\right) \\
&=\; \sum_{t=0}^{T-1}\left(\left(\mathcal{P}_\Lambda\left(0, \lambda^{(t)} - \eta_2(V_g^{(t)}(\rho) - b)\right)\right)^2\,-\,(\lambda^{(t)})^2\right) \\
&\leq\; \sum_{t=0}^{T-1}\left(\left(\lambda^{(t)} - \eta_2(V_g^{(t)}(\rho) - b)\right)^2\,-\,(\lambda^{(t)})^2\right) \\
&=\; 2\eta_2\sum_{t=0}^{T-1}\lambda^{(t)}(b - V_g^{(t)}(\rho))\,+\,\eta_2^2\sum_{t=0}^{T-1}(V_g^{(t)}(\rho) - b)^2 \\
&\leq\; 2\eta_2\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho) - V_g^{(t)}(\rho)\right)\,+\,\frac{\eta_2^2 T}{(1-\gamma)^2}
\end{aligned}
\tag{24}
$$

where the last inequality is due to the feasibility of the optimal policy $\pi^\star$: $V_g^\star(\rho) \geq b$, and $|V_g^{(t)}(\rho) - b| \leq \frac{1}{1-\gamma}$. The above inequality further implies,

$$
-\frac{1}{T}\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho) - V_g^{(t)}(\rho)\right) \;\leq\; \frac{\eta_2}{2(1-\gamma)^2}.
$$

We substitute the above inequality into (23) and use the fact that: $D_{\text{KL}}\left(p \,\|\, q\right) \leq \log|A|$, where $p, q \in \Delta_A$ and q is the uniform distribution to show the optimality gap bound, where we take $\eta_1 = 2\log|A|$ and $\eta_2 = \frac{1-\gamma}{\sqrt{T}}$.

Bounding the constraint violation. By the dual update in (8), for any $\lambda \in \left[0, \frac{2}{(1-\gamma)\xi}\right]$,

$$
\begin{aligned}
|\lambda^{(t+1)} - \lambda|^2 &\leq \left|\lambda^{(t)} - \eta_2\left(V_g^{(t)}(\rho) - b\right) - \lambda\right|^2 \\
&= \left|\lambda^{(t)} - \lambda\right|^2 - 2\eta_2\left(V_g^{(t)}(\rho) - b\right)\left(\lambda^{(t)} - \lambda\right) + \eta_2^2\left(V_g^{(t)}(\rho) - b\right)^2 \\
&\leq \left|\lambda^{(t)} - \lambda\right|^2 - 2\eta_2\left(V_g^{(t)}(\rho) - b\right)\left(\lambda^{(t)} - \lambda\right) + \frac{\eta_2^2}{(1-\gamma)^2}
\end{aligned}
$$

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

$$
0 \leq \frac{1}{T}|\lambda^{(T)} - \lambda|^2 \leq \frac{1}{T}\left|\lambda^{(0)} - \lambda\right|^2 - \frac{2\eta_2}{T}\sum_{t=0}^{T-1}\left(V_g^{(t)}(\rho) - b\right)\left(\lambda^{(t)} - \lambda\right) + \frac{\eta_2^2}{(1-\gamma)^2},
$$

which further implies,

$$
\frac{1}{T}\sum_{t=0}^{T-1}\left(V_g^{(t)}(\rho) - b\right)\left(\lambda^{(t)} - \lambda\right) \leq \frac{1}{2\eta_2 T}\left|\lambda^{(0)} - \lambda\right|^2 + \frac{\eta_2}{2(1-\gamma)^2}.
$$

We now add the above inequality into (23) and note $V_g^\star(\rho) \geq b$,

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right) + \frac{\lambda}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right) \\
&\leq \frac{\log|A|}{\eta_1 T} + \frac{1}{(1-\gamma)^2 T} + \frac{2\eta_2}{(1-\gamma)^3} + \frac{1}{2\eta_2 T}\left|\lambda^{(0)} - \lambda\right|^2 + \frac{\eta_2}{2(1-\gamma)^2}.
\end{aligned}
$$

We take $\lambda = \frac{2}{(1-\gamma)\xi}$ when $\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right) \geq 0$; otherwise $\lambda = 0$. Thus,

$$
\begin{aligned}
&V_r^\star(\rho) - \frac{1}{T}\sum_{t=0}^{T-1}V_r^{(t)}(\rho) + \frac{2}{(1-\gamma)\xi}\left[b - \frac{1}{T}\sum_{t=0}^{T-1}V_g^{(t)}(\rho)\right]_+ \\
&\leq \frac{\log|A|}{\eta_1 T} + \frac{1}{(1-\gamma)^2 T} + \frac{2\eta_2}{(1-\gamma)^3} + \frac{1}{2\eta_2(1-\gamma)^2\xi^2 T} + \frac{\eta_2}{2(1-\gamma)^2}.
\end{aligned}
$$

It should be mentioned that there exists a policy π' such that $V_r^{\pi'}(\rho) = \frac{1}{T}\sum_{t=0}^{T-1}V_r^{(t)}(\rho)$ and $V_g^{\pi'}(\rho) = \frac{1}{T}\sum_{t=0}^{T-1}V_g^{(t)}(\rho)$. This holds in the following way (see [4, Chapter 10]). We recall that $V_r^{(t)}(\rho)$ and $V_g^{(t)}(\rho)$ are linear functions in the occupancy measure induced by policy $\pi^{(t)}$ and transition $P(s' \mid s, a)$. By the convexity of the set of occupancy measures, an average of T occupancy measures is still an occupancy measure that produces a policy π' with value $V_r^{\pi'}$ and $V_g^{\pi'}$.

Therefore,

$$
\begin{aligned}
&V_r^\star(\rho) - V_r^{\pi'}(\rho) + \frac{2}{(1-\gamma)\xi}\left[b - V_g^{\pi'}(\rho)\right]_+ \\
&\leq \frac{\log|A|}{\eta_1 T} + \frac{1}{(1-\gamma)^2 T} + \frac{2\eta_2}{(1-\gamma)^3} + \frac{1}{2\eta_2(1-\gamma)^2\xi^2 T} + \frac{\eta_2}{2(1-\gamma)^2}.
\end{aligned}
$$

We note that $\frac{2}{(1-\gamma)\xi} \geq 2\lambda^\star$. According to Lemma 2, we obtain

$$
\left[b - V_g^{\pi'}(\rho)\right]_+ \leq \frac{\xi\log|A|}{\eta_1 T} + \frac{\xi}{(1-\gamma)T} + \frac{2\eta_2\xi}{(1-\gamma)^2} + \frac{1}{2\eta_2(1-\gamma)\xi T} + \frac{\eta_2\xi}{2(1-\gamma)}
$$

which shows the constraint violation bound by noting $\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right) = b - V_g^{\pi'}(\rho)$ and taking $\eta_1 = 2\log|A|$ and $\eta_2 = \frac{1-\gamma}{\sqrt{T}}$.

E Proof of Theorem 2

We first characterize the effect of the compatible function approximation error on the convergence.

Lemma 8. *Let Assumption 1 and Assumption 2 hold for a policy class $\{\pi_\theta \,|\, \theta \in \Theta\}$. Fix a feasible comparison policy be π, a state distribution ρ, and $T > 0$. Define the induced state-action visitation measure under π: $\nu(s,a) = d_\rho^\pi(s)\pi(a\,|\,s)$. Suppose the iterates $\pi^{(t)}$ and $\lambda^{(t)}$ generated by the algorithm (10), $\theta^{(0)} = 0$, $\lambda^{(0)} = 0$, $\eta_1 = \eta_2 = 1/\sqrt{T}$ satisfy*

$$\frac{1}{T}\sum_{t=0}^{T-1} E^\nu(w^{(t)};\theta^{(t)},\lambda^{(t)}) \;\leq\; \widehat{\epsilon}_{\text{approx}} \;\; \text{and} \;\; \|w^{(t)}\| \;\leq\; W \;\; \text{for all } 0 \leq t < T.$$

Then,

$$\text{(Optimality gap)} \;\; \frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\pi(\rho) - V_r^{(t)}(\rho)\right) \;\leq\; \frac{C_1}{(1-\gamma)^3}\frac{1}{\sqrt{T}} + \frac{1}{1-\gamma}\sqrt{\widehat{\epsilon}_{\text{approx}}} \tag{25a}$$

$$\text{(Constraint violation)} \;\; \left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right)\right]_+ \;\leq\; \frac{C_2}{(1-\gamma)^2}\frac{1}{T^{1/4}} + \frac{\sqrt{2}}{(1-\gamma)^{1/2}}\left(\frac{\widehat{\epsilon}_{\text{approx}}}{T}\right)^{1/4} \tag{25b}$$

where $C_1 := 1 + \log|A| + \beta W^2$ and $C_2 := \sqrt{3 + 2\lambda^\star + 2\log|A| + \beta W^2}$.

Proof. By Assumption 2, application of Taylor's theorem to $\log \pi^{(t)}(a\,|\,s)$ yields

$$\log\frac{\pi^{(t)}(a\,|\,s)}{\pi^{(t+1)}(a\,|\,s)} + \nabla_\theta \log \pi^{(t)}(a\,|\,s)\left(\theta^{(t+1)} - \theta^{(t)}\right) \;\leq\; \frac{\beta}{2}\|\theta^{(t+1)} - \theta^{(t)}\|^2$$

where $\theta^{(t+1)} - \theta^{(t)} = \frac{\eta_1}{1-\gamma}w^{(t)}$. We unload d_ρ^π as d since π and ρ are fixed. Therefore,

$$\begin{aligned}
&\mathbb{E}_{s\sim d}\left(D_{\text{KL}}(\pi(\cdot\,|\,s)\,\|\,\pi^{(t)}(\cdot\,|\,s)) - D_{\text{KL}}(\pi(\cdot\,|\,s)\,\|\,\pi^{(t+1)}(\cdot\,|\,s))\right)\\
&= -\mathbb{E}_{s\sim d}\mathbb{E}_{a\sim\pi(\cdot\,|\,s)}\log\frac{\pi^{(t)}(a\,|\,s)}{\pi^{(t+1)}(a\,|\,s)}\\
&\geq \eta_1\mathbb{E}_{s\sim d}\mathbb{E}_{a\sim\pi(\cdot\,|\,s)}\left[\nabla_\theta\log\pi^{(t)}(a\,|\,s)\,w^{(t)}\right] - \beta\frac{\eta_1^2}{2(1-\gamma)^2}\|w^{(t)}\|^2\\
&= \eta_1\mathbb{E}_{s\sim d}\mathbb{E}_{a\sim\pi(\cdot\,|\,s)}\left[\nabla_\theta\log\pi^{(t)}(a\,|\,s)\,w_r^{(t)}\right]\\
&\quad + \eta_1\lambda^{(t)}\mathbb{E}_{s\sim d}\mathbb{E}_{a\sim\pi(\cdot\,|\,s)}\left[\nabla_\theta\log\pi^{(t)}(a\,|\,s)\,w_g^{(t)}\right] - \beta\frac{\eta_1^2}{2(1-\gamma)^2}\|w^{(t)}\|^2\\
&= \eta_1\mathbb{E}_{s\sim d}\mathbb{E}_{a\sim\pi(\cdot\,|\,s)}A_r^{(t)}(s,a) + \eta_1\lambda^{(t)}\mathbb{E}_{s\sim d}\mathbb{E}_{a\sim\pi(\cdot\,|\,s)}A_g^{(t)}(s,a)\\
&\quad + \eta_1\mathbb{E}_{s\sim d}\mathbb{E}_{a\sim\pi(\cdot\,|\,s)}\left[\nabla_\theta\log\pi^{(t)}(a\,|\,s)\;\left(w_r^{(t)} + \lambda^{(t)}w_g^{(t)}\right) - \left(A_r^{(t)}(s,a) + \lambda^{(t)}A_g^{(t)}(s,a)\right)\right]\\
&\quad - \beta\frac{\eta_1^2}{2(1-\gamma)^2}\|w^{(t)}\|^2\\
&\geq \eta_1(1-\gamma)\left(V_r^\pi(\rho) - V_r^{(t)}(\rho)\right) + \eta_1(1-\gamma)\lambda^{(t)}\left(V_g^\pi(\rho) - V_g^{(t)}(\rho)\right)\\
&\quad - \eta_1\sqrt{\mathbb{E}_{s\sim d}\mathbb{E}_{a\sim\pi(\cdot\,|\,s)}\left[\left(\nabla_\theta\log\pi^{(t)}(a\,|\,s)\,w^{(t)} - A_L^{(t)}(s,a)\right)^2\right]}\\
&\quad - \beta\frac{\eta_1^2}{2(1-\gamma)^2}W^2
\end{aligned}$$

where in the second equality we decompose $w^{(t)} = w_r^{(t)} + \lambda^{(t)}w_g^{(t)}$ for a given $\lambda^{(t)}$, in the last inequality we apply the performance difference lemma, the Jensen's inequality, and $\|w^{(t)}\| \leq W$.

Using notation of $E^\nu(w^{(t)};\theta^{(t)},\lambda^{(t)})$ and rearranging it yield,

$$\begin{aligned}
&V_r^\pi(\rho) - V_r^{(t)}(\rho)\\
&\leq \frac{1}{1-\gamma}\left(\frac{1}{\eta_1}\mathbb{E}_{s\sim d}\left(D_{\text{KL}}(\pi(\cdot\,|\,s)\,\|\,\pi^{(t)}(\cdot\,|\,s)) - D_{\text{KL}}(\pi(\cdot\,|\,s)\,\|\,\pi^{(t+1)}(\cdot\,|\,s))\right)\right)\\
&\quad + \frac{1}{1-\gamma}\sqrt{E^\nu(w^{(t)};\theta^{(t)},\lambda^{(t)})} + \beta\frac{\eta_1}{2(1-\gamma)^3}W^2 - \lambda^{(t)}\left(V_g^\pi(\rho) - V_g^{(t)}(\rho)\right)
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^{\pi}(\rho)-V_r^{(t)}(\rho)\right)\\
\leq\;\; & \frac{1}{(1-\gamma)\eta_1 T}\sum_{t=0}^{T-1}\left(\mathbb{E}_{s\sim d}\left(D_{\mathrm{KL}}(\pi(\cdot\,|\,s)\,\|\,\pi^{(t)}(\cdot\,|\,s))-D_{\mathrm{KL}}(\pi(\cdot\,|\,s)\,\|\,\pi^{(t+1)}(\cdot\,|\,s))\right)\right)\\
&+\frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\sqrt{E^{\nu}(w^{(t)};\theta^{(t)},\lambda^{(t)})}+\beta\frac{\eta_1}{2(1-\gamma)}W^2-\frac{1}{T}\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^{\pi}(\rho)-V_g^{(t)}(\rho)\right)\\
\leq\;\; & \frac{\log|A|}{(1-\gamma)\eta_1 T}\;+\;\frac{1}{1-\gamma}\sqrt{\epsilon_{\text{approx}}}\;+\;\beta\frac{\eta_1}{2(1-\gamma)^3}W^2\;+\;\frac{\eta_2}{2(1-\gamma)^2}
\end{aligned}
$$

where in the last inequality we take telescoping sum of the first sum and drop a non-positive term; we apply the Jensen's inequality to $\sqrt{x}$ for the second sum and use the error bounds, and the last sum is due to,

$$
\begin{aligned}
0\;\leq\;\left(\lambda^{(T)}\right)^2 \;&=\; \sum_{t=0}^{T-1}\left((\lambda^{(t+1)})^2\;-\;(\lambda^{(t)})^2\right)\\
&=\; \sum_{t=0}^{T-1}\left(\left(\max\left(0,\lambda^{(t)}-\eta_2(V_g^{(t)}(\rho)-b)\right)\right)^2\;-\;(\lambda^{(t)})^2\right)\\
&\leq\; \sum_{t=0}^{T-1}\left(\left(\lambda^{(t)}-\eta_2(V_g^{(t)}(\rho)-b)\right)^2\;-\;(\lambda^{(t)})^2\right) \qquad (26)\\
&=\; 2\eta_2\sum_{t=0}^{T-1}\lambda^{(t)}(b-V_g^{(t)}(\rho))\;+\;\eta_2^2\sum_{t=0}^{T-1}(V_g^{(t)}(\rho)-b)^2\\
&\leq\; 2\eta_2\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^{\pi}(\rho)-V_g^{(t)}(\rho)\right)\;+\;\frac{\eta_2^2 T}{(1-\gamma)^2}
\end{aligned}
$$

where the last inequality is due to the feasibility of the comparison policy π: $V_g^{\pi}(\rho)\geq b$, and $|V_g^{(t)}(\rho)-b|\leq\frac{1}{1-\gamma}$. The above inequality further implies,

$$
-\frac{1}{T}\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^{\pi}(\rho)-V_g^{(t)}(\rho)\right)\;\leq\;\frac{\eta_2}{2(1-\gamma)^2}.
$$

Now, we obtain the first bound by taking $\eta_1=\frac{1}{\sqrt{T}}$ and $\eta_2=\frac{1}{\sqrt{T}}$ and some simplification.

We next prove the second bound. By the dual update in (10), we have $\lambda^{(t+1)}-\lambda^{(t)}\geq-\eta_2(V_g^{(t)}(\rho)-b)$. Notice $\lambda^{(T)}\geq 0$. Therefore,

$$
\left[\frac{1}{T}\sum_{t=0}^{T-1}(b\;-\;V_g^{(t)}(\rho))\right]_+\;\leq\;\frac{1}{\eta_2 T}\sum_{t=0}^{T-1}\left(\lambda^{(t+1)}\;-\;\lambda^{(t)}\right)\;=\;\frac{1}{\eta_2 T}\lambda^{(T)}.
$$

It comes down to establishing a bound on $\lambda^{(T)}$. Similar to (26), by the dual update in (10) with $\Lambda=[0,\infty)$,

$$
\begin{aligned}
0\;\leq\;\left(\lambda^{(T)}\right)^2\;&\leq\;\sum_{t=0}^{T-1}\left(\left(\lambda^{(t)}-\eta_2(V_g^{(t)}(\rho)-b)\right)^2\;-\;(\lambda^{(t)})^2\right)\\
&=\;2\eta_2\sum_{t=0}^{T-1}\lambda^{(t)}(b-V_g^{(t)}(\rho))\;+\;\eta_2^2\sum_{t=0}^{T-1}(V_g^{(t)}(\rho)-b)^2\\
&\leq\;2\eta_2\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^{\pi}(\rho)-V_g^{(t)}(\rho)\right)\;+\;\frac{\eta_2^2 T}{(1-\gamma)^2}
\end{aligned}
$$

where the last inequality is due to the feasibility of the optimal policy π^π: $V_g^\pi(\rho) \geq b$, and $|V_g^{(t)}(\rho) - b| \leq \frac{1}{1-\gamma}$. Thus,

$$\left(\lambda^{(T)}\right)^2 \;\leq\; 2\eta_2 \sum_{t=0}^{T-1} \lambda^{(t)} \left(V_g^\pi(\rho) \;-\; V_g^{(t)}(\rho)\right) \;+\; \frac{\eta_2^2 T}{(1-\gamma)^2}.$$

Viewing the above bound, we return back to

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} \lambda^{(t)} \left(V_g^\pi(\rho) - V_g^{(t)}(\rho)\right) \\
\leq\;\; & \frac{1}{(1-\gamma)\eta_1 T}\sum_{t=0}^{T-1}\left(\mathbb{E}_{s\sim d}\left(D_{\rm KL}(\pi(\cdot\,|\,s)\,\|\,\pi^{(t)}(\cdot\,|\,s)) - D_{\rm KL}(\pi(\cdot\,|\,s)\,\|\,\pi^{(t+1)}(\cdot\,|\,s))\right)\right) \\
& + \frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\sqrt{E^\nu(w^{(t)};\theta^{(t)},\lambda^{(t)})} \;+\; \beta\frac{\eta_1}{2(1-\gamma)}W^2 \;-\; \frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\pi(\rho) - V_r^{(t)}(\rho)\right) \\
\leq\;\; & \frac{\log|A|}{(1-\gamma)\eta_1 T} \;+\; \frac{1}{1-\gamma}\sqrt{\widehat{\epsilon}_{\rm approx}} \;+\; \beta\frac{\eta_1}{2(1-\gamma)^3}W^2 \;+\; \frac{\lambda^\star+1}{1-\gamma}
\end{aligned}$$

where in the last inequality we take telescoping sum in the first sum and drop a non-positive term; we apply the Jensen's inequality to $\sqrt{x}$ for the second sum and use the error bounds, and the last sum is due to

$$\begin{aligned}
V_r^\pi(\rho) \;&=\; V_r^\star(\rho) + (V_r^\pi(\rho) - V_r^\star(\rho)) \\
&\geq\; V_r^\star(\rho) - \frac{1}{1-\gamma} \\
&=\; V_D^\star(\rho) - \frac{1}{1-\gamma} \\
&=\; \underset{\pi}{\text{maximize}}\;\; V_r^\pi(\rho) + \lambda^\star\,(V_g^\pi(\rho) - b) - \frac{1}{1-\gamma} \\
&\geq\; V_r^{(t)}(\rho) + \lambda^\star\,(V_g^{(t)}(\rho) - b) - \frac{1}{1-\gamma} \\
&\geq\; V_r^{(t)}(\rho) - \frac{\lambda^\star+1}{1-\gamma}
\end{aligned}$$

where in the second equality we apply the strong duality in Lemma 1, the first and last inequalities are due to the boundedness of $|V_r^\pi(\rho) - V_r^\star(\rho)| \leq \frac{1}{1-\gamma}$ and $|V_g^{(t)}(\rho) - b| \leq \frac{1}{1-\gamma}$.

Therefore,

$$\frac{1}{\eta_2 T}\,\lambda^{(T)} \;\leq\; \frac{1}{\eta_2 T}\sqrt{2\eta_2 T\left(\frac{\log|A|}{(1-\gamma)\eta_1 T} \;+\; \frac{\sqrt{\widehat{\epsilon}_{\rm approx}}}{1-\gamma} \;+\; \frac{\eta_1\beta W^2}{2(1-\gamma)^3} \;+\; \frac{\lambda^\star+1}{1-\gamma}\right) \;+\; \frac{\eta_2^2 T}{(1-\gamma)^2}}$$

which leads to the desired bound by taking $\eta_1 = \frac{1}{\sqrt{T}}$ and $\eta_2 = \frac{1}{\sqrt{T}}$ and some simplification. $\square$

In Lemma 8, the compatible function approximation error shows up as an additive term in the upper bound for the optimality gap (25a) or the constraint violation (25b).

We now prove Theorem 2. It follows from the proof of Lemma 8 with an application of the inequality,

$$E^{\nu^\star}(w^{(t)};\theta^{(t)},\lambda^{(t)}) \;\leq\; \left\|\frac{\nu^\star}{\nu^{(t)}}\right\|_\infty E^{\nu^{(t)}}(w^{(t)};\theta^{(t)},\lambda^{(t)}) \;\leq\; \left\|\frac{\nu^\star}{\nu^{(t)}}\right\|_\infty \epsilon_{\rm approx} \;\leq\; \frac{\epsilon_{\rm approx}}{1-\gamma}\left\|\frac{\nu^\star}{\nu_0}\right\|_\infty$$

where the last inequality is due to $\nu^{(t)}(s,a) \geq (1-\gamma)\nu_0(s,a)$.

F Sample-Based NPG-PD Algorithm with Function Approximation

We describe a sample-based NPG-PD algorithm with function approximation in Algorithm 1. We note the computational complexity of Algorithm 1: each round has expected length $2/(1-\gamma)$ so the expected number of total samples is $4KT/(1-\gamma)$; the total number of gradient computations $\nabla_\theta \log \pi^{(t)}(a\,|\,s)$ is $2KT$; the total number of scalar multiplies, divides, and additions is $O(dKT + KT/(1-\gamma))$.

Algorithm 1 Sample-based NPG-PD Algorithm with Function Approximation

Initialization: Learning rates η_1 and η_2, SGD learning rate α, number of SGD iterations K, and simulation access to CMDP$(S, A, P, r, g, b, \gamma, \rho)$ under initial state-action distribution ν_0.

for $t = 0, \ldots, T-1$ **do**

  Initialize $\theta^{(0)} = 0$, $\lambda^{(0)} = 0$, $w_0 = 0$.

  for $k = 0, 1, \ldots, K-1$ **do**

    Draw $(s, a) \sim \nu^{(t)}$.

    Execute policy $\pi^{(t)}$ starting from (s, a) with a termination probability $1-\gamma$ and estimate,

$$\widehat{Q}_L^{(t)}(s,a) \;=\; \sum_{k=0}^{K'-1} \Big(r(s_k, a_k) + \lambda^{(t)} g(s_k, a_k) \Big) \;\; \text{where } s_0 = s, a_0 = a, K' \sim \text{Geo}(1-\gamma).$$

    Start from s, execute policy $\pi^{(t)}$ with a termination probability $1-\gamma$ and estimate,

$$\widehat{V}_L^{(t)}(s) \;=\; \sum_{k=0}^{K'-1} \Big(r(s_k, a_k) + \lambda^{(t)} g(s_k, a_k) \Big) \;\; \text{where } s_0 = s, K' \sim \text{Geo}(1-\gamma).$$

    $\widehat{A}_L^{(t)}(s,a) = \widehat{Q}_L^{(t)}(s,a) - \widehat{V}_L^{(t)}(s)$.

    SGD update $w_{k+1} \;=\; w_k \;-\; \alpha\, G_k$, where

$$G_k \;=\; 2 \Big(w_k \cdot \nabla_\theta \log \pi^{(t)}(a \,|\, s) - \widehat{A}_L^{(t)}(s,a) \Big) \nabla_\theta \log \pi^{(t)}(a \,|\, s).$$

  end for

  Set $\widehat{w}^{(t)} = \frac{1}{K} \sum_{k=1}^{K} w_k$.

  Initialize $\widehat{V}_g^{(t)}(\rho) = 0$.

  for $k = 0, 1, \ldots, K-1$ **do**

    Draw $s \sim \rho$ and draw $a \sim \pi^{(t)}(\cdot \,|\, s)$.

    Execute policy $\pi^{(t)}$ starting from s with a termination probability $1-\gamma$ and estimate,

$$\widehat{V}_g^{(t)}(s) \;=\; \sum_{k=0}^{K'-1} g(s_k, a_k) \;\; \text{where } s_0 = s, a_0 = a, K' \sim \text{Geo}(1-\gamma).$$

    Update $\widehat{V}_g^{(t)}(\rho) = \widehat{V}_g^{(t)}(\rho) + \frac{1}{K} \widehat{V}_g^{(t)}(s)$.

  end for

  Natural policy gradient primal-dual update

$$\begin{aligned} \theta^{(t+1)} &\;=\; \theta^{(t)} \;+\; \eta_1 \widehat{w}^{(t)} \\ \lambda^{(t+1)} &\;=\; \mathcal{P}_{[0,\infty)} \Big(\lambda^{(t)} - \eta_2 \big(\widehat{V}_g^{(t)}(\rho) - b \big) \Big). \end{aligned}$$

end for

We provide several unbiased estimates that are useful in our convergence proof.

$$
\begin{aligned}
\mathbb{E}\left[\widehat{V}_g^{(t)}(s)\right] &= \mathbb{E}\left[\sum_{k=0}^{K'-1} g(s_k,a_k)\,|\,\theta^{(t)}, s_0=s\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{\infty} \mathbb{I}\{K'-1\geq k\geq 0\} g(s_k,a_k)\,|\,\theta^{(t)}, s_0=s\right] \\
&= \sum_{k=0}^{\infty} \mathbb{E}\left[\mathbb{E}_{K'}\left[\mathbb{I}\{K'-1\geq k\geq 0\}\right] g(s_k,a_k)\,|\,\theta^{(t)}, s_0=s\right] \\
&= \sum_{k=0}^{\infty} \mathbb{E}\left[\gamma^k g(s_k,a_k)\,|\,\theta^{(t)}, s_0=s\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k g(s_k,a_k)\,|\,\theta^{(t)}, s_0=s\right] \\
&= V_g^{(t)}(s)
\end{aligned}
$$

where we apply the Monotone Convergence Theorem and the Dominated Convergence Theorem for the third equality and the fifth equality to swap the expectation and the infinite sum, and in the fourth equality we use $\mathbb{E}_{K'}\left[\mathbb{I}\{K'-1\geq k\geq 0\}\right] = 1 - P\left(K'<k\right) = \gamma^k$ since K' follows a geometric distribution Geo$(1-\gamma)$.

By a similar agument as above,

$$
\begin{aligned}
\mathbb{E}\left[\widehat{Q}_L^{(t)}(s,a)\right] &= \mathbb{E}\left[\sum_{k=0}^{K'-1} \Big(r(s_k,a_k)+\lambda^{(t)} g(s_k,a_k)\Big)\,|\,\theta^{(t)}, s_0=s, a_0=a\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{K'-1} r(s_k,a_k)\,|\,\theta^{(t)}, s_0=s, a_0=a\right] \\
&\quad + \lambda^{(t)}\mathbb{E}\left[\sum_{k=0}^{K'-1} g(s_k,a_k)\,|\,\theta^{(t)}, s_0=s, a_0=a\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{\infty} \mathbb{I}\{K'-1\geq k\geq 0\} r(s_k,a_k)\,|\,\theta^{(t)}, s_0=s, a_0=a\right] \\
&\quad + \lambda^{(t)}\mathbb{E}\left[\sum_{k=0}^{\infty} \mathbb{I}\{K'-1\geq k\geq 0\} g(s_k,a_k)\,|\,\theta^{(t)}, s_0=s, a_0=a\right] \\
&= \sum_{k=0}^{\infty} \mathbb{E}\left[\mathbb{E}_{K'}\left[\mathbb{I}\{K'-1\geq k\geq 0\}\right] r(s_k,a_k)\,|\,\theta^{(t)}, s_0=s, a_0=a\right] \\
&\quad + \lambda^{(t)}\sum_{k=0}^{\infty} \mathbb{E}\left[\mathbb{E}_{K'}\left[\mathbb{I}\{K'-1\geq k\geq 0\}\right] g(s_k,a_k)\,|\,\theta^{(t)}, s_0=s, a_0=a\right] \\
&= \sum_{k=0}^{\infty} \mathbb{E}\left[\gamma^k r(s_k,a_k)\,|\,\theta^{(t)}, s_0=s, a_0=a\right] \\
&\quad + \lambda^{(t)}\sum_{k=0}^{\infty} \mathbb{E}\left[\gamma^k g(s_k,a_k)\,|\,\theta^{(t)}, s_0=s, a_0=a\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(s_k,a_k)\,|\,\theta^{(t)}, s_0=s, a_0=a\right] \\
&\quad + \lambda^{(t)}\mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k g(s_k,a_k)\,|\,\theta^{(t)}, s_0=s, a_0=a\right] \\
&= Q_r^{(t)}(s,a) + \lambda^{(t)} Q_g^{(t)}(s,a) \\
&= Q_L^{(t)}(s,a).
\end{aligned}
$$

Therefore,

$$\mathbb{E}\left[\widehat{A}_L^{(t)}(s,a)\right] \;=\; \mathbb{E}\left[\widehat{Q}_L^{(t)}(s,a)\right] \,-\, \mathbb{E}\left[\widehat{V}_L^{(t)}(s)\right] \;=\; Q_L^{(t)}(s,a) \,-\, V_L^{(t)}(s) \;=\; A_L^{(t)}(s,a).$$

We also provide a bound on the variance of $\widehat{V}_g^{(t)}(s)$.

$$\begin{aligned}
\text{Var}\left[\widehat{V}_g^{(t)}(s)\right] &= \mathbb{E}\left[\left(\widehat{V}_g^{(t)}(s) - V_g^{(t)}(s)\right)^2 \mid \theta^{(t)}, s_0 = s\right] \\
&= \mathbb{E}\left[\left(\textstyle\sum_{k=0}^{K'-1} g(s_k,a_k) - V_g^{(t)}(s)\right)^2 \mid \theta^{(t)}, s_0 = s\right] \\
&= \mathbb{E}_{K'}\left[\mathbb{E}\left[\left(\textstyle\sum_{k=0}^{K'-1} g(s_k,a_k) - V_g^{(t)}(s)\right)^2\right] \mid K'\right] \\
&\leq \mathbb{E}_{K'}\left[(K')^2 \mid K'\right] \\
&= \frac{1}{(1-\gamma)^2}
\end{aligned}$$

where the first inequality is due to $0 \leq g(x_k, a_k) \leq 1$ and $V_g^{(t)}(s) \geq 0$ and the last equality is clear from $K' \sim \text{Geo}(1-\gamma)$.

G Proof of Theorem 3

We split the proof into two parts. We state the roadmap here for readers' convenience. In the first part, we establish the following two bounds for the optimality gap and the constraint violation,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right)\right] \;\leq\;& \frac{\log|A|}{(1-\gamma)\eta_1 T} + \frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\mathbb{E}\left[\sqrt{E^{\nu^\star}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})}\right] \\
&+\beta\frac{\eta_1}{2(1-\gamma)^3}\widehat{W}^2 + \frac{2\eta_2}{(1-\gamma)^2}
\end{aligned} \tag{27}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}(b - V_g^{(t)}(\rho))\right]_+ \;\leq \\
&\frac{1}{\eta_2 T}\sqrt{2\eta_2 T\left(\frac{\log|A|}{(1-\gamma)\eta_1 T} + \frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\mathbb{E}\left[\sqrt{E^{\nu^\star}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})}\right] + \frac{\eta_1\beta\widehat{W}^2}{2(1-\gamma)^3} + \frac{\lambda^\star+1}{1-\gamma}\right) + \frac{4\eta_2^2 T}{(1-\gamma)^2}}.
\end{aligned} \tag{28}$$

In the second part, we are seeking to control the error $\mathbb{E}\left[E^{\nu^\star}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})\right]$,

$$\mathbb{E}\left[E^{\nu^\star}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})\right] \;\leq\; \frac{1}{1-\gamma}\left\|\frac{\nu^\star}{\nu_0}\right\|_\infty\left(\epsilon_{\text{approx}} + \frac{2\left(\sqrt{d}\,WL_\pi + \frac{\sqrt{d}}{1-\gamma} + WL_\pi\right)^2}{K}\right). \tag{29}$$

Finally, we combine two parts to complete the proof by taking $\eta_1 = \frac{1}{\sqrt{T}}$ and $\eta_2 = \frac{1}{\sqrt{T}}$ and noting

$$\mathbb{E}\left[\sqrt{E^{\nu^\star}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})}\right] \;\leq\; \sqrt{\mathbb{E}\left[E^{\nu^\star}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})\right]}.$$

Let us begin with the first part. By Assumption 2, application of Taylor's theorem to $\log\pi^{(t)}(a\,|\,s)$ yields

$$\log\frac{\pi^{(t)}(a\,|\,s)}{\pi^{(t+1)}(a\,|\,s)} \,+\, \nabla_\theta\log\pi^{(t)}(a\,|\,s)\left(\theta^{(t+1)} - \theta^{(t)}\right) \;\leq\; \frac{\beta}{2}\|\theta^{(t+1)} - \theta^{(t)}\|^2$$

where $\theta^{(t+1)} - \theta^{(t)} = \frac{\eta_1}{1-\gamma}\widehat{w}^{(t)}$. We unload $d_\rho^{\pi^\star}$ as $d^\star$ since $\pi^\star$ and ρ are fixed. Therefore,

$$
\begin{aligned}
&\mathbb{E}_{s\,\sim\, d^\star}\left(D_{\mathrm{KL}}(\pi^\star(\cdot\,|\,s)\,\|\,\pi^{(t)}(\cdot\,|\,s)) - D_{\mathrm{KL}}(\pi^\star(\cdot\,|\,s)\,\|\,\pi^{(t+1)}(\cdot\,|\,s))\right)\\
&= \; -\mathbb{E}_{s\,\sim\, d^\star}\mathbb{E}_{a\,\sim\,\pi^\star(\cdot\,|\,s)}\log\frac{\pi^{(t)}(a\,|\,s)}{\pi^{(t+1)}(a\,|\,s)}\\
&\geq \; \eta_1\mathbb{E}_{s\,\sim\, d^\star}\mathbb{E}_{a\,\sim\,\pi^\star(\cdot\,|\,s)}\left[\nabla_\theta\log\pi^{(t)}(a\,|\,s)\,\widehat{w}^{(t)}\right] \,-\, \beta\frac{\eta_1^2}{2(1-\gamma)^2}\|\widehat{w}^{(t)}\|^2\\
&= \; \eta_1\mathbb{E}_{s\,\sim\, d^\star}\mathbb{E}_{a\,\sim\,\pi^\star(\cdot\,|\,s)}\left[\nabla_\theta\log\pi^{(t)}(a\,|\,s)\,\widehat{w}_r^{(t)}\right]\\
&\quad + \eta_1\lambda^{(t)}\mathbb{E}_{s\,\sim\, d^\star}\mathbb{E}_{a\,\sim\,\pi^\star(\cdot\,|\,s)}\left[\nabla_\theta\log\pi^{(t)}(a\,|\,s)\,\widehat{w}_g^{(t)}\right]\\
&\quad - \beta\frac{\eta_1^2}{2(1-\gamma)^2}\|\widehat{w}^{(t)}\|^2\\
&= \; \eta_1\mathbb{E}_{s\,\sim\, d^\star}\mathbb{E}_{a\,\sim\,\pi^\star(\cdot\,|\,s)}A_r^{(t)}(s,a) \,+\, \eta_1\lambda^{(t)}\mathbb{E}_{s\,\sim\, d^\star}\mathbb{E}_{a\,\sim\,\pi^\star(\cdot\,|\,s)}A_g^{(t)}(s,a)\\
&\quad + \eta_1\mathbb{E}_{s\,\sim\, d^\star}\mathbb{E}_{a\,\sim\,\pi^\star(\cdot\,|\,s)}\Big[\nabla_\theta\log\pi^{(t)}(a\,|\,s)\,\left(\widehat{w}_r^{(t)}+\lambda^{(t)}\widehat{w}_g^{(t)}\right) - \left(A_r^{(t)}(s,a)+\lambda^{(t)}A_g^{(t)}(s,a)\right)\Big]\\
&\quad - \beta\frac{\eta_1^2}{2(1-\gamma)^2}\|\widehat{w}^{(t)}\|^2\\
&\geq \; \eta_1(1-\gamma)\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right) \,+\, \eta_1(1-\gamma)\lambda^{(t)}\left(V_g^\star(\rho) - V_g^{(t)}(\rho)\right)\\
&\quad - \eta_1\sqrt{\mathbb{E}_{s\,\sim\, d^\star}\mathbb{E}_{a\,\sim\,\pi^\star(\cdot\,|\,s)}\left[\left(\nabla_\theta\log\pi^{(t)}(a\,|\,s)\,\widehat{w}^{(t)} - A_L^{(t)}(s,a)\right)^2\right]}\\
&\quad - \frac{\eta_1^2\beta\widehat{W}^2}{2(1-\gamma)^2}
\end{aligned}
$$

where in the second equality we decompose $\widehat{w}^{(t)} = \widehat{w}_r^{(t)} + \lambda^{(t)}\widehat{w}_g^{(t)}$ for a given $\lambda^{(t)}$, in the last inequality we apply the performance difference lemma, the Jensen's inequality, and $\|\widehat{w}^{(t)}\| \leq \widehat{W}$. Using notation of $E^{\nu^\star}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})$ and rearranging it yields

$$
\begin{aligned}
&V_r^\star(\rho) - V_r^{(t)}(\rho)\\
&\leq \; \frac{1}{1-\gamma}\left(\frac{1}{\eta_1}\mathbb{E}_{s\,\sim\, d^\star}\left(D_{\mathrm{KL}}(\pi^\star(\cdot\,|\,s)\,\|\,\pi^{(t)}(\cdot\,|\,s)) - D_{\mathrm{KL}}(\pi^\star(\cdot\,|\,s)\,\|\,\pi^{(t+1)}(\cdot\,|\,s))\right)\right)\\
&\quad + \frac{1}{1-\gamma}\sqrt{E^{\nu^\star}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})} \,+\, \beta\frac{\eta_1}{2(1-\gamma)^3}\widehat{W}^2 \,-\, \lambda^{(t)}\left(V_g^\star(\rho) - V_g^{(t)}(\rho)\right)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\frac{1}{T}\sum_{t\,=\,0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right)\\
&\leq \; \frac{1}{(1-\gamma)\eta_1 T}\sum_{t\,=\,0}^{T-1}\left(\mathbb{E}_{s\,\sim\, d^\star}\left(D_{\mathrm{KL}}(\pi^\star(\cdot\,|\,s)\,\|\,\pi^{(t)}(\cdot\,|\,s)) - D_{\mathrm{KL}}(\pi^\star(\cdot\,|\,s)\,\|\,\pi^{(t+1)}(\cdot\,|\,s))\right)\right)\\
&\quad + \frac{1}{(1-\gamma)T}\sum_{t\,=\,0}^{T-1}\sqrt{E^{\nu^\star}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})} + \frac{\eta_1\beta\widehat{W}^2}{2(1-\gamma)} - \frac{1}{T}\sum_{t\,=\,0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho) - V_g^{(t)}(\rho)\right)\\
&\leq \; \frac{\log|A|}{(1-\gamma)\eta_1 T} \,+\, \frac{1}{(1-\gamma)T}\sum_{t\,=\,0}^{T-1}\sqrt{E^{\nu^\star}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})}\\
&\quad + \frac{\eta_1\beta\widehat{W}^2}{2(1-\gamma)^3} \,-\, \frac{1}{T}\sum_{t\,=\,0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho) - V_g^{(t)}(\rho)\right)
\end{aligned}
\tag{30}
$$

where in the last inequality we take telescoping sum of the first sum and drop a non-positive term. Taking expectation over randomness in $\theta^{(t)}$ and $\lambda^{(t)}$ yields

$$\begin{aligned}\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho)-V_r^{(t)}(\rho)\right)\right] \;\leq\; & \frac{\log|A|}{(1-\gamma)\eta_1 T}+\mathbb{E}\left[\frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\sqrt{E^{\nu^\star}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})}\right]\\ & +\frac{\eta_1\beta\widehat{W}^2}{2(1-\gamma)^3}-\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho)-V_g^{(t)}(\rho)\right)\right].\end{aligned} \tag{31}$$

By the dual update in (12),

$$\begin{aligned} 0 \;\leq\; \left(\lambda^{(T)}\right)^2 \;&=\; \sum_{t=0}^{T-1}\left((\lambda^{(t+1)})^2-(\lambda^{(t)})^2\right)\\ &=\; \sum_{t=0}^{T-1}\left(\left(\max\left(0,\lambda^{(t)}-\eta_2(\widehat{V}_g^{(t)}(\rho)-b)\right)\right)^2-(\lambda^{(t)})^2\right)\\ &\leq\; \sum_{t=0}^{T-1}\left(\left(\lambda^{(t)}-\eta_2(\widehat{V}_g^{(t)}(\rho)-b)\right)^2-(\lambda^{(t)})^2\right)\\ &=\; 2\eta_2\sum_{t=0}^{T-1}\lambda^{(t)}(b-\widehat{V}_g^{(t)}(\rho))+\eta_2^2\sum_{t=0}^{T-1}(\widehat{V}_g^{(t)}(\rho)-b)^2\\ &\leq\; 2\eta_2\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho)-V_g^{(t)}(\rho)\right)+2\eta_2\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^{(t)}(\rho)-\widehat{V}_g^{(t)}(\rho)\right)\\ &\quad\;+\eta_2^2\sum_{t=0}^{T-1}(\widehat{V}_g^{(t)}(\rho)-b)^2\end{aligned} \tag{32}$$

where the last inequality is due to the feasibility of the policy $\pi^\star$: $V_g^\star(\rho)\geq b$. Since $V_g^{(t)}(\rho)$ is a population quantity and $\widehat{V}_g^{(t)}(\rho)$ is an estimate that is independent of $\lambda^{(t)}$ given $\theta^{(t-1)}$, $\lambda^{(t)}$ is independent of $V_g^{(t)}(\rho)-\widehat{V}_g^{(t)}(\rho)$ at time t and thus $\mathbb{E}\big[\lambda^{(t)}\big(V_g^{(t)}(\rho)-\widehat{V}_g^{(t)}(\rho)\big)\big]=0$ due to the fact $\mathbb{E}\big[\widehat{V}_g^{(t)}(\rho)\big]=V_g^{(t)}(\rho)$ (see Appendix F). Therefore,

$$\begin{aligned}-\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho)-V_g^{(t)}(\rho)\right)\right] \;&\leq\; \mathbb{E}\left[\frac{\eta_2}{2T}\sum_{t=0}^{T-1}(\widehat{V}_g^{(t)}(\rho)-b)^2\right]\\ &\leq\; \frac{\eta_2}{2(1-\gamma)^2}\left(1+\frac{K+1}{K}\right)\end{aligned}$$

where in the second inequality we drop a non-positive term and use the fact (see Appendix F),

$$\mathbb{E}\left[\widehat{V}_g^{(t)}(\rho)\right] \;=\; V_g^{(t)}(\rho)$$

and

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{V}_g^{(t)}(\rho)\right)^2\right] \;&=\; \frac{1}{K}\mathbb{E}\left[\left(\widehat{V}_g^{(t)}(s)\right)^2\right]+\frac{K-1}{K}\mathbb{E}\left[\widehat{V}_g^{(t)}(s)\right]\mathbb{E}\left[\widehat{V}_g^{(t)}(s)\right]\\ &=\; \frac{1}{K}\left(\mathrm{Var}\left[\widehat{V}_g^{(t)}(s)\right]+\left(\mathbb{E}\left[\widehat{V}_g^{(t)}(s)\right]\right)^2\right)+\frac{K-1}{K(1-\gamma)^2}\\ &=\; \frac{1}{K}\left(\mathrm{Var}\left[\widehat{V}_g^{(t)}(s)\right]+\left(V_g^{(t)}(s)\right)^2\right)+\frac{K-1}{K(1-\gamma)^2}\\ &\leq\; \frac{2}{K(1-\gamma)^2}+\frac{K-1}{K(1-\gamma)^2}\end{aligned}$$

where the first equality is due to line 16 of Algorithm 1; the last inequality is due to $\mathrm{Var}\big[\widehat{V}_g^{(t)}(s)\big]\leq\frac{1}{(1-\gamma)^2}$ (see Appendix F) and $0\leq V_g^{(t)}(s)\leq\frac{1}{1-\gamma}$.

We now return to (31) and apply $1+\frac{K+1}{K}\leq 4$ to obtain (27).

On the other hand, by the dual update in (12), we have $\lambda^{(t+1)} - \lambda^{(t)} \geq -\eta_2(\widehat{V}_g^{(t)}(\rho) - b)$. Therefore,

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}(b - V_g^{(t)}(\rho)) &\leq \left[\frac{1}{T}\sum_{t=0}^{T-1}(b - \widehat{V}_g^{(t)}(\rho))\right]_+ + \frac{1}{T}\sum_{t=0}^{T-1}\left(\widehat{V}_g^{(t)}(\rho) - V_g^{(t)}(\rho)\right) \\
&\leq \frac{1}{\eta_2 T}\sum_{t=0}^{T-1}\left(\lambda^{(t+1)} - \lambda^{(t)}\right) + \frac{1}{T}\sum_{t=0}^{T-1}\left(\widehat{V}_g^{(t)}(\rho) - V_g^{(t)}(\rho)\right) \\
&= \frac{1}{\eta_2 T}\lambda^{(T)} + \frac{1}{T}\sum_{t=0}^{T-1}\left(\widehat{V}_g^{(t)}(\rho) - V_g^{(t)}(\rho)\right).
\end{aligned}
$$

By $\mathbb{E}\big[\widehat{V}_g^{(t)}(\rho)\big] = V_g^{(t)}(\rho)$ (see Appendix F),

$$
\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}(b - V_g^{(t)}(\rho))\right]_+ \leq \frac{1}{\eta_2 T}\mathbb{E}\left[\lambda^{(T)}\right] \tag{33}
$$

Also, by (32),

$$
\begin{aligned}
\mathbb{E}\left[\left(\lambda^{(T)}\right)^2\right] &\leq 2\eta_2\mathbb{E}\left[\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho) - V_g^{(t)}(\rho)\right)\right] + 2\eta_2\mathbb{E}\left[\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^{(t)}(\rho) - \widehat{V}_g^{(t)}(\rho)\right)\right] \\
&\quad + \eta_2^2\mathbb{E}\left[\sum_{t=0}^{T-1}(\widehat{V}_g^{(t)}(\rho) - b)^2\right] \\
&\leq 2\eta_2\mathbb{E}\left[\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho) - V_g^{(t)}(\rho)\right)\right] + \frac{4\eta_2^2 T}{(1-\gamma)^2}
\end{aligned}
$$

where we use arguments similar to those right below (32): $\mathbb{E}\big[\lambda^{(t)}\big(V_g^{(t)}(\rho) - \widehat{V}_g^{(t)}(\rho)\big)\big] = 0$ and $\mathbb{E}\big[\eta_2^2\sum_{t=0}^{T-1}(\widehat{V}_g^{(t)}(\rho) - b)^2\big] \leq \frac{\eta_2^2 T}{(1-\gamma)^2}\left(2 + \frac{K+1}{K}\right) \leq \frac{4\eta_2^2 T}{(1-\gamma)^2}$.

It should be noticed that $V_r^\star(\rho) = V_r^{\pi_\theta^\star}(\rho)$. Viewing the bound in (30), we return back to

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho) - V_g^{(t)}(\rho)\right) \\
&\leq \frac{1}{(1-\gamma)\eta_1 T}\sum_{t=0}^{T-1}\left(\mathbb{E}_{s\sim d^\star}\left(D_{\mathrm{KL}}(\pi^\star(\cdot\,|\,s)\,\|\,\pi^{(t)}(\cdot\,|\,s)) - D_{\mathrm{KL}}(\pi^\star(\cdot\,|\,s)\,\|\,\pi^{(t+1)}(\cdot\,|\,s))\right)\right) \\
&\quad + \frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\sqrt{E^{\nu^\star}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})} + \frac{\eta_1\beta\widehat{W}^2}{2(1-\gamma)^3} \\
&\quad - \frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right) \\
&\leq \frac{\log|A|}{(1-\gamma)\eta_1 T} + \frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\sqrt{E^{\nu^\star}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})} + \frac{\eta_1\beta\widehat{W}^2}{2(1-\gamma)^3} + \frac{\lambda^\star+1}{1-\gamma}
\end{aligned}
$$

where in the last inequality we take telescoping sum in the first sum and drop off a non-positive term; the last sum is due to

$$
\begin{aligned}
V_r^{\pi_\theta^\star}(\rho) &= V_r^\star(\rho) + (V_r^{\pi_\theta^\star}(\rho) - V_r^\star(\rho)) \\
&\geq V_r^\star(\rho) - \frac{1}{1-\gamma} \\
&= V_D^\star(\rho) - \frac{1}{1-\gamma} \\
&= \underset{\pi}{\text{maximize}}\;\; V_r^\pi(\rho) + \lambda^\star\left(V_g^\pi(\rho) - b\right) - \frac{1}{1-\gamma} \\
&\geq V_r^{(t)}(\rho) + \lambda^\star\,(V_g^{(t)}(\rho) - b) - \frac{1}{1-\gamma} \\
&\geq V_r^{(t)}(\rho) - \frac{\lambda^\star+1}{1-\gamma}
\end{aligned}
$$

where in the second equality we apply the strong duality in Lemma 1, the first and last inequalities are due to the boundedness of $|V_r^{\pi_\theta^\star}(\rho) - V_r^\star(\rho)| \leq \frac{1}{1-\gamma}$ and $|V_g^{(t)}(\rho) - b| \leq \frac{1}{1-\gamma}$. In the above context, we abuse notation $V_r^\star(\rho)$ a bit: $V_r^\star(\rho)$ is described by Lemma 1.

Notice $\left(\mathbb{E}\left[\lambda^{(T)}\right]\right)^2 \leq \mathbb{E}\left[\left(\lambda^{(T)}\right)^2\right]$. Therefore,

$$
\frac{1}{\eta_2 T}\mathbb{E}\left[\lambda^{(T)}\right] \;\leq\;
\frac{1}{\eta_2 T}\sqrt{2\eta_2 T\left(\frac{\log|A|}{(1-\gamma)\eta_1 T} + \frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\mathbb{E}\left[\sqrt{E^{\nu^\star}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})}\right] + \frac{\eta_1\beta\widehat{W}^2}{2(1-\gamma)^3} + \frac{\lambda^\star+1}{1-\gamma}\right) + \frac{4\eta_2^2 T}{(1-\gamma)^2}}
$$

which leads to the desired bound (28) via (33).

The second part seeks to upper bound the expecation of $E^{\nu^\star}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})$,

$$
\begin{aligned}
&\mathbb{E}\left[E^{\nu^\star}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})\right]\\
\leq\;& \mathbb{E}\left[\left\|\frac{\nu^\star}{\nu^{(t)}}\right\|_\infty E^{\nu^{(t)}}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})\right]\\
\leq\;& \frac{1}{1-\gamma}\mathbb{E}\left[\left\|\frac{\nu^\star}{\nu_0}\right\|_\infty E^{\nu^{(t)}}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})\right]\\
=\;& \frac{1}{1-\gamma}\left\|\frac{\nu^\star}{\nu_0}\right\|_\infty\left(\mathbb{E}\left[E_\star^{\nu^{(t)}}(\theta^{(t)},\lambda^{(t)})\right] + \mathbb{E}\left[E^{\nu^{(t)}}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})\right] - \mathbb{E}\left[E_\star^{\nu^{(t)}}(\theta^{(t)},\lambda^{(t)})\right]\right)\\
\leq\;& \frac{1}{1-\gamma}\left\|\frac{\nu^\star}{\nu_0}\right\|_\infty\left(\epsilon_{\text{approx}} + \mathbb{E}\left[E^{\nu^{(t)}}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})\right] - \mathbb{E}\left[E_\star^{\nu^{(t)}}(\theta^{(t)},\lambda^{(t)})\right]\right).
\end{aligned}
$$

To upper bound $\mathbb{E}\left[E^{\nu^{(t)}}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)})\right] - \mathbb{E}\left[E_\star^{\nu^{(t)}}(\theta^{(t)},\lambda^{(t)})\right]$, we analyze the SGD update in line 9 of Algorithm 1. The SGD update performs minimizing the objective $E^{\nu^{(t)}}(w;\theta^{(t)},\lambda^{(t)})$ with an unbiased estimate of the gradient $\nabla_w E^{\nu^{(t)}}(w_k;\theta^{(t)},\lambda^{(t)})$,

$$
\begin{aligned}
&\mathbb{E}\left[G_k \,|\, w_k\right]\\
=\;& 2\mathbb{E}_{(s,a)\sim\nu^{(t)}}\left[\left(w_k\cdot\nabla_\theta\log\pi^{(t)}(a\,|\,s) - \widehat{A}_L^{(t)}(s,a)\right)\nabla_\theta\log\pi^{(t)}(a\,|\,s)\right]\\
=\;& 2\mathbb{E}_{(s,a)\sim\nu^{(t)}}\left[\left(w_k\cdot\nabla_\theta\log\pi^{(t)}(a\,|\,s) - \mathbb{E}\big[\widehat{A}_L^{(t)}(s,a)\,|\,s,a\big]\right)\nabla_\theta\log\pi^{(t)}(a\,|\,s)\right]\\
=\;& 2\mathbb{E}_{(s,a)\sim\nu^{(t)}}\left[\left(w_k\cdot\nabla_\theta\log\pi^{(t)}(a\,|\,s) - A_L^{(t)}(s,a)\right)\nabla_\theta\log\pi^{(t)}(a\,|\,s)\right]\\
=\;& \nabla_w E^{\nu^{(t)}}(w_k;\theta^{(t)},\lambda^{(t)})
\end{aligned}
$$

where the last equality is due to the fact that: $\widehat{A}_L^{(t)}(s,a)$ is an unbiased estimate of $A_L^{(t)}(s,a)$ in line 8 of Algorithm 1.

By the fast SGD result [5, Theorem 1] with $\alpha = \frac{1}{L_\pi}$ and Assumption 3,

$$
\mathbb{E}\left[E^{\nu^{(t)}}(\widehat{w}^{(t)};\theta^{(t)},\lambda^{(t)}) - E_\star^{\nu^t}(\theta^{(t)},\lambda^{(t)})\right] \;\leq\; \frac{2\left(\sigma\sqrt{d}+WL_\pi\right)^2}{K}
$$

where σ is an uniform bound on the minimum variance,

$$
\mathbb{E}_{(s,a)\sim\nu^{(t)}}\left[G_\star^{(t)}\left(G_\star^{(t)}\right)^\top\right] \;\leq\; \sigma^2\,\nabla_w^2 E^{\nu^{(t)}}(w^{(t)};\theta^{(t)},\lambda^{(t)})
$$

$$
G_\star^{(t)} \;=\; \left(w^{(t)}\cdot\nabla_\theta\log\pi^{(t)}(a\,|\,s) - \widehat{A}_L^{(t)}(s,a)\right)\nabla_\theta\log\pi^{(t)}(a\,|\,s).
$$

We complete the second part by noting $\sigma < L_\pi\left\|w^{(t)}\right\| + \frac{1}{1-\gamma} \leq WL_\pi + \frac{1}{1-\gamma}$.

H Sample-Based NPG-PD Algorithm with Softmax Parametrization

We describe a sample-based NPG-PD algorithm with softmax parametrization in Algorithm 2. Regarding the computational complexity of Algorithm 2: each round has expected length $2/(1-\gamma)$ so the expected number of total samples is $2(2|S|+|S||A|)KT/(1-\gamma)$; the total number of scalar multiplies, divides, and additions is $O(|S||A|KT + KT/(1-\gamma))$.

Algorithm 2 Sample-Based NPG-PD Algorithm with Softmax Parametrization

Initialization: Learning rates η_1 and η_2, number of rounds K, and simulation access to CMDP$(S, A, P, r, g, b, \gamma, \rho)$.
for $t = 0, \ldots, T-1$ **do**
  Initialize $\theta^{(0)} = 0$, $\lambda^{(0)} = 0$, $w_0 = 0$.
  Initialize $\widehat{V}_L^{(t)}(s) = 0$ for all $s \in S$ and $\widehat{Q}_L^{(t)}(s,a) = 0$ for all $(s,a) \in S \times A$.
  for $k = 0, 1, \ldots, K-1$ **do**
    Starting from each $s \in S$, execute policy $\pi^{(t)}$ with a termination probability $1-\gamma$ and estimate,

$$\widehat{V}_L^{(t)}(s) \;=\; \sum_{k=0}^{K'-1} \Big(r(s_k, a_k) + \lambda^{(t)} g(s_k, a_k) \Big) \;\text{ where } s_0 = s, K' \sim \text{Geo}(1-\gamma).$$

    Starting from each $(s,a) \in S \times A$, execute policy $\pi^{(t)}$ with a termination probability $1-\gamma$ and estimate,

$$\widehat{Q}_L^{(t)}(s,a) \;=\; \sum_{k=0}^{K'-1} \Big(r(s_k, a_k) + \lambda^{(t)} g(s_k, a_k) \Big) \;\text{ where } s_0 = s, a_0 = a, K' \sim \text{Geo}(1-\gamma).$$

    Update $\widehat{V}_L^{(t)}(s) = \widehat{V}_L^{(t)}(s) + \frac{1}{K}\widehat{V}_L^{(t)}(s)$ for all $s \in S$.
    Update $\widehat{Q}_L^{(t)}(s,a) = \widehat{Q}_L^{(t)}(s,a) + \frac{1}{K}\widehat{Q}_L^{(t)}(s,a)$ for all $(s,a) \in S \times A$.
  end for
  Estimate $\widehat{A}_L^{(t)} = \widehat{Q}_L^{(t)} - \widehat{V}_L^{(t)}$.
  Initialize $\widehat{V}_g^{(t)}(\rho) = 0$.
  for $k = 0, 1, \ldots, K-1$ **do**
    Draw $s \sim \rho$ and draw $a \sim \pi^{(t)}(\cdot \,|\, s)$.
    Execute policy $\pi^{(t)}$ starting from s with a termination probability $1-\gamma$ and compute the estimate,

$$\widehat{V}_g^{(t)}(s) \;=\; \sum_{k=0}^{K'-1} g(s_k, a_k) \;\text{ where } s_0 = s, a_0 = a, K' \sim \text{Geo}(1-\gamma).$$

    Update $\widehat{V}_g^{(t)}(\rho) = \widehat{V}_g^{(t)}(\rho) + \frac{1}{K}\widehat{V}_g^{(t)}(s)$.
  end for
  Natural policy gradient primal-dual update

$$\begin{aligned} \theta^{(t+1)} &= \theta^{(t)} + \tfrac{\eta_1}{1-\gamma}\widehat{A}_L^{(t)} \\ \lambda^{(t+1)} &= \mathcal{P}_{[0,1/((1-\gamma)\xi)]}\left(\lambda^{(t)} - \eta_2\big(\widehat{V}_g^{(t)}(\rho) - b\big) \right). \end{aligned} \tag{34}$$

end for

Similar to Appendix F, we have unbiased estimates,

$$\mathbb{E}\left[\widehat{V}_L^{(t)}(s)\right] \;=\; V_L^{(t)}(s) \;\text{ and }\; \mathbb{E}\left[\widehat{Q}_L^{(t)}(s,a)\right] \;=\; Q_L^{(t)}(s,a) \;\text{ and }\; \mathbb{E}\left[\widehat{V}_g^{(t)}(s)\right] \;=\; V_g^{(t)}(s)$$

and a variance bound,

$$\text{Var}\left[\widehat{V}_g^{(t)}(s)\right] \;\leq\; \frac{1}{(1-\gamma)^2}.$$

I Proof of Theorem 4

The proof idea is similar to Theorem 1, we repeat it for readers' convenience.

We highlight some different steps here. The proof is based on similar results as Lemma 6 and Lemma 7 in which essentially we replace the population quantities by the empirical ones estimated by Algorithm 2. It is noted that the trajectory samplings in Algorithm 2 are independent at different times t and all estimates are unbiased. By Lemma 6 and Lemma 7,

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) \,-\, V_r^{(t)}(\rho)\right)\right] + \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho) \,-\, V_g^{(t)}(\rho)\right)\right] \\ &\leq\; \frac{\log|A|}{\eta_1 T} \,+\, \frac{1}{(1-\gamma)^2 T} \,+\, \frac{2\eta_2}{(1-\gamma)^3}. \end{aligned} \tag{35}$$

Bounding the optimality gap. By the dual update in (34),

$$\begin{aligned} 0 \;\leq\; \left(\lambda^{(T)}\right)^2 \;&=\; \sum_{t=0}^{T-1}\left((\lambda^{(t+1)})^2 \,-\, (\lambda^{(t)})^2\right) \\ &=\; \sum_{t=0}^{T-1}\left(\left(\mathcal{P}_\Lambda\left(\,0, \lambda^{(t)} - \eta_2(\widehat{V}_g^{(t)}(\rho) - b)\,\right)\right)^2 \,-\, (\lambda^{(t)})^2\right) \\ &\leq\; \sum_{t=0}^{T-1}\left(\left(\lambda^{(t)} - \eta_2(\widehat{V}_g^{(t)}(\rho) - b)\right)^2 \,-\, (\lambda^{(t)})^2\right) \\ &=\; 2\eta_2\sum_{t=0}^{T-1}\lambda^{(t)}(b - \widehat{V}_g^{(t)}(\rho)) \,+\, \eta_2^2\sum_{t=0}^{T-1}(\widehat{V}_g^{(t)}(\rho) - b)^2 \\ &\leq\; 2\eta_2\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho) - V_g^{(t)}(\rho)\right) \,+\, 2\eta_2\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^{(t)}(\rho) - \widehat{V}_g^{(t)}(\rho)\right) \\ &\quad\; + \eta_2^2\sum_{t=0}^{T-1}(\widehat{V}_g^{(t)}(\rho) - b)^2 \end{aligned} \tag{36}$$

where the last inequality is due to the feasibility of the policy $\pi^\star$: $V_g^\star(\rho) \geq b$. Since $V_g^{(t)}(\rho)$ is a population quantity and $\widehat{V}_g^{(t)}(\rho)$ is an estimate that is independent of $\lambda^{(t)}$ given $\theta^{(t-1)}$, $\lambda^{(t)}$ is independent of $V_g^{(t)}(\rho) - \widehat{V}_g^{(t)}(\rho)$ at time t and thus $\mathbb{E}\big[\lambda^{(t)}\big(V_g^{(t)}(\rho) - \widehat{V}_g^{(t)}(\rho)\big)\big] = 0$ due to the fact $\mathbb{E}\big[\widehat{V}_g^{(t)}(\rho)\big] = V_g^{(t)}(\rho)$ (see Appendix H). Therefore,

$$\begin{aligned} -\,\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho) - V_g^{(t)}(\rho)\right)\right] \;&\leq\; \mathbb{E}\left[\frac{\eta_2}{2T}\sum_{t=0}^{T-1}(\widehat{V}_g^{(t)}(\rho) - b)^2\right] \\ &\leq\; \frac{\eta_2}{2(1-\gamma)^2}\left(1 + \frac{K+1}{K}\right) \end{aligned}$$

where in the second inequality we drop a non-positive term and use the fact (see Appendix F),

$$\mathbb{E}\left[\widehat{V}_g^{(t)}(\rho)\right] \;=\; V_g^{(t)}(\rho)$$

and

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{V}_g^{(t)}(\rho)\right)^2\right] &= \frac{1}{K}\mathbb{E}\left[\left(\widehat{V}_g^{(t)}(s)\right)^2\right] + \frac{K-1}{K}\mathbb{E}\left[\widehat{V}_g^{(t)}(s)\right]\mathbb{E}\left[\widehat{V}_g^{(t)}(s)\right] \\ &= \frac{1}{K}\left(\text{Var}\left[\widehat{V}_g^{(t)}(s)\right] + \left(\mathbb{E}\left[\widehat{V}_g^{(t)}(s)\right]\right)^2\right) + \frac{K-1}{K(1-\gamma)^2} \\ &= \frac{1}{K}\left(\text{Var}\left[\widehat{V}_g^{(t)}(s)\right] + \left(V_g^{(t)}(s)\right)^2\right) + \frac{K-1}{K(1-\gamma)^2} \\ &\leq \frac{2}{K(1-\gamma)^2} + \frac{K-1}{K(1-\gamma)^2}\end{aligned}$$

where the first equality is due to line 16 of Algorithm 2; the last inequality is due to $\text{Var}\big[\widehat{V}_g^{(t)}(s)\big] \leq \frac{1}{(1-\gamma)^2}$ (see Appendix F) and $0 \leq V_g^{(t)}(s) \leq \frac{1}{1-\gamma}$.

To finish this part, we now return to (35) with $\eta_1 = 2\log|A|$ and $\eta_2 = \frac{1-\gamma}{\sqrt{T}}$,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right)\right] \leq \frac{\log|A|}{\eta_1 T} + \frac{1}{(1-\gamma)^2 T} + \frac{2\eta_2}{(1-\gamma)^3} + \frac{\eta_2}{(1-\gamma)^2} + \frac{\eta_2}{2(1-\gamma)^2 K}.$$

Bounding the constraint violation. By the dual update in (34), for any $\lambda \in \big[0, \frac{1}{(1-\gamma)\xi}\big]$,

$$\begin{aligned}&\mathbb{E}\left[|\lambda^{(t+1)} - \lambda|^2\right] \\ &= \mathbb{E}\left[\left|\mathcal{P}_\Lambda\left(\lambda^{(t)} - \eta_2\left(\widehat{V}_g^{(t)}(\rho) - b\right)\right) - \mathcal{P}_\Lambda(\lambda)\right|^2\right] \\ &\leq \mathbb{E}\left[\left|\lambda^{(t)} - \eta_2\left(\widehat{V}_g^{(t)}(\rho) - b\right) - \lambda\right|^2\right] \\ &= \mathbb{E}\left[\left|\lambda^{(t)} - \lambda\right|^2\right] - 2\eta_2\mathbb{E}\left[\left(\widehat{V}_g^{(t)}(\rho) - b\right)\left(\lambda^{(t)} - \lambda\right)\right] + \eta_2^2\mathbb{E}\left[\left(\widehat{V}_g^{(t)}(\rho) - b\right)^2\right] \\ &\leq \mathbb{E}\left[\left|\lambda^{(t)} - \lambda\right|^2\right] - 2\eta_2\mathbb{E}\left[\left(\widehat{V}_g^{(t)}(\rho) - b\right)\left(\lambda^{(t)} - \lambda\right)\right] + \frac{2\eta_2^2}{(1-\gamma)^2} + \frac{\eta_2^2}{K(1-\gamma)^2}\end{aligned}$$

where the first inequality is due to the non-expansiveness of $\mathcal{P}_\Lambda$ and the last inequality is due to $\mathbb{E}\big[(\widehat{V}_g^{(t)}(\rho) - b)^2\big] \leq \frac{2}{(1-\gamma)^2} + \frac{1}{K(1-\gamma)^2}$. Summing it up from $t = 0$ to $t = T-1$ and dividing it by T yield

$$\begin{aligned}0 &\leq \frac{1}{T}\mathbb{E}\left[|\lambda^{(T)} - \lambda|^2\right] \\ &\leq \frac{1}{T}\mathbb{E}\left[\left|\lambda^{(0)} - \lambda\right|^2\right] - \frac{2\eta_2}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left(\widehat{V}_g^{(t)}(\rho) - b\right)\left(\lambda^{(t)} - \lambda\right)\right] + \frac{2\eta_2^2}{(1-\gamma)^2} + \frac{\eta_2^2}{K(1-\gamma)^2}\end{aligned}$$

which further implies,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_g^{(t)}(\rho) - b\right)\left(\lambda^{(t)} - \lambda\right)\right] \leq \frac{1}{2\eta_2 T}\mathbb{E}\left[\left|\lambda^{(0)} - \lambda\right|^2\right] + \frac{\eta_2}{(1-\gamma)^2} + \frac{\eta_2}{2K(1-\gamma)^2}$$

where we use $\mathbb{E}\big[\widehat{V}_g^{(t)}(\rho)\big] = V_g^{(t)}(\rho)$ and $\lambda^{(t)}$ is independent of $\widehat{V}_g^{(t)}(\rho)$ given $\theta^{(t-1)}$. We now add the above inequality into (35) and note $V_g^\star(\rho) \geq b$,

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right)\right] + \lambda\,\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right)\right] \\ &\leq \frac{\log|A|}{\eta_1 T} + \frac{1}{(1-\gamma)^2 T} + \frac{2\eta_2}{(1-\gamma)^3} + \frac{1}{2\eta_2 T}\left|\lambda^{(0)} - \lambda\right|^2 + \frac{\eta_2}{(1-\gamma)^2} + \frac{\eta_2}{2K(1-\gamma)^2}.\end{aligned}$$

We take $\lambda = \frac{2}{(1-\gamma)\xi}$ when $\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right) \geq 0$; otherwise $\lambda = 0$. Thus,

$$\begin{aligned}&\mathbb{E}\left[V_r^\star(\rho) - \frac{1}{T}\sum_{t=0}^{T-1}V_r^{(t)}(\rho)\right] + \frac{2}{(1-\gamma)\xi}\mathbb{E}\left[b - \frac{1}{T}\sum_{t=0}^{T-1}V_g^{(t)}(\rho)\right]_+ \\ &\leq \frac{\log|A|}{\eta_1 T} + \frac{1}{(1-\gamma)^2 T} + \frac{2\eta_2}{(1-\gamma)^3} + \frac{1}{2\eta_2(1-\gamma)^2\xi^2 T} + \frac{\eta_2}{(1-\gamma)^2} + \frac{\eta_2}{2K(1-\gamma)^2}.\end{aligned}$$

Similar to the proof of Theorem 1, there exists a policy π' such that $V_r^{\pi'}(\rho) = \frac{1}{T}\sum_{t=0}^{T-1} V_r^{(t)}(\rho)$ and $V_g^{\pi'}(\rho) = \frac{1}{T}\sum_{t=0}^{T-1} V_g^{(t)}(\rho)$. Therefore,

$$\begin{aligned} &\mathbb{E}\left[V_r^\star(\rho) \,-\, V_r^{\pi'}(\rho)\right] \,+\, \frac{2}{(1-\gamma)\xi}\mathbb{E}\left[b \,-\, V_g^{\pi'}(\rho)\right]_+ \\ &\leq\; \frac{\log|A|}{\eta_1 T} \,+\, \frac{1}{(1-\gamma)^2 T} \,+\, \frac{2\eta_2}{(1-\gamma)^3} \,+\, \frac{1}{2\eta_2(1-\gamma)^2\xi^2 T} \,+\, \frac{\eta_2}{(1-\gamma)^2} \,+\, \frac{\eta_2}{2K(1-\gamma)^2}. \end{aligned}$$

According to Theorem 6 in Appendix C, we obtain

$$\mathbb{E}\left[b \,-\, V_g^{\pi'}(\rho)\right]_+ \;\leq\; \frac{\xi\log|A|}{\eta_1 T} + \frac{\xi}{(1-\gamma)T} + \frac{2\eta_2\xi}{(1-\gamma)^2} + \frac{1}{2\eta_2(1-\gamma)\xi T} + \frac{\eta_2\xi}{(1-\gamma)} + \frac{\eta_2\xi}{2K(1-\gamma)}$$

which shows the constraint violation bound by noting $\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right) = b - V_g^{\pi'}(\rho)$ and taking $\eta_1 = 2\log|A|$ and $\eta_2 = \frac{1-\gamma}{\sqrt{T}}$.

J Experimental Results

In this section, we provide additional experimental results to support our convergence theory. Our CMDP simulation is based on the shared MDP code [9]. We generate CMDPs with random transitions, uniform rewards, and utilities in $[0, 1]$. We simulate our algorithms with random initializations. Given $T > 0$, the total number of optimization iterations, our stepsizes in theorems become constants and multiplying them with positive constants does not affect convergence rates.

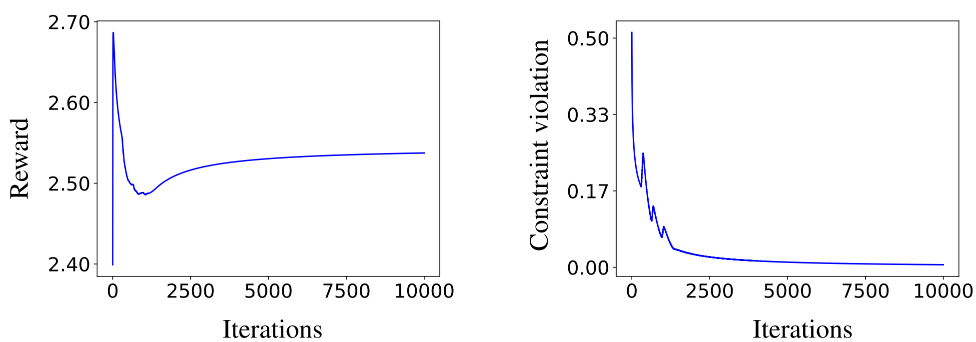

Figure 4: Convergence of the NPG-PD method (10). In this experiment, we have randomly generated a CMDP with $|S| = 20$, $|A| = 10$, $\gamma = 0.8$, and $b = 3$, and chosen: $\eta_1 = \eta_2 = 0.1$ and $d = 150$.

We show simulation results for algorithms with the general smooth parametrization. We consider a class of linear softmax policies,

$$\pi_\theta(a\,|\,s) \;=\; \frac{\exp(\theta\cdot\phi_{s,a})}{\sum_{a'\,\in\, A}\exp(\theta\cdot\phi_{s',a'})}$$

where $\phi_{s,a} \in \mathbb{R}^d$ is the feature map with $\|\phi_{s,a}\| \leq \beta$. We compute $\nabla_\theta \log\pi_\theta(a\,|\,s) = \phi_{s,a} - \mathbb{E}_{a'\,\sim\,\pi_\theta(\cdot\,|\,s)}[\phi_{s,a'}] := \widetilde{\phi}_{s,a}$ and the compatible function approximation error,

$$E^\nu(w;\theta,\lambda) \;=\; \mathbb{E}_{s,a\,\sim\,\nu}\left[\left(A_L^{\pi_\theta,\lambda}(s,a) - w\cdot\widetilde{\phi}_{s,a}\right)^2\right].$$

In this experiment, we take d canonical bases in $\mathbb{R}^d$ as our feature maps. Since $d < |S||A|$, they can't capture the advantage function and will introduce function approximation errors. In Figure 4, we only show the convergence of the reward value function to a stationary value that could be sub-optimal due to the function approximation error. By contrast, the constraint violation converges to zero sublinearly. It verifies Theorem 2 that the function approximation error does not dominate the constraint violation.

Last but not least, we show the objective and the constraint violation for running the sample-based NPG-PD algorithm (10): Algorithm 1, using two different sample sizes. We see that both reward value functions converge, and both constraint violations decrease to be negative. The large sample size of $K = 200$ performs better, especially for the constraint violation. It confirms Theorem 3 that the constraint violation is insusceptible to the function approximation error.

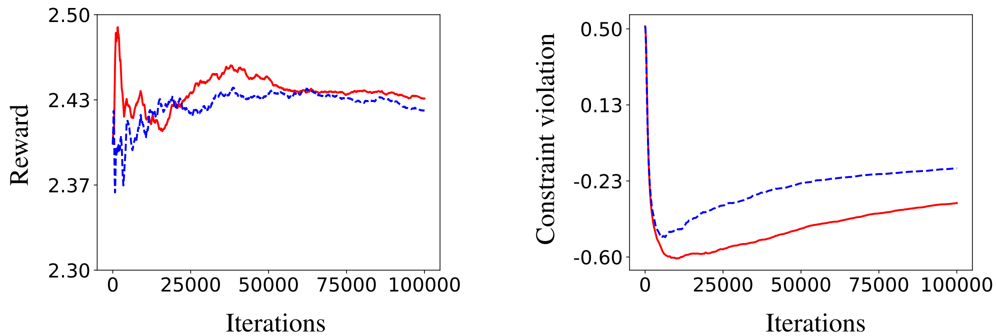

Figure 5: Convergence of the sample-based NPG-PD algorithm (10): Algorithm 1, using different sample sizes: $K = 100$ (– –) and $K = 200$ (—). In this experiment, we have randomly generated a CMDP with $|S| = 20$, $|A| = 10$, $\gamma = 0.8$, and $b = 3$. We have chosen parameters for Algorithm 1: $\eta_1 = \eta_2 = 0.1$, $\alpha = 0.1$, and $d = 150$.