Journal of Machine Learning Research 26 (2025) 1-76 Submitted 6/22; Revised 8/24; Published 10/25

Convergence and Sample Complexity of Natural Policy Gradient Primal-Dual Methods for Constrained MDPs

Dongsheng Ding DONGSHED@UTK.EDU
University of Tennessee, Knoxville
Knoxville, TN 37996, USA

Kaiqing Zhang KAIQING@UMD.EDU
University of Maryland, College Park
College Park, MD 20742, USA

Jiali Duan JDUAN24@APPLE.COM
Apple Inc.
Cupertino, CA 95014, USA

Tamer Başar BASAR1@ILLINOIS.EDU
University of Illinois Urbana-Champaign
Champaign, IL 61820, USA

Mihailo R. Jovanović MIHAILO@USC.EDU
University of Southern California
Los Angeles, CA 90089, USA

Editor: Ambuj Tewari

Abstract

We study the sequential decision making problem of maximizing 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 optimal control problem for Constrained Markov Decision Processes (constrained MDPs). Specifically, we propose a new Natural Policy Gradient Primal-Dual (NPG-PD) method that updates the primal variable via natural policy gradient ascent and the dual variable via projected subgradient descent. Although 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 convergence is independent of the size of the state-action space, i.e., it is dimension-free. Furthermore, for log-linear and general smooth policy parametrizations, we establish sublinear convergence rates up to a function approximation error caused by restricted policy parametrization. We also provide convergence and finite-sample complexity guarantees for two sample-based NPG-PD algorithms. We use a set of computational experiments to showcase the effectiveness of our approach.

Keywords: Constrained Markov decision processes, Natural policy gradient, Constrained nonconvex optimization, Method of Lagrange multipliers, Primal-dual algorithms

1 Introduction

Reinforcement learning (RL) studies sequential decision-making problems with the objective of maximizing the expected total reward while interacting with an unknown environment (Sutton and Barto, 2018). Markov Decision Processes (MDPs) are typically used to model the dynamics of the environment. However, in many safety-critical applications, e.g., in autonomous driving (Fisac et al., 2018), robotics (Ono et al., 2015), cyber-security (Zhang et al., 2019), and financial management (Abe et al., 2010), the control system is also subject to constraints on its utilities or costs. In this setting, constrained Markov Decision Processes (constrained MDPs) are used to model the environment dynamics (Altman, 1999) and, in addition to maximizing the expected total reward it is also important to take into account the constraints on the expected total utility/cost as an extra learning objective.

Policy gradient (PG) (Sutton et al., 2000) and natural policy gradient (NPG) (Kakade, 2002) methods have enjoyed substantial empirical successes in solving MDPs (Schulman et al., 2015; Lillicrap et al., 2015; Mnih et al., 2016; Schulman et al., 2017; Sutton and Barto, 2018). PG methods, or more generally *direct policy search* methods, have also been used to solve constrained MDPs (Uchibe and Doya, 2007; Borkar, 2005; Bhatnagar and Lakshmanan, 2012; Chow et al., 2017; Tessler et al., 2019; Liang et al., 2018; Paternain et al., 2022; Achiam et al., 2017; Spooner and Savani, 2020), but most existing theoretical guarantees are asymptotic and/or only provide local convergence around stationary-point policies. On the other hand, it is desirable to show that, for arbitrary initial condition, a solution that enjoys ϵ-optimality gap and ϵ-constraint violation is computed using a finite number of iterations and/or samples. It is thus imperative to establish global convergence guarantees for PG methods when solving constrained MDPs.

In this work, we provide a theoretical foundation for the non-asymptotic global convergence of PG methods for solving constrained MDPs, and answer the following questions:

(i) Can we employ PG methods to solve optimal control problems for constrained MDPs?

(ii) Do PG methods converge to the globally optimal solution that satisfies constraints?

(iii) What is the convergence rate of PG methods and the effect of the function approximation error caused by a restricted policy parametrization?

(iv) What is the sample complexity of model-free PG methods?

1.1 Preview of key contributions

Our key contributions are:

(i) We propose a simple but effective primal-dual policy gradient algorithm for solving discounted infinite-horizon optimal control problems for constrained MDPs. Our Natural Policy Gradient Primal-Dual (NPG-PD) method employs natural policy gradient ascent to update the primal variable and projected subgradient descent to update the dual variable.

(ii) We exploit the structure of the softmax policy parametrization to establish global convergence guarantees in spite of the fact that the objective function in the maximization problem is not concave and the constraint set is not convex. In particular, we prove that our NPG-PD method achieves global convergence at a rate of $O(1/\sqrt{T})$ for 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) We establish sublinear convergence at a rate of $O(1/\sqrt{T})$ in both the optimality gap and the constraint violation for log-linear and general smooth policy parametrizations, up to a function approximation error caused by restricted policy parametrization. This is accomplished by providing a new regret-type primal-dual analysis in the function approximation case, thereby eliminating the need for in-policy class comparison.

(iv) We provide convergence and finite-sample complexity guarantees for two sample-based NPG-PD algorithms. The new sample complexity of $O(1/\epsilon^4)$ for generating an ϵ-optimal policy results from our new regret-type primal-dual analysis, accompanied by a practical stochastic gradient ascent method. We utilize a set of computational experiments to showcase the effectiveness of our approach.

At this point it is worth highlighting the main differences between the results of this paper (as an extended version of our earlier NeurIPS paper (Ding et al., 2020)) and those in Ding et al. (2020). Although our algorithmic framework here builds on the NPG-PD method of Ding et al. (2020), our new characterization of function approximation error, which is based on the estimation-transfer error decomposition in both the optimality gap and the constraint violation, facilitates the derivation of convergence and sample complexity results described in (iii) and (iv) above. In contrast, the earlier function approximation study of Ding et al. (2020) utilizes the classical notion of *compatible function approximation* which is often challenging to control. Furthermore, Ding et al. (2020) adopts a standard drift analysis of constraint violation in online optimization, assuming in-policy class feasibility. This approach not only yields sub-optimal rates relative to the tabular case, but it also leaves the optimality of a comparison policy within policy class unjustified.

Our results as summarized in (iii) and (iv) above extend the PG methods studied in Agarwal et al. (2021) and provide a novel contribution to the constrained MDP setting. In contrast to our earlier work (Ding et al., 2020), we establish here the optimal rate for log-linear and general smooth policy parameterizations up to a function approximation error, and eliminate the in-policy class feasibility assumption. By providing a new regret-type primal-dual analysis, we show that the derived rate for the function approximation case matches the optimal rate for the tabular case. Furthermore, in contrast to the sample complexity result of Ding et al. (2020), which holds only when the estimates of value functions are bounded, we employ and analyze a more practical version of stochastic gradient method that does not require this boundedness assumption. Our new analysis allows us to establish an improved sample complexity of $O(1/\epsilon^4)$, compared to the previous $O(1/\epsilon^8)$, where ϵ is the desired level of accuracy.

In addition to these technical differences, we also characterize the zero constraint violation performance (on average) of our method in both the tabular and the function approximation settings, and conduct computational experiments on a set of benchmark robotic simulation tasks to demonstrate the effectiveness of our approach. These are all new results compared to our conference version (Ding et al., 2020).

1.2 Related work

Our work builds on Lagrangian-based constrained MDP algorithms (Altman, 1999; Abad et al., 2002; Abad and Krishnamurthy, 2003; Borkar, 2005). However, convergence guarantees of these algorithms are either local (to stationary-point or locally optimal policies) (Bhatnagar and Lakshmanan, 2012; Chow et al., 2017; Tessler et al., 2019) or asymptotic (Borkar, 2005). In the tabular setting, we compare the convergence rates in Table 1 by assuming the exact evaluation of policy gradients. When function approximation is used for policy parametrization, Yu et al. (2019) 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 convergence to a globally optimal solution in spite of the lack of convexity. References (Paternain et al., 2019, 2022) are closely related to our work. Paternain et al. (2019) provided a duality analysis for constrained MDPs in the policy space and proposed a provably convergent dual descent algorithm by assuming access to a nonconvex optimization oracle. However, it is not clear how to obtain a solution to a primal nonconvex optimization problem and the global convergence guarantees are not established. Paternain et al. (2022) proposed a primal-dual algorithm and provided empirical results, but did not offer a convergence analysis. In spite of the lack of convexity, our work provides global convergence guarantees for a new primal-dual algorithm without using any optimization oracles. For the function approximation setting, we compare the convergence rates and sample complexities in Table 2. Other related policy optimization methods include CPG (Uchibe and Doya, 2007), accelerated PDPO (Liang et al., 2018), CPO (Achiam et al., 2017; Yang et al., 2020b), FOCOPS (Zhang et al., 2020c), IPPO (Liu et al., 2020b), and CUP (Yang et al., 2022) but theoretical guarantees for these algorithms are still lacking. Recently, optimism principles have been used for efficient exploration in constrained MDPs (Singh et al., 2022; Zheng and Ratliff, 2020; Ding et al., 2021; Qiu et al., 2020; Efroni et al., 2020; Bai et al., 2023a; Yu et al., 2021; Liu et al., 2021a; Wei et al., 2022). In comparison, our work focuses on the optimization landscape within a primal-dual framework in both model-based and model-free settings.

Our work is also pertinent to the global convergence results of PG methods. Fazel et al. (2018); Malik et al. (2020); Mohammadi et al. (2019, 2020, 2021, 2022) provided global convergence guarantees and quantified sample complexity of (natural) PG methods for the nonconvex linear quadratic regulator problem of both discrete- and continuous-time systems. Zhang et al. (2020b) showed that locally optimal policies for MDPs are achievable using PG methods with reward reshaping. Wang et al. (2019) demonstrated that (natural) PG methods converge to the globally optimal value when overparametrized neural networks are used. A variant of NPG, trust-region policy optimization (TRPO) (Schulman et al., 2015), converges to the globally optimal policy with overparametrized neural networks (Liu et al., 2019) and for regularized MDPs (Shani et al., 2020). Bhandari and Russo (2024, 2021) studied global optimality and convergence of PG methods from a policy iteration perspective. Agarwal et al. (2021) characterized global convergence properties of (natural) PG methods and studied computational, approximation, and sample size issues. Additional recent advances along these lines include (Mei et al., 2020; Zhang et al., 2020a; Cen et al., 2022; Liu et al., 2020a; Khodadadian et al., 2022). While all these references handled the lack

Algorithm	Iteration/Sample complexities
PG-PD (Abad et al., 2002)	asymptotic
PG-PD (Borkar, 2005)	asymptotic
NPG-PD (Theorem 10, Theorem 29)	$O\left(1/\epsilon^2\right)$ / $O\left(1/\epsilon^4\right)$

Table 1: Complexity comparison of our NPG-PD method with closely related algorithms for the tabular case with finitely many states/actions. The iteration complexity is determined by the number of gradient-based updates that an algorithm takes to achieve ϵ-optimality gap and ϵ-constraint violation, $\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right) \leq \epsilon$ and $\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right)\right]_+ \leq \epsilon$, and the sample complexity is determined by the number of trajectory rollouts.

of convexity in the objective function, additional effort is required to deal with *nonconvex constraints* that arise in constrained MDPs. Our paper addresses this challenge.

We also remark on some recent work on Lagrangian-based policy optimization. Liu et al. (2021b); Li et al. (2024); Ying et al. (2022) examined a two-timescale scheme for updating the primal-dual variables by updating policy in an inner loop via an NPG-style subroutine for each dual iterate. In spite of the improved convergence that results from the proposed modifications of the Lagrangian and the dual update, the double-loop scheme often increases computational cost and introduces difficulty in parameter tuning. Additionally, Liu et al. (2021b,a); Bai et al. (2022) proposed policy optimization algorithms that offer zero constraint violation at the end of training, and Ding et al. (2023); Müller et al. (2024); Montenegro et al. (2024) introduced regularization to single-timescale primal-dual algorithms that achieve policy-iterate convergence, which is orthogonal to our work.

1.3 Paper outline

In Section 2, we formulate an optimal control problem for constrained Markov decision processes and provide necessary background material. In Section 3, we describe our natural policy gradient primal-dual method. We provide convergence guarantees for our algorithm under the tabular softmax policy parametrization in Section 4 and under log-linear and general smooth policy parametrizations in Section 5. We establish convergence and finite-sample complexity guarantees for two model-free primal-dual algorithms in Section 6 and provide computational experiments in Section 7. We close the paper with remarks in Section 8.

2 Problem setup

In Section 2.1, we introduce constrained Markov decision processes. In Section 2.2, we present the method of Lagrange multipliers, formulate a saddle-point problem for the constrained

Algorithm	**Iteration/Sample complexities**
PDO (Chow et al., 2017)	asymptotic
RCPO (Tessler et al., 2019)	asymptotic
CBP (Jain et al., 2022)	$O\left(1/\epsilon^2\right)$ / —
C-NPG-PD (Bai et al., 2023b)	$O\left(1/\epsilon^2\right)$ / $O\left(1/\epsilon^4\right)$
NPG-PD (Theorem 16, Theorem 29)	$O\left(1/\epsilon^2\right)$ / $O\left(1/\epsilon^4\right)$
NPG-PD (Theorem 24, Theorem 28)	$O\left(1/\epsilon^2\right)$ / $O\left(1/\epsilon^4\right)$

Table 2: Complexity comparison of our NPG-PD method with closely related algorithms for the function approximation case with potentially infinitely many states/actions. The iteration complexity is determined by the number of gradient-based iterations an algorithm takes to ensure ϵ-optimality gap and ϵ-constraint violation up to a function approximation error ϵ_{fa}, $\mathbb{E}\big[\,\frac{1}{T}\sum_{t=0}^{T-1}\big(V_r^\star(\rho)-V_r^{(t)}(\rho)\big)\,\big]\leq\epsilon+\sqrt{\epsilon_{\text{fa}}}$ and $\mathbb{E}\big[\,\big[\,\frac{1}{T}\sum_{t=0}^{T-1}\big(b-V_g^{(t)}(\rho)\big)\,\big]_+\,\big]\leq\epsilon+\sqrt{\epsilon_{\text{fa}}}$. The sample complexity is determined by the number of trajectory rollouts to ensure ϵ-optimality gap and ϵ-constraint violation up to a function approximation error ϵ_{fa} that is given either by the bias error $\bar{\epsilon}_{\text{bias}}$ (C-NPG-PD) or the transfer error ϵ_{bias} (NPG-PD). The bias error $\bar{\epsilon}_{\text{bias}}$ contains the transfer error ϵ_{bias}, which captures how well an approximation function class covers the true value function, and the error of policy representation.

policy optimization, and exhibit several problem properties: strong duality, boundedness of the optimal dual variable, and constraint violation. In Section 2.3, we introduce a parametrized formulation of the constrained policy optimization problem, provide an example of a constrained MDP that is not convex, and present several useful policy parametrizations.

2.1 Constrained Markov decision processes

We consider an infinite-horizon discounted Constrained Markov Decision Process (Piunovskiy, 1997; Altman, 1999):

$$\text{CMDP}\,(\,S,\,A,\,P,\,r,\,g,\,b,\,\gamma,\,\rho\,)$$

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 $a \in A$, $r\colon\, S\times A \to [0,1]$ is a reward function, $g\colon\, S\times A\to[0,1]$ is a utility function, b is a constraint offset, $\gamma\in[0,1)$ is a discount factor, and ρ is an initial distribution over S.

For any state s_t, a stochastic policy $\pi\colon\, S\to\Delta_A$ is a function in the probability simplex Δ_A over the action space A, i.e., $a_t\sim\pi(\cdot\,|\,s_t)$ at time t. Let Π be a set of all possible policies. A policy $\pi\in\Pi$, together with the 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, $V_g^\pi\colon\, S\to\mathbb{R}$ associated with the reward r or the utility g are determined by the expected values of total discounted rewards or utilities received under policy π:

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

where the expectation $\mathbb{E}$ is taken over the randomness of the trajectory τ induced by π. Starting from an arbitrary state-action pair (s,a) and following a policy π, we also introduce the state-action value functions $Q_r^\pi(s,a)$, $Q_g^\pi(s,a)\colon\, S\times A\to\mathbb{R}$ together with their advantage functions A_r^π, $A_g^\pi\colon\, S\times A\to\mathbb{R}$:

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

where the symbol $\diamond$ represents either r or g. Since r, $g\in[0,1]$, we have

$$V_\diamond^\pi(s)\;\in\;\left[0,\,\frac{1}{1-\gamma}\right]$$

and their expected values under the initial distribution ρ are determined by

$$V_\diamond^\pi(\rho)\;:=\;\mathbb{E}_{s_0\sim\rho}\,[\,V_\diamond^\pi(s_0)\,]\,.$$

Having defined a policy as well as the state-action value functions for the discounted constrained MDP, the objective is to find a policy that maximizes the expected reward value over all policies, subject to a constraint on the expected utility value:

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

In view of the aforementioned boundedness of $V_r^{\pi}(s)$ and $V_g^{\pi}(s)$, we set the constraint offset $b \in (0, 1/(1-\gamma)]$ to make Problem (1) meaningful.

Remark 1 *For notational convenience, we consider a single constraint in Problem (1), but our convergence guarantees are readily generalizable to problems with multiple constraints.*

2.2 Method of Lagrange multipliers

By dualizing constraints (Luenberger and Ye, 1984; Bertsekas, 2014), we cast Problem (1) into the following max-min problem:

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

where $V_L^{\pi,\lambda}(\rho) := V_r^{\pi}(\rho) + \lambda\,(V_g^{\pi}(\rho) - b)$ is the Lagrangian of Problem (1), π is the primal variable, and λ is the Lagrange multiplier or dual variable which is nonnegative. The associated dual objective function is defined as

$$V_D^{\lambda}(\rho) \;:=\; \underset{\pi \in \Pi}{\text{maximize}} \; V_L^{\pi,\lambda}(\rho).$$

Instead of utilizing the linear-programming-based method (Piunovskiy, 1997; Altman, 1999), we employ *direct policy search* method to solve Problem (2). Direct policy search is attractive for three reasons: (i) it allows us to directly optimize/monitor the value functions that we are interested in; (ii) it can deal with large state-action spaces via policy parameterization, e.g., neural nets; and (iii) it can utilize policy gradient estimates via simulations of the policy. Since Problem (1) is a nonconcave constrained maximization problem and the policy space is often infinite-dimensional, Problems (1) and (2) are challenging.

In spite of these challenges, Problem (1) has nice properties in the policy space when it is strictly feasible. We adapt the standard Slater condition in constrained optimization (Bertsekas, 2014) and assume strict feasibility of Problem (1) throughout the paper.

Assumption 2 (Slater condition) *There exist a constant $\xi > 0$ and a policy $\bar{\pi} \in \Pi$ such that $V_g^{\bar{\pi}}(\rho) - b \geq \xi$ holds.*

The Slater condition is mild in practice because 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.

Let $\pi^\star$ denote an optimal solution to Problem (1), let $\lambda^\star$ be an optimal dual variable:

$$\lambda^\star \;\in\; \underset{\lambda \geq 0}{\text{argmin}} \; V_D^{\lambda}(\rho)$$

and let the set of all optimal dual variables be $\Lambda^\star$. 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. We recall the strong duality for constrained MDPs and we prove boundedness of an optimal dual variable $\lambda^\star$.

Lemma 3 (Strong duality and boundedness of $\lambda^\star$) *Let Assumption 2 hold. Then*

(i) $V_r^\star(\rho) \,=\, V_D^\star(\rho)$;

(ii) $0 \,\leq\, \lambda^\star \,\leq\, \left(V_r^\star(\rho) - V_r^{\bar{\pi}}(\rho)\right)/\xi$.

Proof The proof of (i) is standard; e.g., see Altman (1999, Theorem 3.6) or Paternain et al. (2022, Theorem 3). The proof of (ii) builds on the constrained convex optimization (Beck, 2017, Section 8.5). Let $\Lambda_a := \{\lambda \geq 0 \,|\, V_D^\lambda(\rho) \leq a\}$ be a sublevel set of the dual objective function for $a \in \mathbb{R}$. For any $\lambda \in \Lambda_a$, we have

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

where $\bar{\pi}$ is a Slater point. Thus, $\lambda \leq \left(a - V_r^{\bar{\pi}}(\rho)\right)/\xi$. If we take $a = V_r^\star(\rho) = V_D^\star(\rho)$, then $\Lambda_a = \Lambda^\star$, which proves (ii). ■

Remark 4 *In the proof of Lemma 3 (ii), we choose a particular sublevel set of the dual objective function based on the strong duality (i). However, the boundedness of an optimal dual variable $\lambda^\star$ does not necessarily depend on the strong duality (i). In general, we have weak duality $V_D^\star(\rho) \geq V_r^\star(\rho)$. In this case, the choice of $a = V_r^\star(\rho)$ yields an empty sublevel set $\Lambda_a = \emptyset$, and the choice of $a = V_D^\star(\rho)$ leads to $0 \leq \lambda^\star \leq \left(V_D^\star(\rho) - V_r^{\bar{\pi}}(\rho)\right)/\xi$, which depends on an optimal dual variable.*

Let the value function associated with Problem (1) be determined by

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

Using the concavity of $v(\tau)$ (e.g., see Paternain et al. (2019, Proposition 1)), in Lemma 5 we establish a bound on the constraint violation, thereby extending a result from the constrained convex optimization (Beck, 2017, Section 8.5) to a constrained nonconvex setting.

Lemma 5 (Constraint violation) *Let Assumption 2 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 $[\,b - V_g^\pi(\rho)\,]_+ \leq 2\delta/C$, where $[\,x\,]_+ := \max(x, 0)$.*

Proof By the definition of $v(\tau)$, we have $v(0) = V_r^\star(\rho)$. We also note that $v(\tau)$ is concave (see the proof of Paternain et al. (2019, Proposition 1)). First, we show that $-\lambda^\star \in \partial v(0)$. By the definition of $V_L^{\pi,\lambda}(\rho)$ and the strong duality in Lemma 3,

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

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

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

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

$$v(0) \, - \, \tau\lambda^\star \;\geq\; v(\tau) \tag{3}$$

and, thus, $-\lambda^\star \in \partial v(0)$.

On the other hand, if we take $\tau = -\left[b - V_g^\pi(\rho)\right]_+$, then

$$V_r^\pi(\rho) \;\leq\; V_r^\star(\rho) \;=\; v(0) \;\leq\; v(\tau). \tag{4}$$

Combining (3) and (4) yields $V_r^\pi(\rho) - V_r^\star(\rho) \leq -\tau\lambda^\star$. Thus,

$$(\, C \, - \, \lambda^\star \,) \, |\tau| \;=\; -\,\lambda^\star \, |\tau| \, + \, C \, |\tau| \;=\; \tau\lambda^\star \, + \, C \, |\tau| \;\leq\; V_r^\star(\rho) \, - \, V_r^\pi(\rho) \, + \, C \, |\tau|$$

which completes the proof by applying the assumed condition on π. ■

Aided by the above properties implied by the Slater condition, we target the max-min Problem (2) using the primal-dual method.

2.3 Policy parametrization

Introduction of a set of parametrized policies $\{\pi_\theta \,|\, \theta \in \Theta\}$ brings Problem (1) into a constrained optimization problem over a finite-dimensional parameter space Θ:

$$\begin{aligned} \underset{\theta \,\in\, \Theta}{\text{maximize}} \quad & V_r^{\pi_\theta}(\rho) \\ \text{subject to} \quad & V_g^{\pi_\theta}(\rho) \;\geq\; b. \end{aligned} \tag{5}$$

A parametric version of Problem (2) is given by

$$\underset{\theta \,\in\, \Theta}{\text{maximize}} \;\; \underset{\lambda \,\geq\, 0}{\text{minimize}} \;\; V_r^{\pi_\theta}(\rho) \, + \, \lambda \left(V_g^{\pi_\theta}(\rho) \, - \, b \right) \tag{6}$$

where $V_L^{\pi_\theta,\lambda}(\rho) := V_r^{\pi_\theta}(\rho) + \lambda(V_g^{\pi_\theta}(\rho) - b)$ is the associated Lagrangian and λ is the Lagrange multiplier. The dual objective function is determined by $V_D^\lambda(\rho) := \text{maximize}_\theta \, V_L^{\pi_\theta,\lambda}(\rho)$. The primal maximization problem (5) is finite-dimensional but not concave, even in the absence of a constraint (Agarwal et al., 2021). In Lemma 6 we prove that, in general, Problem (5) is not convex because it involves maximization of a nonconcave objective function over a nonconvex constraint set. The proof is provided in Appendix A and it utilizes an example of a constrained MDP in Figure 1.

Lemma 6 (Lack of convexity) *There exists a constrained MDP for which the objective function $V_r^{\pi_\theta}(s)$ in Problem (5) is not concave and the constraint set $\{\theta \in \Theta \,|\, V_g^{\pi_\theta}(s) \geq b\}$ is not convex.*

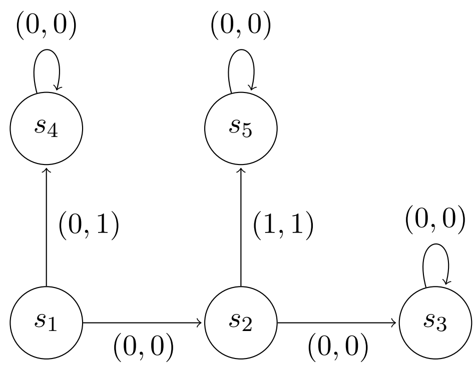

Figure 1: An example of a constrained MDP for which the objective function $V_r^{\pi_\theta}(s)$ in Problem (5) is not concave and the constraint set $\{\theta \in \Theta \,|\, V_g^{\pi_\theta}(s) \geq b\}$ is not convex. A pair (r, g) associated with a directed arrow represents the (reward, utility) received when an action in a certain state is taken. This example is utilized in the proof of Lemma 6.

In general, the Lagrangian $V_L^{\pi_\theta,\lambda}(\rho)$ in Problem (6) is convex in λ but not concave in θ. While many algorithms for solving max-min optimization problems, e.g., those proposed in Lin et al. (2020); Nouiehed et al. (2019); Yang et al. (2020a), require extra assumptions on the max-min structure or only guarantee convergence to a stationary point, we exploit problem structure and propose a new primal-dual method to compute a globally optimal solution to Problem (6). Before doing that, we first introduce several useful classes of policies.

2.3.1 Direct policy parametrization

A direct parametrization of a policy is a probability distribution:

$$\pi_\theta(a\,|\,s) \;=\; \theta_{s,a} \;\;\text{for all}\; \theta \,\in\, \Delta_A^{|S|}$$

where $\theta_s \in \Delta_A$ for any $s \in S$, i.e., $\theta_{s,a} \geq 0$ and $\sum_{a\,\in\,A}\theta_{s,a} = 1$. This policy class is complete since it directly represents any stochastic policy. Even though it is a challenging class of policy to work with from both theoretical and computational viewpoints (Mei et al., 2020; Agarwal et al., 2021), it offers a useful sanity check for many policy search methods.

2.3.2 Softmax policy parametrization

This class of policies is parametrized by a softmax function:

$$\pi_\theta(a\,|\,s) \;=\; \frac{\exp(\theta_{s,a})}{\sum_{a'\in A}\exp(\theta_{s,a'})} \;\text{ for all }\; \theta\in\mathbb{R}^{|S||A|}. \tag{7}$$

The softmax policy can be used to represent any stochastic policy, and its closure contains all stationary policies. It has been utilized to study the convergence properties of many RL algorithms (Bhandari and Russo, 2024; Agarwal et al., 2021; Mei et al., 2020; Cen et al., 2022; Khodadadian et al., 2022), and it offers several algorithmic advantages: (i) it equips the policy with a special structure so that the NPG update works like the classical multiplicative weights update in online learning (Freund and Schapire, 1997; Cesa-Bianchi and Lugosi, 2006); (ii) it can be used to interpret the function approximation error (Agarwal et al., 2021).

2.3.3 Log-linear policy parametrization

A log-linear policy is given by

$$\pi_\theta(a\,|\,s) \;=\; \frac{\exp(\theta^\top\phi_{s,a})}{\sum_{a'\in A}\exp(\theta^\top\phi_{s,a'})} \;\text{ for all }\; \theta\in\mathbb{R}^{d} \tag{8}$$

where $\phi_{s,a}\in\mathbb{R}^d$ is the feature map at a state-action pair (s,a). The log-linear policy builds on the softmax policy by applying the softmax function to a set of linear functions in a given feature space. More importantly, it exactly characterizes the linear function approximation via policy parametrization (Agarwal et al., 2021); see Miryoosefi and Jin (2022); Amani et al. (2021) for solving constrained MDPs with linear function approximation.

2.3.4 General policy parametrization

A general class of stochastic policies is given by $\{\pi_\theta\,|\,\theta\in\Theta\}$ with $\Theta\subset\mathbb{R}^d$ without specifying a parametric structure of the policy π_θ. The parameter space has dimension d and this policy class covers the settings that utilize nonlinear function approximation, such as (deep) neural networks (Liu et al., 2019; Wang et al., 2019).

When we choose the dimension d such that $d\ll|S||A|$ in either the log-linear policy or the general nonlinear policy, the policy class has limited expressiveness and may not contain all stochastic policies. Motivated by this observation, the theory that we develop in Section 5 establishes global convergence up to a function approximation error caused by restricted policy parametrization.

3 Natural policy gradient primal-dual method

In Section 3.1, we provide a brief summary of three basic optimization methods that have been used to solve the constrained policy optimization problem (5). In Section 3.2, we propose a natural policy gradient primal-dual method which represents an extension of the natural policy gradient method to the constrained problems.

3.1 Constrained policy optimization methods

We summarize three basic optimization methods that can be employed to solve the primal problem (5). We assume that the value function and the policy gradient can be evaluated exactly for any given policy.

We first introduce some useful definitions. The discounted visitation distribution $d_{s_0}^{\pi}$ of a policy π and its expectation over the initial distribution ρ are, respectively, given by

$$\begin{aligned} d_{s_0}^{\pi}(s) &:= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P^{\pi}(s_t = s \,|\, s_0) \\ d_{\rho}^{\pi}(s) &:= \mathbb{E}_{s_0 \sim \rho}\left[\, d_{s_0}^{\pi}(s)\,\right] \end{aligned} \tag{9}$$

where $P^{\pi}(s_t = s \,|\, s_0)$ is the probability of visiting state s at time t under the policy π with an initial state s_0. When the use of parametrized policy π_θ is clear from the context, we use $V_r^{\theta}(\rho)$ to denote $V_r^{\pi_\theta}(\rho)$. When $\pi_\theta(\cdot\,|\,s)$ is differentiable and when it belongs to the probability simplex, i.e., $\pi_\theta \in \Delta_A^{|S|}$ for all θ, the policy gradient of the Lagrangian (6) is determined by

$$\begin{aligned} \nabla_\theta V_L^{\theta,\lambda}(s_0) &= \nabla_\theta V_r^{\theta}(s_0) + \lambda \nabla_\theta V_g^{\theta}(s_0) \\ &= \frac{1}{1-\gamma}\,\mathbb{E}_{s \sim d_{s_0}^{\pi_\theta}}\,\mathbb{E}_{a \sim \pi_\theta(\cdot\,|\,s)}\left[\, A_L^{\theta,\lambda}(s,a)\nabla_\theta \log \pi_\theta(a\,|\,s)\,\right] \end{aligned} \tag{10}$$

where $A_L^{\theta,\lambda}(s,a) := A_r^{\theta}(s,a) + \lambda A_g^{\theta}(s,a)$.

3.1.1 Dual method

When the strong duality in Lemma 3 holds, it is convenient to work with the dual formulation of Problem (5):

$$\underset{\lambda \geq 0}{\text{minimize}}\;\; V_D^{\lambda}(\rho). \tag{11}$$

While the dual objective function is convex regardless of the concavity of the primal maximization problem (5), it is often non-differentiable (Bertsekas, 2008). Thus, a projected dual subgradient descent is used to solve the dual problem:

$$\lambda^{(t+1)} = \mathcal{P}_+\left(\lambda^{(t)} - \eta\, g_\lambda^{(t)}\right)$$

where $\mathcal{P}_+(\cdot)$ is the projection to the non-negative real axis, $\eta > 0$ is the stepsize, and $g_\lambda^{(t)}$ is a subgradient of the dual objective function evaluated at $\lambda = \lambda^{(t)}$, i.e., $g_\lambda^{(t)} \in \partial_\lambda V_D^{\lambda^{(t)}}(\rho)$.

The dual method works in the space of dual variables and it requires efficient evaluation of the subgradient of the dual objective function. We note that computing the dual objective function $V_D^{\lambda}(\rho)$ for a given $\lambda = \lambda^{(t)}$ in each step amounts to a standard unconstrained RL problem (Paternain et al., 2019). In spite of global convergence guarantees for several policy search methods in the tabular setting, it is challenging to obtain the dual objective function and/or to compute its subgradient, e.g., when the problem dimension is high and/or when the state space is continuous. Although the primal problem can be approximated using the first-order Taylor expansion (Achiam et al., 2017; Yang et al., 2020b), inverting Hessian matrices becomes a computational burden, and it is costly to implement the dual method.

3.1.2 Primal method

In the primal method, a policy search strategy works directly on the primal problem (5) by seeking an optimal policy in a feasible region. The key challenge is to ensure the feasibility of the next policy iterate in the search direction, which is similar to the use of the primal method in nonlinear programming (Luenberger and Ye, 1984).

An intuitive approach is to check the feasibility of each policy iterate and determine whether the constraint is active (Xu et al., 2021). If the policy iterate is feasible or the constraint is inactive, we move towards maximizing the single objective function; otherwise, we look for a feasible direction. For the softmax policy parametrization (7), this can be accomplished using a simple first-order gradient-based method:

$$\begin{aligned} \theta_{s,a}^{(t+1)} &= \theta_{s,a}^{(t)} + \eta\, G_{s,a}^{(t)}(\rho) \\ G_{s,a}^{(t)}(\rho) &:= \begin{cases} \dfrac{1}{1-\gamma}\, A_g^{(t)}(s,a), & \text{when } V_g^{(t)}(\rho) < b - \epsilon_b \\ \dfrac{1}{1-\gamma}\, A_r^{(t)}(s,a), & \text{when } V_g^{(t)}(\rho) \geq b - \epsilon_b \end{cases} \end{aligned} \tag{12}$$

where we use shorthand $A_r^{(t)}(s,a)$ and $A_g^{(t)}(s,a)$ to denote $A_r^{\theta^{(t)}}(s,a)$ and $A_g^{\theta^{(t)}}(s,a)$, respectively, $G_{s,a}^{(t)}(\rho)$ is the gradient-ascent direction determined by the scaled version of the advantage functions, and ϵ_b is the relaxation parameter for the constraint $V_g^{\pi_\theta}(\rho) \geq b$. When the iterate violates the relaxed constraint, $V_g^{\pi_\theta}(\rho) \geq b - \epsilon_b$ with $\epsilon_b > 0$, it maximizes the constraint function to gain feasibility. More reliable evaluation of the feasibility often demands a more tractable characterization of the constraint, e.g., by utilizing Lyapunov function (Chow et al., 2018), Gaussian process modeling (Sui et al., 2018), backward value function (Satija et al., 2020), and logarithmic penalty function (Liu et al., 2020b). Hence, the primal method offers the adaptability of adjusting a policy to satisfy the constraint, which is desirable in safe training applications. However, global convergence theory is still lacking, and recent progress (Xu et al., 2021) requires a careful relaxation of the constraint.

3.1.3 Primal-dual method

The primal-dual method simultaneously updates the primal and dual variables (Arrow, 1958). With the direct parametrization $\pi_\theta(a\,|\,s) = \theta_{s,a}$, a basic primal-dual method (Abad and Krishnamurthy, 2003) performs the following Policy Gradient Primal-Dual (PG-PD) update:

$$\begin{aligned} \theta^{(t+1)} &= \mathcal{P}_\Theta\left(\theta^{(t)} + \eta_1 \nabla_\theta V_L^{\theta^{(t)},\lambda^{(t)}}(\rho)\right) \\ \lambda^{(t+1)} &= \mathcal{P}_\Lambda\left(\lambda^{(t)} - \eta_2\left(V_g^{\theta^{(t)}}(\rho) - b\right)\right) \end{aligned} \tag{13}$$

where $\nabla_\theta V_L^{\theta^{(t)},\lambda^{(t)}}(\rho) := \nabla_\theta V_r^{\theta^{(t)}}(\rho) + \lambda^{(t)} \nabla_\theta V_g^{\theta^{(t)}}(\rho)$, $\eta_1 > 0$ and $\eta_2 > 0$ are the stepsizes, $\mathcal{P}_\Theta$ is the projection onto a probability simplex $\Theta := \Delta_A^{|S|}$, and $\mathcal{P}_\Lambda$ is the projection that will be specified later. For the max-min formulation (6), the PG-PD method (13) directly performs projected gradient ascent in the policy parameter θ and descent in the dual variable λ, both over the Lagrangian $V_L^{\theta,\lambda}(\rho)$. The primal-dual method overcomes the disadvantages of the primal and dual methods either by relaxing the precise calculation of the subgradient of the dual objective function, or by changing the descent direction via tuning of the dual variable. While this simple method provides a foundation for solving constrained MDPs (Chow et al., 2017; Tessler et al., 2019), the lack of convexity in (6) makes it challenging to establish the global convergence theory for the primal-dual method, which is our primary objective.

We first leverage the structure of the constrained policy optimization problem (5) to provide a positive result in terms of both the optimality gap and the constraint violation.

Theorem 7 (Restrictive convergence: direct policy parametrization) *Let Assumption 2 hold with a policy class $\{\pi_\theta = \theta \,|\, \theta \in \Theta\}$ and let $\Lambda = [\,0, 2/((1-\gamma)\xi)\,]$, $\rho > 0$, $\lambda^{(0)} = 0$, and $\theta^{(0)}$ be such that $V_r^{\theta^{(0)}}(\rho) \geq V_r^\star(\rho)$. If we choose $\eta_1 = \Theta(1)$ and $\eta_2 = \Theta(1/\sqrt{T})$, then the iterates $\theta^{(t)}$ generated by the PG-PD method (13) satisfy*

$$\text{(Optimality gap)} \quad \frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) \,-\, V_r^{(t)}(\rho)\right) \;\leq\; C_1 \frac{|A|\,|S|}{(1-\gamma)^6\, T^{1/4}} \left\| d_\rho^{\pi^\star}/\rho \right\|_\infty^2$$

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

where C_1 and C_2 are absolute constants that are independent of T.

For tabular constrained MDPs with direct policy parametrization, Theorem 7 guarantees that, on average, the optimality gap $V_r^\star(\rho) - V_r^{(t)}(\rho)$ and the constraint violation $b - V_g^{(t)}(\rho)$ decay to zero at a sublinear rate $1/T^{1/4}$. However, this rate explicitly depends on the sizes of state/action spaces $|S|$ and $|A|$, and the distribution shift $\|d_\rho^{\pi^\star}/\rho\|_\infty$, which characterizes the exploration factor. A careful initialization $\theta^{(0)}$ that satisfies $V_r^{\theta^{(0)}}(\rho) \geq V_r^{\theta^\star}(\rho)$ is also required. We leave it as future work to prove a tighter rate for this tabular setting.

The proof of Theorem 7, which first appeared in Ding et al. (2022a), is provided in Appendix B for completeness. We exploit the problem structure that casts the primal problem (5) as a linear program in the occupancy measure (Altman, 1999) and apply the convex optimization analysis. This method is not well-suited for large-scale problems, and the projection onto a high-dimensional probability simplex is not desirable in practice. We next introduce a natural policy gradient primal-dual method to overcome these challenges and provide stronger convergence guarantees.

3.2 Natural policy gradient primal-dual (NPG-PD) method

The Fisher information matrix induced by a parametrized policy π_θ, denoted

$$F_\rho(\theta) \;:=\; \mathbb{E}_{s \sim d_\rho^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot\,|\,s)} \left[\nabla_\theta \log \pi_\theta(a\,|\,s) \left(\nabla_\theta \log \pi_\theta(a\,|\,s)\right)^\top\right],$$

is now used in the update of the primal variable in our primal-dual algorithm. The expectations are taken over the randomness of the state-action trajectory induced by π_θ, and the Natural Policy Gradient Primal-Dual (NPG-PD) method for solving Problem (6) is given by

$$\begin{aligned} \theta^{(t+1)} \;&=\; \theta^{(t)} \,+\, \eta_1\, F_\rho^\dagger(\theta^{(t)})\, \nabla_\theta V_L^{\theta^{(t)},\lambda^{(t)}}(\rho) \\ \lambda^{(t+1)} \;&=\; \mathcal{P}_\Lambda\left(\lambda^{(t)} \,-\, \eta_2\left(V_g^{\theta^{(t)}}(\rho) \,-\, b\right)\right) \end{aligned} \tag{14}$$

where $\dagger$ denotes the Moore-Penrose inverse of a given matrix, $\mathcal{P}_\Lambda(\cdot)$ denotes the projection to an interval Λ that will be specified later, and (η_1, η_2) are the constant stepsizes in the updates of the primal and dual variables. The primal update $\theta^{(t+1)}$ is obtained using a preconditioned gradient ascent via the natural policy gradient $F_\rho^\dagger(\theta^{(t)})\nabla_\theta V_L^{(t)}(\rho)$ and it represents the policy gradient of the Lagrangian $V_L^{(t)}(\rho)$ in the manifold induced by the Fisher information matrix $F_\rho(\theta^{(t)})$. On the other hand, the dual update $\lambda^{(t+1)}$ is obtained using a projected subgradient descent by collecting the constraint violation $b - V_g^{(t)}(\rho)$, where, for brevity, we use $V_L^{(t)}(\rho)$ and $V_g^{(t)}(\rho)$ to denote $V_L^{\theta^{(t)},\lambda^{(t)}}(\rho)$ and $V_g^{\theta^{(t)}}(\rho)$, respectively.

Throughout the paper, we abbreviate the policy notation $\pi_{\theta^{(t)}}$ as $\pi^{(t)}$, or as $\pi_\theta^{(t)}$ when it is necessary to emphasize the dependence on θ.

In Section 4, we establish the global convergence of NPG-PD (14) under the softmax policy parametrization. In Section 5, we examine the general policy parametrization and, in Section 6, we analyze the sample complexity of two sample-based implementations of the NPG-PD method (14).

Remark 8 *The performance difference lemma (Kakade and Langford, 2002; Agarwal et al., 2021), which quantifies the difference between $V_\diamond^\pi(s_0)$ and $V_\diamond^{\pi'}(s_0)$ for any two policies π and π' and any state s_0 as follows*

$$V_\diamond^\pi(s_0) \,-\, V_\diamond^{\pi'}(s_0) \;=\; \frac{1}{1-\gamma}\,\mathbb{E}_{s\,\sim\, d_{s_0}^\pi,\, a\,\sim\,\pi(\cdot\,|\,s)}\left[\,A_\diamond^{\pi'}(s,a)\,\right],$$

is utilized in our analysis, where the symbol $\diamond$ denotes either r or g.

4 Tabular softmax parametrization: dimension-free global convergence

We first examine the NPG-PD method (14) under the softmax policy parametrization (7). Strong duality in Lemma 3 holds on the closure of the softmax policy class, because of the completeness of the softmax policy class. Even though the maximization problem (5) is not concave, we establish global convergence of our algorithm with dimension-independent convergence rates.

We first exploit the softmax policy structure to show that the primal update in (14) can be expressed in a more compact form; see Appendix C for the proof.

Lemma 9 (Primal update as MWU) *Let $\Lambda := [\,0, 2/((1-\gamma)\xi)\,]$ and let $A_L^{(t)}(s,a) := A_r^{(t)}(s,a) + \lambda^{(t)} A_g^{(t)}(s,a)$. Under the softmax parametrized policy (7), the NPG-PD algo-*

rithm (14) can be brought to the following form

$$
\begin{aligned}
\theta_{s,a}^{(t+1)} &= \theta_{s,a}^{(t)} + \frac{\eta_1}{1-\gamma} A_L^{(t)}(s,a) \\
\lambda^{(t+1)} &= \mathcal{P}_\Lambda \left(\lambda^{(t)} - \eta_2 \left(V_g^{(t)}(\rho) - b \right) \right).
\end{aligned}
\tag{15a}
$$

Furthermore, the primal update in (15a) can be equivalently expressed as

$$
\pi^{(t+1)}(a\,|\,s) \;=\; \pi^{(t)}(a\,|\,s)\, \frac{\exp\left(\frac{\eta_1}{1-\gamma} A_L^{(t)}(s,a)\right)}{Z^{(t)}(s)} \tag{15b}
$$

where $Z^{(t)}(s) := \sum_{a\,\in\,A} \pi^{(t)}(a\,|\,s) \exp\left(\frac{\eta_1}{1-\gamma} A_L^{(t)}(s,a)\right)$.

The primal update in (15a) does not depend on the state distribution $d_\rho^{\pi^{(t)}}$ that appears in the NPG-PD algorithm (14) through the policy gradient (10). This is because of the Moore-Penrose inverse of the Fisher information matrix in (14). Furthermore, the policy update (15b) is given by the multiplicative weights update (MWU), which is commonly used in online linear optimization (Cesa-Bianchi and Lugosi, 2006). In contrast, an advantage function appears in the MWU policy update at each iteration in (15b). Such a connection to MWU has been first identified in the unconstrained MDP case by Agarwal et al. (2021).

In Theorem 10, we establish global convergence of the NPG-PD algorithm (15a) with respect to both the optimality gap $V_r^\star(\rho) - V_r^{(t)}(\rho)$ and the constraint violation $b - V_g^{(t)}(\rho)$. Even though we set $\theta_{s,a}^{(0)} = 0$ and $\lambda^{(0)} = 0$ in the proof of Theorem 10 in Section 4.1, the global convergence can be established for arbitrary initial conditions.

Theorem 10 (Global convergence: softmax policy parametrization) *Let Assumption 2 hold for $\xi > 0$ and let us fix $T > 0$ and $\rho \in \Delta_S$. If we choose $\eta_1 = 2\log|A|$ and $\eta_2 = 2(1-\gamma)/\sqrt{T}$, then the iterates $\{\pi^{(t)}\}_{t=0}^{T-1}$ generated by the algorithm (15) satisfy*

$$
\begin{aligned}
&\text{(Optimality gap)} \quad \frac{1}{T}\sum_{t\,=\,0}^{T-1} \left(V_r^\star(\rho) - V_r^{(t)}(\rho) \right) \;\le\; \frac{7}{(1-\gamma)^2}\frac{1}{\sqrt{T}} \\
&\text{(Constraint violation)} \quad \left[\frac{1}{T}\sum_{t\,=\,0}^{T-1} \left(b - V_g^{(t)}(\rho) \right) \right]_+ \;\le\; \frac{2/\xi + 4\,\xi}{(1-\gamma)^2}\frac{1}{\sqrt{T}}.
\end{aligned}
$$

Theorem 10 demonstrates that, on average, the reward value function converges to its globally optimal value and that the constraint violation decays to zero. In other words, for a desired accuracy ϵ, it takes $O(1/\epsilon^2)$ iterations to compute the solution which is ϵ away from a globally optimal one (with respect to both the optimality gap and the constraint violation). We note that the required number of iterations only depends on the desired accuracy ϵ and is *independent of the sizes of the state and action spaces*. Although the maximization problem (5) is not concave, our rates of $(1/\sqrt{T}, 1/\sqrt{T})$ for the optimality gap and constraint violation outperform the classical ones of $(1/\sqrt{T}, 1/T^{3/4})$ (Mahdavi et al., 2012), and match the achievable rates for solving online *convex* minimization problems with *convex* constraint sets (Yu et al., 2017). Moreover, in contrast to the bounds established for the PG-PD algorithm (13) in Theorem 7, the bounds in Theorem 10 for the NPG-PD algorithm (14) under the softmax policy parameterization do not depend on the initial state distribution ρ.

As shown in Lemma 11 in Section 4.1, the reward and utility value functions are coupled, and the natural policy gradient method in the unconstrained setting does not provide monotonic improvement to either of them (Agarwal et al., 2021, Section 5.3). To address this challenge, we introduce a new line of nonconvex analysis by bridging the online regret analysis in unconstrained MDPs (Even-Dar et al., 2009; Agarwal et al., 2021) and the Lagrangian methods in constrained optimization (Beck, 2017). To bound the optimality gap, via a drift analysis of the dual update, we first establish the bounded average performance in Lemma 12 in Section 4.1. Furthermore, instead of using methods from constrained convex optimization (Mahdavi et al., 2012; Yu et al., 2017; Wei et al., 2020; Yuan and Lamperski, 2018), which either require extra assumptions or have a slow convergence rate, under strong duality, we establish that the constraint violation for the nonconvex problem (5) converges with the same rate as the optimality gap. To the best of our knowledge, this appears to be the first such result for nonconvex constrained optimization problems.

4.1 Proof of Theorem 10

We employ the performance difference lemma in Remark 8 to show the joint policy improvement per iteration in the reward and utility value functions. Neither of them is necessarily monotonic.

Lemma 11 (Non-monotonic improvement) *For any distribution of the initial state μ, the iterates $(\pi^{(t)}, \lambda^{(t)})$ of the algorithm (15) satisfy*

$$V_r^{(t+1)}(\mu) - V_r^{(t)}(\mu) + \lambda^{(t)} \left(V_g^{(t+1)}(\mu) - V_g^{(t)}(\mu) \right) \;\geq\; \frac{1-\gamma}{\eta_1} \mathbb{E}_{s\sim\mu} \left[\log Z^{(t)}(s) \right] \;\geq\; 0. \quad (16)$$

Proof Let $d_\mu^{(t+1)} := d_\mu^{\pi^{(t+1)}}$. The performance difference lemma in conjunction with the multiplicative weights update in (15b) yield

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

where the last equality follows from the definition of the Kullback-Leibler divergence or relative entropy between distributions p and q, $D_{\rm KL}(p\,\|\,q) := \mathbb{E}_{x\,\sim\,p}\log(p(x)/q(x))$. Furthermore,

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

is a consequence of the performance difference lemma (see Remark 8), where we drop a nonnegative term in (a) and (b). The first inequality in (16) follows from a component-wise inequality $d_\mu^{(t+1)} \geq (1-\gamma)\mu$, which is obtained using (9).

Now we prove that $\log Z^{(t)}(s) \geq 0$. From the definition of $Z^{(t)}(s)$, we have

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

where in (a) we apply the Jensen's inequality to the concave function $\log(x)$. On the other hand, the last equality follows from the definitions of $A_r^{(t)}(s,a)$ and $A_g^{(t)}(s,a)$, which yield

$$
\begin{aligned}
\sum_{a\,\in\,A}\pi^{(t)}(a\,|\,s)A_r^{(t)}(s,a) \;&=\; \sum_{a\,\in\,A}\pi^{(t)}(a\,|\,s)\big(Q_r^{(t)}(s,a)-V_r^{(t)}(s)\big) \;=\; 0 \\
\sum_{a\,\in\,A}\pi^{(t)}(a\,|\,s)A_g^{(t)}(s,a) \;&=\; 0,
\end{aligned}
$$

completing the proof. ■

Lemma 11 states that each primal update (15b) improves the Lagrangian-like term $V_r^{(t)}(\mu)+\lambda^{(t)}V_g^{(t)}(\mu)$ to $V_r^{(t+1)}(\mu)+\lambda^{(t)}V_g^{(t+1)}(\mu)$, with improvement depending on the previous primal-dual update $(\pi^{(t)},\lambda^{(t)})$. This lemma can be viewed as a constrained version of the policy improvement established for the unconstrained case (Agarwal et al., 2021), resulting from setting $\lambda^{(t)} = 0$. In fact, the dual iterate $\lambda^{(t)}$ captures how the constraint violation of policy improvement affects the reward value function, which is a unique feature of constrained policy optimization. Because of this superimposed effect, there is no monotonic improvement in the reward or utility value functions.

In constrained convex optimization, the primal iterate cannot reduce the unconstrained objective function, monotonically, and some averaging scheme has to be imposed (Beck, 2017). In our nonconvex context, we examine the average of value functions, which is similar to the regret analysis in online optimization. We next compare the average value functions of the policy iterates generated by the algorithm (15) with the ones that result from the use of an optimal policy.

Lemma 12 (Bounded average performance) *Let Assumption 2 hold and let us fix $T > 0$ and $\rho \in \Delta_S$. Then the iterates $\{(\pi^{(t)}, \lambda^{(t)})\}_{t=0}^{T-1}$ generated by the algorithm (15) satisfy*

$$\frac{1}{T}\sum_{t=0}^{T-1}\Big(\big(V_r^\star(\rho) - V_r^{(t)}(\rho)\big) + \lambda^{(t)}\big(V_g^\star(\rho) - V_g^{(t)}(\rho)\big)\Big) \;\le\; \frac{\log|A|}{\eta_1 T} + \frac{1}{(1-\gamma)^2 T} + \frac{2\eta_2}{(1-\gamma)^3}. \tag{17}$$

Proof Let $d^\star := d_\rho^{\pi^\star}$. The performance difference lemma in conjunction with the multiplicative weights update in (15b) yield

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

Application of the definition of the Kullback–Leibler divergence or relative entropy between distributions p and q, $D_{\rm KL}(p\,\|\,q) := \mathbb{E}_{x\sim p}\log(p(x)/q(x))$, and the performance difference lemma again yields

$$\begin{aligned}
V_r^\star(\rho) \,-\, V_r^{(t)}(\rho) &= \frac{1}{\eta_1}\,\mathbb{E}_{s\sim d^\star}\Big[D_{\rm KL}\Big(\pi^\star(\cdot\,|\,s)\,\|\,\pi^{(t)}(\cdot\,|\,s)\Big) - D_{\rm KL}\Big(\pi^\star(\cdot\,|\,s)\,\|\,\pi^{(t+1)}(\cdot\,|\,s)\Big)\Big] \\
&\quad + \frac{1}{\eta_1}\,\mathbb{E}_{s\sim d^\star}\Big[\log Z^{(t)}(s)\Big] \,-\, \frac{\lambda^{(t)}}{1-\gamma}\,\mathbb{E}_{s\sim d^\star}\left[\sum_{a\in A}\pi^\star(a\,|\,s)A_g^{(t)}(s,a)\right] \\
&= \frac{1}{\eta_1}\,\mathbb{E}_{s\sim d^\star}\Big[D_{\rm KL}\Big(\pi^\star(\cdot\,|\,s)\,\|\,\pi^{(t)}(\cdot\,|\,s)\Big) - D_{\rm KL}\Big(\pi^\star(\cdot\,|\,s)\,\|\,\pi^{(t+1)}(\cdot\,|\,s)\Big)\Big] \\
&\quad + \frac{1}{\eta_1}\,\mathbb{E}_{s\sim d^\star}\Big[\log Z^{(t)}(s)\Big] \,-\, \lambda^{(t)}\big(V_g^\star(\rho) \,-\, V_g^{(t)}(\rho)\big).
\end{aligned} \tag{18}$$

On the other hand, the first inequality in (16) with $\mu = d^\star$ becomes

$$V_r^{(t+1)}(d^\star) \,-\, V_r^{(t)}(d^\star) \,+\, \lambda^{(t)}\big(V_g^{(t+1)}(d^\star) \,-\, V_g^{(t)}(d^\star)\big) \;\ge\; \frac{1-\gamma}{\eta_1}\,\mathbb{E}_{s\sim d^\star}\Big[\log Z^{(t)}(s)\Big]. \tag{19}$$

Hence, application of (19) to the average of (18) over $t = 0, 1, \ldots, T-1$ leads to

$$
\begin{aligned}
& \frac{1}{T} \sum_{t=0}^{T-1} \left(V_r^\star(\rho) \,-\, V_r^{(t)}(\rho) \right) \\
= \;& \frac{1}{\eta_1 T} \sum_{t=0}^{T-1} \mathbb{E}_{s \sim d^\star} \left[D_{\mathrm{KL}} \left(\pi^\star(\cdot \,|\, s) \,\|\, \pi^{(t)}(\cdot \,|\, s) \right) - D_{\mathrm{KL}} \left(\pi^\star(\cdot \,|\, s) \,\|\, \pi^{(t+1)}(\cdot \,|\, s) \right) \right] \\
& + \frac{1}{\eta_1 T} \sum_{t=0}^{T-1} \mathbb{E}_{s \sim d^\star} \left[\log Z^{(t)}(s) \right] \,-\, \frac{1}{T} \sum_{t=0}^{T-1} \lambda^{(t)} \left(V_g^\star(\rho) \,-\, V_g^{(t)}(\rho) \right) \\
\leq \;& \frac{1}{\eta_1 T} \sum_{t=0}^{T-1} \mathbb{E}_{s \sim d^\star} \left[D_{\mathrm{KL}} \left(\pi^\star(\cdot \,|\, s) \,\|\, \pi^{(t)}(\cdot \,|\, s) \right) - D_{\mathrm{KL}} \left(\pi^\star(\cdot \,|\, s) \,\|\, \pi^{(t+1)}(\cdot \,|\, s) \right) \right] \\
& + \frac{1}{(1-\gamma)T} \sum_{t=0}^{T-1} \left(V_r^{(t+1)}(d^\star) \,-\, V_r^{(t)}(d^\star) \right) \\
& + \frac{1}{(1-\gamma)T} \sum_{t=0}^{T-1} \lambda^{(t)} \left(V_g^{(t+1)}(d^\star) \,-\, V_g^{(t)}(d^\star) \right) \,-\, \frac{1}{T} \sum_{t=0}^{T-1} \lambda^{(t)} \left(V_g^\star(\rho) \,-\, V_g^{(t)}(\rho) \right).
\end{aligned} \tag{20}
$$

From the dual update in (15a), we have

$$
\begin{aligned}
& \frac{1}{T} \sum_{t=0}^{T-1} \lambda^{(t)} \left(V_g^{(t+1)}(\mu) \,-\, V_g^{(t)}(\mu) \right) \\
= \;& \frac{1}{T} \sum_{t=0}^{T-1} \left(\lambda^{(t+1)} V_g^{(t+1)}(\mu) \,-\, \lambda^{(t)} V_g^{(t)}(\mu) \right) \,+\, \frac{1}{T} \sum_{t=0}^{T-1} \left(\lambda^{(t)} \,-\, \lambda^{(t+1)} \right) V_g^{(t+1)}(\mu) \\
\overset{(a)}{\leq} \;& \frac{1}{T} \lambda^{(T)} V_g^{(T)}(\mu) \,+\, \frac{1}{T} \sum_{t=0}^{T-1} \left| \lambda^{(t)} \,-\, \lambda^{(t+1)} \right| V_g^{(t+1)}(\mu) \\
\overset{(b)}{\leq} \;& \frac{2\eta_2}{(1-\gamma)^2}
\end{aligned} \tag{21}
$$

where we take a telescoping sum for the first sum in (a) and drop a non-positive term, and in (b) we utilize $|\lambda^{(T)}| \leq \eta_2 T/(1-\gamma)$ and $|\lambda^{(t)} - \lambda^{(t+1)}| \leq \eta_2/(1-\gamma)$, which follows from the dual update in (15a), the non-expansiveness of the projection $\mathcal{P}_\Lambda$, and the boundedness of the value function $V_g^{(t)}(\mu) \leq 1/(1-\gamma)$. Application of (21) with $\mu = d^\star$ and the use of telescoping sum to (20) yield

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

Finally, we use $D_{\mathrm{KL}}(p \,\|\, q) \leq \log|A|$ for $p \in \Delta_A$ and $q = \mathrm{Unif}_A$, $V_r^{(T)}(d^\star) \leq 1/(1-\gamma)$, and $V_g^\star(\rho) \geq b$ to complete the proof. ■

Lemma 12 shows that the average difference between $(V_r^\star(\rho), V_g^\star(\rho))$ and $(V_r^{(t)}(\rho), V_g^{(t)}(\rho))$ can be bounded by a (T, η_1, η_2)-term. As aforementioned, when there is no constraint (e.g., $\eta_2 = 0$), it is straightforward to strengthen Lemma 12 as the fast rate result in the unconstrained case (Agarwal et al., 2021, Theorem 16). We also note that this average performance analysis generalizes to the function approximation setting in Section 5, with an additional characterization of function approximation errors.

Proof [Proof of Theorem 10]

Bounding the optimality gap. From the dual update in (15a), we have

$$
\begin{array}{rcl}
0 \;\leq\; \big(\lambda^{(T)}\big)^2 & = & \displaystyle\sum_{t=0}^{T-1} \big((\lambda^{(t+1)})^2 \,-\, (\lambda^{(t)})^2 \big) \\
& = & \displaystyle\sum_{t=0}^{T-1} \Big(\big(\mathcal{P}_\Lambda\,(\,\lambda^{(t)} - \eta_2\,(V_g^{(t)}(\rho) - b\,)\,)\big)^2 \,-\, (\lambda^{(t)})^2 \Big) \\
& \overset{(a)}{\leq} & \displaystyle\sum_{t=0}^{T-1} \Big(\big(\,\lambda^{(t)} \,-\, \eta_2\,(V_g^{(t)}(\rho) - b\,)\,\big)^2 \,-\, (\lambda^{(t)})^2 \Big) \\
& = & 2\eta_2 \displaystyle\sum_{t=0}^{T-1} \lambda^{(t)}\big(\,b \,-\, V_g^{(t)}(\rho)\,\big) \,+\, \eta_2^2 \sum_{t=0}^{T-1} \big(\,V_g^{(t)}(\rho) \,-\, b\,\big)^2 \\
& \overset{(b)}{\leq} & 2\eta_2 \displaystyle\sum_{t=0}^{T-1} \lambda^{(t)}\big(\,V_g^\star(\rho) \,-\, V_g^{(t)}(\rho)\,\big) \,+\, \frac{\eta_2^2\, T}{(1-\gamma)^2}
\end{array}
\tag{22a}
$$

where (a) holds because of the projection $\mathcal{P}_\Lambda$, (b) is because of the feasibility of an optimal policy $\pi^\star$: $V_g^\star(\rho) \geq b$, and $|V_g^{(t)}(\rho) - b| \leq 1/(1-\gamma)$. Hence,

$$
-\frac{1}{T} \sum_{t=0}^{T-1} \lambda^{(t)}\big(\,V_g^\star(\rho) \,-\, V_g^{(t)}(\rho)\,\big) \;\leq\; \frac{\eta_2}{2(1-\gamma)^2}. \tag{22b}
$$

To obtain the optimality gap bound, we now substitute (22b) into (17), apply $D_{\rm KL}\,(p\,\|\,q) \leq \log|A|$ for $p \in \Delta_A$ and $q = {\rm Unif}_A$, and take $\eta_1 = 2\log|A|$ and $\eta_2 = 2(1-\gamma)/\sqrt{T}$.

Bounding the constraint violation. For any $\lambda \in \big[\,0,\, 2/((1-\gamma)\xi)\,\big]$, from the dual update in (15a), we have

$$
\begin{array}{rcl}
|\lambda^{(t+1)} \,-\, \lambda|^2 & \overset{(a)}{\leq} & \big|\lambda^{(t)} \,-\, \eta_2\,\big(\,V_g^{(t)}(\rho) \,-\, b\,\big) \,-\, \lambda\big|^2 \\
& = & \big|\lambda^{(t)} - \lambda\big|^2 \,-\, 2\eta_2\big(\,V_g^{(t)}(\rho) \,-\, b\,\big)\big(\,\lambda^{(t)} \,-\, \lambda\,\big) \,+\, \eta_2^2\,\big(\,V_g^{(t)}(\rho) \,-\, b\,\big)^2 \\
& \overset{(b)}{\leq} & \big|\lambda^{(t)} \,-\, \lambda\big|^2 \,-\, 2\eta_2\big(\,V_g^{(t)}(\rho) \,-\, b\,\big)\big(\,\lambda^{(t)} \,-\, \lambda\,\big) \,+\, \dfrac{\eta_2^2}{(1-\gamma)^2}
\end{array}
$$

where (a) is because of the non-expansiveness of projection $\mathcal{P}_\Lambda$ and (b) is because of $(V_g^{(t)}(\rho) - b)^2 \leq 1/(1-\gamma)^2$. Averaging the above inequality over $t = 0, \ldots, T-1$ yields

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

which implies

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(V_g^{(t)}(\rho)\,-\,b\right)\left(\lambda^{(t)}\,-\,\lambda\right)\;\leq\;\frac{1}{2\eta_2 T}\left|\lambda^{(0)}-\lambda\right|^2\,+\,\frac{\eta_2}{2(1-\gamma)^2}. \tag{23}$$

We now add (23) to (17) on both sides of the inequality, and utilize $V_g^\star(\rho)\geq b$ to obtain

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho)\,-\,V_r^{(t)}(\rho)\right)\,+\,\frac{\lambda}{T}\sum_{t=0}^{T-1}\left(b\,-\,V_g^{(t)}(\rho)\right)\\ &\leq\;\frac{\log|A|}{\eta_1 T}\,+\,\frac{1}{(1-\gamma)^2T}\,+\,\frac{2\eta_2}{(1-\gamma)^3}\,+\,\frac{1}{2\eta_2 T}\left|\lambda^{(0)}\,-\,\lambda\right|^2\,+\,\frac{\eta_2}{2(1-\gamma)^2}.\end{aligned} \tag{24}$$

Taking $\lambda=2/((1-\gamma)\xi)$ when $\sum_{t=0}^{T-1}\left(b-V_g^{(t)}(\rho)\right)\geq 0$ and $\lambda=0$ otherwise, we obtain

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

Note that both $V_r^{(t)}(\rho)$ and $V_g^{(t)}(\rho)$ can be expressed as linear functions in the same occupancy measure (Altman, 1999, Chapter 10) that is induced by the policy $\pi^{(t)}$ and the transition $P(s'\,|\,s,a)$. The convexity of the set of occupancy measures shows that the average of T occupancy measures is an occupancy measure that produces a policy π' with values $V_r^{\pi'}$ and $V_g^{\pi'}$. Hence, there exists a policy π' such that $V_r^{\pi'}(\rho)=\frac{1}{T}\sum_{t=0}^{T-1}V_r^{(t)}(\rho)$ and $V_g^{\pi'}(\rho)=\frac{1}{T}\sum_{t=0}^{T-1}V_g^{(t)}(\rho)$. Thus,

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

From Lemma 3 (ii), we have $\lambda^\star\leq 1/((1-\gamma)\xi)$. Application of Lemma 5 with $C=2/((1-\gamma)\xi)$ yields

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

which leads to our constraint violation bound if we further utilize $\frac{1}{T}\sum_{t=0}^{T-1}\left(b-V_g^{(t)}(\rho)\right)=b-V_g^{\pi'}(\rho)$, $\eta_1=2\log|A|$, and $\eta_2=2(1-\gamma)/\sqrt{T}$. ■

4.2 Zero constraint violation

In practice, it is natural to employ a conservative constraint $V_g^\pi(\rho)\geq b+\delta$ for some $\delta>0$ in Problem (1). When our desired accuracy ϵ is small enough, there exists some δ for the algorithm (15) to get zero constraint violation on average.

Corollary 13 (Zero constraint violation: softmax policy parametrization) *Let Assumption 2 hold for $\xi > 0$ and let us fix $\rho \in \Delta_S$ and replace the constraint of Problem (1) by $V_g^\pi(\rho) \geq \bar{b}$, where $\bar{b} := b + \delta$ for some $\delta > 0$. For $\epsilon < \xi/2$, there exists a $\delta = \Theta(\epsilon)$ such that if we choose $T = \Omega(1/\epsilon^2)$, $\eta_1 = 2\log|A|$, and $\eta_2 = 2(1-\gamma)/\sqrt{T}$, then the iterates $\{\pi^{(t)}\}_{t=0}^{T-1}$ generated by the algorithm (15) satisfy*

$$\text{(Optimality gap)} \quad \frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) \,-\, V_r^{(t)}(\rho)\right) \;=\; O(\epsilon)$$

$$\text{(Constraint violation)} \quad \left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b \,-\, V_g^{(t)}(\rho)\right)\right]_+ \;\leq\; 0.$$

Proof The proof idea is similar to the one used in the proof of Theorem 10. Using the new constraint $V_g^\pi(\rho) \geq \bar{b}$, Problem (1) satisfies Assumption 2 for $\bar{\xi} := \xi - \delta$ where $\delta < \xi$, and there exists an optimal policy $\bar{\pi}^\star$. Without loss of generality, by restricting $\delta < \xi/2$, we can replace Λ by $\bar{\Lambda} := [\,0, 4/((1-\gamma)\xi)\,]$, which contains $[\,0, 2/((1-\gamma)\bar{\xi})\,]$ for any such a $\bar{\xi}$. Thus, we can apply the NPG-PD algorithm (14) to this conservative problem using the projection set $\bar{\Lambda}$. It is straightforward to check that Lemma 12 holds for $V_r^{\bar{\pi}^\star}(\rho)$ and $V_g^{\bar{\pi}^\star}(\rho)$. Thus, bounding of the optimality gap in the proof of Theorem 10 proves that after $T = \Omega(1/\epsilon^2)$ iterations,

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^{\bar{\pi}^\star}(\rho) \,-\, V_r^{(t)}(\rho)\right) \;=\; O(\epsilon). \tag{25}$$

Let $q^\star$ and $\bar{q}^\star$ be the occupancy measures induced by the policies $\pi^\star$ and $\bar{\pi}^\star$, respectively. In the occupancy measure space, Problem (1) becomes a linear program and, thus, $V_r^{\pi^\star}(\rho) = \langle r, q^\star\rangle$ and $V_r^{\bar{\pi}^\star}(\rho) = \langle r, \bar{q}^\star\rangle$. By the continuity of optimal objective function in convex optimization (Terazono and Matani, 2015), $|V_r^{\pi^\star}(\rho) - V_r^{\bar{\pi}^\star}(\rho)| \leq 2\epsilon/((1-\gamma)\xi)$ for $\delta = \epsilon$. Therefore, we can replace $V_r^{\bar{\pi}^\star}(\rho)$ in (25) by $V_r^\star(\rho)$ to bound the optimality gap by the same desired accuracy ϵ up to some problem-dependent constant.

To establish the bound on constraint violation, the key change begins with (24). Since we use $\bar{b} = b+\delta$ and $V_r^{\bar{\pi}^\star}(\rho)$, the right-hand side of (24) contains an extra term $2\epsilon/((1-\gamma)\xi) - \lambda\delta$. Similarly, there are two options for selecting λ: $\lambda = 4/((1-\gamma)\xi)$ when $\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right) \geq 0$ and $\lambda = 0$ otherwise. In the first case, if we set $\delta = \epsilon$, then the extra term $-2\epsilon/((1-\gamma)\xi)$ cancels the error $O(1/\sqrt{T})$ for $T = \Omega(1/\epsilon^2)$, concluding zero constraint violation according to Lemma 5. On the other hand, the second case is exactly the zero constraint violation. ■

5 Function approximation: convergence rate and optimality

Let us consider a general form of the NPG-PD algorithm (14):

$$\begin{aligned}\theta^{(t+1)} &= \theta^{(t)} + \frac{\eta_1}{1-\gamma}\, w^{(t)} \\ \lambda^{(t+1)} &= \mathcal{P}_\Lambda\left(\lambda^{(t)} \,-\, \eta_2\left(V_g^{(t)}(\rho) \,-\, b\right)\right)\end{aligned} \tag{26}$$

where $w^{(t)}/(1-\gamma)$ denotes either the exact natural policy gradient or its sample-based approximation. For a general policy class, $\{\pi_\theta \,|\, \theta \in \Theta\}$, with the parameter space $\Theta \subset \mathbb{R}^d$, the strong duality in Lemma 3 does not necessarily hold and our analysis of Section 4 does not apply directly. Let the parametric dual objective function $V_D^{\lambda_\theta}(\rho) := \text{maximize}_{\theta\,\in\,\Theta}\, V_L^{\pi_\theta,\lambda}(\rho)$ be minimized at the optimal dual variable $\lambda_\theta^\star$. Under the Slater condition of Assumption 2, the parametrization gap (Paternain et al., 2019, Theorem 2) is determined by

$$V_r^{\pi^\star}(\rho) \;=\; V_D^{\lambda^\star}(\rho) \;\geq\; V_D^{\lambda_\theta^\star}(\rho) \;\geq\; V_r^{\pi^\star}(\rho) \,-\, M\epsilon_\pi$$

where $\epsilon_\pi := \max_s \|\pi(\cdot\,|\,s) - \pi_\theta(\cdot\,|\,s)\|_1$ is the policy approximation error and $M > 0$ is a problem-dependent constant. Application of the item (ii) in Lemma 3 to the set of all optimal dual variables $\lambda_\theta^\star$ yields $\lambda_\theta^\star \in [\,0, 2/((1-\gamma)\xi)\,]$ and, thus, $\Lambda = [\,0, 2/((1-\gamma)\xi)\,]$.

To quantify the error caused by the restricted policy parametrization, let us first generalize NPG. For a distribution over state-action pair $\nu \in \Delta_{S\times A}$, we introduce the *compatible function approximation error* (Kakade, 2002) as the following regression objective:

$$E^\nu(w;\theta,\lambda) \;:=\; \mathbb{E}_{(s,a)\,\sim\,\nu}\left[\left(A_L^{\theta,\lambda}(s,a) - w^\top \nabla_\theta \log \pi_\theta(a\,|\,s)\right)^2\right]$$

where $A_L^{\theta,\lambda}(s,a) \;:=\; A_r^\theta(s,a) + \lambda A_g^\theta(s,a)$. We can view NPG in (14) as a minimizer of $E^\nu(w;\theta,\lambda)$ for $\nu(s,a) = d_\rho^{\pi_\theta}(s)\pi_\theta(a\,|\,s)$:

$$(1-\gamma)F_\rho^\dagger(\theta)\nabla_\theta V_L^{\theta,\lambda}(\rho) \;\in\; \underset{w}{\text{argmin}}\; E^\nu(w;\theta,\lambda). \tag{27}$$

Expression (27) follows from the first-order optimality condition and the use of $\nabla_\theta V_L^{\theta,\lambda}(\rho) := \nabla_\theta V_r^\theta(\rho) + \lambda \nabla_\theta V_g^\theta(\rho)$ allows us to rewrite (27) as a linear combination of

$$(1-\gamma)F_\rho^\dagger(\theta)\nabla_\theta V_\diamond^\theta(\rho) \;\in\; \underset{w_\diamond}{\text{argmin}}\; E_\diamond^\nu(w_\diamond;\theta) \tag{28}$$

where $\diamond$ denotes r or g, and the compatible function approximation error $E_\diamond^\nu(w_\diamond;\theta)$ reads

$$E_\diamond^\nu(w_\diamond;\theta) \;:=\; \mathbb{E}_{(s,a)\,\sim\,\nu}\left[\left(A_\diamond^\theta(s,a) - w_\diamond^\top \nabla_\theta \log \pi_\theta(a\,|\,s)\right)^2\right]. \tag{29}$$

Let the minimal error be $E_{\diamond,\star}^\nu := \text{minimize}_{w_\diamond}\, E_\diamond^\nu(w_\diamond;\theta)$.

When the compatible function approximation error is zero, the global convergence follows from Theorem 10. However, this is not the case for a general policy class because it may not include all possible policies (e.g., if we take $d \ll |S||A|$ for tabular constrained MDPs). The intuition behind *compatibility* is that any minimizer of $E_\diamond^\nu(w_\diamond;\theta)$ can be used as the NPG direction without affecting the global convergence property; also see more discussions in Kakade (2002); Sutton et al. (2000); Agarwal et al. (2021).

Since the state-action measure ν of some feasible comparison policy π is not known, we introduce an exploratory initial distribution ν_0 over state-action pairs and define a state-action visitation distribution $\nu_{\nu_0}^\pi$ of a policy π as

$$\nu_{\nu_0}^\pi(s,a) \;=\; (1-\gamma)\mathbb{E}_{(s_0,a_0)\,\sim\,\nu_0}\left[\sum_{t\,=\,0}^\infty \gamma^t P^\pi\left(s_t = s, a_t = a\,|\,s_0,a_0\right)\right]$$

where $P^{\pi}(s_t = s, a_t = a \,|\, s_0, a_0)$ is the probability of visiting a state-action pair (s, a) under policy π for an initial state-action pair (s_0, a_0). Whenever clear from context, we use $\nu^{(t)}$ to denote $\nu_{\nu_0}^{\pi^{(t)}}$ for notational convenience. When the minimizer is computed exactly, we can update $w^{(t)}$ in (26) using $w^{(t)} = w_r^{(t)} + \lambda^{(t)} w_g^{(t)}$, where $w_r^{(t)}$ and $w_g^{(t)}$ are given by

$$w_\diamond^{(t)} \;\in\; \operatorname*{argmin}_{w_\diamond} \; E_\diamond^{\nu^{(t)}}\big(w_\diamond; \theta^{(t)}\big). \tag{30}$$

Even though the exact computation of a minimizer in (30) may not be feasible, we can use a sample-based algorithm to approximately solve its empirical version. By characterizing the errors that result from the sample-based solutions and from the function approximation, we next prove the convergence of the algorithm (26) for the log-linear and the general smooth policy classes.

5.1 Log-linear policy class

We first consider the policies π_θ in the log-linear class (8), with the feature maps $\phi_{s,a} \in \mathbb{R}^d$. In this case, the gradient $\nabla_\theta \log \pi_\theta(a \,|\, s)$ becomes a shifted version of the feature $\phi_{s,a}$:

$$\nabla_\theta \log \pi_\theta(a \,|\, s) \;=\; \phi_{s,a} \,-\, \mathbb{E}_{a' \sim \pi_\theta(\cdot \,|\, s)}[\, \phi_{s,a'} \,] \;=:\; \bar{\phi}_{s,a}. \tag{31}$$

Thus, the compatible function approximation error (29) captures how well the linear function $\theta^\top \bar{\phi}_{s,a}$ approximates the advantage function $A_r^\theta(s,a)$ or $A_g^\theta(s,a)$ under the state-action distribution ν. We also introduce the compatible function approximation error with respect to the state-action value function $Q_\diamond^\theta(s,a)$:

$$\mathcal{E}_\diamond^{\nu}(w_\diamond; \theta) \;:=\; \mathbb{E}_{(s,a)\sim\nu}\left[\, \big(\, Q_\diamond^\theta(s,a) \,-\, w_\diamond^\top \phi_{s,a} \,\big)^2 \,\right].$$

When there are no compatible function approximation errors, the log-linear policy update in (26) for $w^{(t)}$ that is determined by (30) is given by $w^{(t)} = w_r^{(t)} + \lambda^{(t)} w_g^{(t)}$, $w_\diamond^{(t)} \in \operatorname{argmin}_{w_\diamond} \mathcal{E}_\diamond^{\nu^{(t)}}\big(w_\diamond; \theta^{(t)}\big)$ for $\diamond = r$ or g, where $\nu^{(t)}(s,a) = d_\rho^{(t)}(s) \pi_\theta^{(t)}(a \,|\, s)$ is an on-policy state-action visitation distribution. This is because the softmax function is invariant to any terms that are independent of the action.

Let us consider an approximate solution

$$w_\diamond^{(t)} \;\approx\; \operatorname*{argmin}_{\|w_\diamond\|_2 \,\le\, W} \; \mathcal{E}_\diamond^{\nu^{(t)}}\big(w_\diamond; \theta^{(t)}\big) \tag{32}$$

where the constraint with a norm bound $W > 0$ can be viewed as an L_2-regularization. We restrict the domain to make the approximate solution well-defined even when it is not well-posed, which is similar to imposing an L_2-regularization in practice. Let an exact minimizer be $w_{\diamond,\star}^{(t)} \in \operatorname{argmin}_{\|w_\diamond\|_2 \le W} \mathcal{E}_\diamond^{\nu^{(t)}}(w_\diamond; \theta^{(t)})$. Fixing a state-action distribution $\nu^{(t)}$, the estimation error in $w_\diamond^{(t)}$ arises from the discrepancy between $w_\diamond^{(t)}$ and $w_{\diamond,\star}^{(t)}$, which comes from the randomness in a sample-based optimization algorithm and the mismatch between the linear function and the true state-action value function. We represent the estimation error as

$$\mathcal{E}_{\diamond,\text{est}}^{(t)} \;:=\; \mathbb{E}\left[\, \mathcal{E}_\diamond^{\nu^{(t)}}\big(w_\diamond^{(t)}; \theta^{(t)}\big) \,-\, \mathcal{E}_\diamond^{\nu^{(t)}}\big(w_{\diamond,\star}^{(t)}; \theta^{(t)}\big) \,\right]$$

where the expectation $\mathbb{E}$ is taken over the randomness of the approximate algorithm used to solve (32). The estimation error simplifies when the state-action value function is linear (Ding and Jovanović, 2022).

Note that the state-action distribution $\nu^{(t)}$ is on-policy. To characterize the effect of distribution shift on $w_{\diamond,\star}^{(t)}$, we first introduce some notation. We represent a fixed distribution over state-action pairs (s,a) by

$$\nu^\star(s,a) \;:=\; d_\rho^{\pi^\star}(s) \circ \mathrm{Unif}_A(a). \tag{33}$$

The fixed distribution $\nu^\star$ samples a state from $d_\rho^{\pi^\star}(s)$ and an action uniformly from $\mathrm{Unif}_A(a)$. We characterize the error in $w_{\diamond,\star}^{(t)}$ that arises from the distribution shift via the transfer error

$$\mathcal{E}_{\diamond,\mathrm{bias}}^{(t)} \;:=\; \mathbb{E}\left[\,\mathcal{E}_\diamond^{\nu^\star}\big(w_{\diamond,\star}^{(t)};\theta^{(t)}\big)\,\right].$$

The transfer error characterizes the expressiveness of function approximation that is affected by the feature maps $\phi_{s,a} \in \mathbb{R}^d$ and the quality of the exact minimizer $w_{\diamond,\star}^{(t)}$.

Assumption 14 (Estimation error and transfer error) *Both the estimation error and the transfer error are bounded, i.e.,* $\mathcal{E}_{\diamond,\mathrm{est}}^{(t)} \leq \epsilon_{\mathrm{est}}$ *and* $\mathcal{E}_{\diamond,\mathrm{bias}}^{(t)} \leq \epsilon_{\mathrm{bias}}$ *for all* $t \geq 0$*, where* $\diamond$ *denotes either* r *or* g*.*

We also point out that it is possible to remove the domain restriction in (32) when some regularity assumptions on the feature maps are made in the sample-based algorithm (Bach and Moulines, 2013). Let $\bar{w}_{\diamond,\star}^{(t)} \in \mathrm{argmin}_{w_\diamond} \mathcal{E}_\diamond^{\nu^{(t)}}\big(w_\diamond;\theta^{(t)}\big)$. Since the expressiveness of function approximation is captured by the transfer error, the gap between the exact minimizers $w_{\diamond,\star}^{(t)}$ and $\bar{w}_{\diamond,\star}^{(t)}$ is contained in the transfer error.

When we apply a sample-based algorithm to (32), it is standard to have $\epsilon_{\mathrm{est}} = O(1/\sqrt{K})$, where K is the number of samples; e.g., see Shalev-Shwartz and Ben-David (2014, Theorem 14.8). A special case is the exact tabular softmax policy parametrization for which $\epsilon_{\mathrm{bias}} = \epsilon_{\mathrm{est}} = 0$, since the features $\phi_{s,a} \in \mathbb{R}^d$ now reduce to indicator functions of the state/action spaces.

For any state-action distribution ν, we define $\Sigma_\nu := \mathbb{E}_{(s,a)\sim\nu}\left[\,\phi_{s,a}\phi_{s,a}^\top\,\right]$, and following (Agarwal et al., 2021, Assumption 6.2), to compare ν_0 with $\nu^\star$, we introduce the notion of *relative condition number*:

$$\kappa \;:=\; \sup_{w\,\in\,\mathbb{R}^d} \; \frac{w^\top \Sigma_{\nu^\star} w}{w^\top \Sigma_{\nu_0} w}.$$

Assumption 15 (Bounded relative condition number) *For an initial state-action distribution* ν_0 *and* $\nu^\star$ *determined by (33), the relative condition number* κ *is finite.*

With the estimation error ϵ_{est}, the transfer error ϵ_{bias}, and the relative condition number κ in place, in Theorem 16 we establish convergence guarantees for the algorithm (26) using the approximate update (32). Even though we set $\theta^{(0)} = 0$ and $\lambda^{(0)} = 0$ in the proof of Theorem 16, global convergence can be established for arbitrary initial conditions.

Theorem 16 (Convergence and optimality: log-linear policy parametrization) *Let Assumption 2 hold for $\xi > 0$ and let us fix a state distribution ρ and a state-action distribution ν_0. If the iterates $\{(\theta^{(t)}, \lambda^{(t)})\}_{t=0}^{T-1}$ generated by the algorithm (26) using (32) with $\|\phi_{s,a}\| \leq B$ and $\eta_1 = \eta_2 = 1/\sqrt{T}$ satisfy Assumptions 14 and 15, then*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right)\right] \;\leq\; \frac{C_3}{(1-\gamma)^4}\frac{1}{\sqrt{T}} + \frac{2+4/\xi}{(1-\gamma)^2}\left(\sqrt{|A|\,\epsilon_{\text{bias}}} + \sqrt{\frac{\kappa\,|A|\,\epsilon_{\text{est}}}{1-\gamma}}\right)$$

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right)\right]_+ \;\leq\; \frac{C_4}{(1-\gamma)^3}\frac{1}{\sqrt{T}} + \left(\frac{4+2\xi}{1-\gamma}\right)\left(\sqrt{|A|\,\epsilon_{\text{bias}}} + \sqrt{\frac{\kappa\,|A|\,\epsilon_{\text{est}}}{1-\gamma}}\right)$$

where $C_3 := 1 + \log|A| + 5B^2W^2/\xi^2$ and $C_4 := (1 + \log|A| + B^2W^2)\xi + (2 + 4B^2W^2)/\xi$.

Theorem 16 shows that, on average, the reward value function converges to its globally optimal value and that the constraint violation decays to zero (up to an estimation error ϵ_{est} and a transfer error ϵ_{bias}). When $\epsilon_{\text{bias}} = \epsilon_{\text{est}} = 0$, the rate $(1/\sqrt{T}, 1/\sqrt{T})$ matches the result in Theorem 10 for the exact tabular softmax case. Compared to (Ding et al., 2020, Theorem 2), the improved rate $1/\sqrt{T}$ in the constraint violation benefits from a new regret-type primal-dual analysis in Section 5.2, which departs from the previous drift analysis of constraint violation. In contrast to the optimality gap, the lower order of effective horizon $1/(1-\gamma)$ in the constraint violation yields a tighter error bound.

Remark 17 *By a natural error decomposition (as also used in Agarwal et al. (2021))*

$$\mathcal{E}_\diamond^{\nu^{(t)}}\left(w_\diamond^{(t)};\theta^{(t)}\right) \;=\; \mathcal{E}_\diamond^{\nu^{(t)}}\left(w_\diamond^{(t)};\theta^{(t)}\right) - \mathcal{E}_\diamond^{\nu^{(t)}}\left(w_{\diamond,\star}^{(t)};\theta^{(t)}\right) + \mathcal{E}_\diamond^{\nu^{(t)}}\left(w_{\diamond,\star}^{(t)};\theta^{(t)}\right),$$

the difference term is the standard estimation error that results from the discrepancy between $w_\diamond^{(t)}$ and $w_{\diamond,\star}^{(t)}$, and the last term characterizes the approximation error in $w_{\diamond,\star}^{(t)}$. In Corollary 18, we repeat Theorem 16 in terms of an upper bound ϵ_{approx} on the approximation error

$$\mathcal{E}_{\diamond,\text{approx}}^{(t)} \;:=\; \mathbb{E}\left[\mathcal{E}_\diamond^{\nu^{(t)}}\left(w_{\diamond,\star}^{(t)};\theta^{(t)}\right)\right].$$

Since $\mathcal{E}_{\diamond,\text{approx}}^{(t)}$ utilizes an on-policy state-action distribution $\nu^{(t)}$, the error bounds in Corollary 18 depend on the worst-case distribution mismatch coeffeceint $\|\nu^\star/\nu_0\|_\infty$. In contrast, application of estimation and transfer errors in Theorem 16 does not involve the distribution mismatch coefficient. Therefore, the error bounds in Theorem 16 are tighter than the ones in Corollary 18 that utilizes this natural error decomposition.

Corollary 18 (Convergence and optimality: log-linear policy parametrization) *Let Assumption 2 hold for $\xi > 0$ and let us fix a state distribution ρ and a state-action distribution ν_0. If the iterates $\{(\theta^{(t)}, \lambda^{(t)})\}_{t=0}^{T-1}$ generated by the algorithm (26) using (32) with $\|\phi_{s,a}\| \leq B$ and $\eta_1 = \eta_2 = 1/\sqrt{T}$ satisfy Assumption 14 except for $\mathcal{E}_{\diamond,\text{bias}}^{(t)}$, Assumption 15,* *and additionally* $\mathcal{E}_{\diamond,\text{approx}}^{(t)} \leq \epsilon_{\text{approx}}$ *(for both* $\diamond = r$ *and* g*), then*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right)\right] \leq \frac{C_3}{(1-\gamma)^4}\frac{1}{\sqrt{T}} + C_3'\left(\sqrt{\frac{|A|\,\epsilon_{\text{approx}}}{1-\gamma}\left\|\frac{\nu^\star}{\nu_0}\right\|_\infty} + \sqrt{\frac{\kappa\,|A|\,\epsilon_{\text{est}}}{1-\gamma}}\right)$$

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right)\right]_+ \leq \frac{C_4}{(1-\gamma)^3}\frac{1}{\sqrt{T}} + C_4'\left(\sqrt{\frac{|A|\,\epsilon_{\text{approx}}}{1-\gamma}\left\|\frac{\nu^\star}{\nu_0}\right\|_\infty} + \sqrt{\frac{\kappa\,|A|\,\epsilon_{\text{est}}}{1-\gamma}}\right)$$

where $C_3 := 1 + \log|A| + 5B^2W^2/\xi^2$*,* $C_4 := (1+\log|A| + B^2W^2)\xi + (2+4B^2W^2)/\xi$*,* $C_3' := (2+4/\xi)/(1-\gamma)^2$*, and* $C_4' := (4+2\xi)/(1-\gamma)$*.*

Proof From the definitions of $\mathcal{E}_\diamond^{\nu^\star}$ and $\mathcal{E}_\diamond^{\nu^{(t)}}$, we have

$$\mathcal{E}_\diamond^{\nu^\star}\left(w_{\diamond,\star}^{(t)};\theta^{(t)}\right) \leq \left\|\frac{\nu^\star}{\nu^{(t)}}\right\|_\infty \mathcal{E}_\diamond^{\nu^{(t)}}\left(w_{\diamond,\star}^{(t)};\theta^{(t)}\right) \leq \frac{1}{1-\gamma}\left\|\frac{\nu^\star}{\nu_0}\right\|_\infty \mathcal{E}_\diamond^{\nu^{(t)}}\left(w_{\diamond,\star}^{(t)};\theta^{(t)}\right)$$

where the second inequality is because of $(1-\gamma)\nu_0 \leq \nu^{(t)}$. Thus,

$$\mathcal{E}_{\diamond,\text{bias}}^{(t)} \leq \frac{1}{1-\gamma}\left\|\frac{\nu^\star}{\nu_0}\right\|_\infty \mathcal{E}_{\diamond,\text{approx}}^{(t)}$$

which allows us to replace $\mathcal{E}_{\diamond,\text{bias}}^{(t)}$ in the proof of Theorem 16 by $\mathcal{E}_{\diamond,\text{approx}}^{(t)}$. ■

5.2 Proof of Theorem 16

We provide a regret-type analysis for a general class of smooth policies that subsumes the log-linear policy class as a special case, in Lemma 19. Using the property of policy smoothness, we first generalize Lemma 12 to the function approximation setting. Then, we can utilize the function approximation error to contain the duality gap and characterize the regret and the constraint violation performance.

Lemma 19 (Regret/Violation lemma) *Let Assumption 2 hold for* $\xi > 0$*, let us fix a state distribution* ρ *and* $T > 0$*, and let* $\log\pi_\theta(a\,|\,s)$ *be* β*-smooth in* θ *for any* (s,a)*. If the iterates* $\{(\theta^{(t)},\lambda^{(t)})\}_{t=0}^{T-1}$ *are generated by the algorithm (26) with* $\theta^{(0)} = 0$*,* $\lambda^{(0)} = 0$*,* $\eta_1 = \eta_2 = 1/\sqrt{T}$*, and* $\|w_\diamond^{(t)}\| \leq W$*, then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right) \leq \frac{C_3}{(1-\gamma)^4}\frac{1}{\sqrt{T}} + \sum_{t=0}^{T-1}\frac{\text{err}_r^{(t)}(\pi^\star)}{(1-\gamma)T} + \sum_{t=0}^{T-1}\frac{2\times\text{err}_g^{(t)}(\pi^\star)}{(1-\gamma)^2\xi T}$$

$$\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right)\right]_+ \leq \frac{C_4}{(1-\gamma)^3}\frac{1}{\sqrt{T}} + \sum_{t=0}^{T-1}\frac{\xi\times\text{err}_r^{(t)}(\pi^\star)}{T} + \sum_{t=0}^{T-1}\frac{2\times\text{err}_g^{(t)}(\pi^\star)}{(1-\gamma)T}$$

where $C_3 := 1+\log|A| + 5\beta W^2/\xi^2$*,* $C_4 := (1+\log|A|+\beta W^2)\xi + (2+4\beta W^2)/\xi$*, and*

$$\text{err}_\diamond^{(t)}(\pi) := \left|\mathbb{E}_{s\sim d_\rho^\pi}\mathbb{E}_{a\sim\pi(\cdot\,|\,s)}\left[A_\diamond^{(t)}(s,a) - (w_\diamond^{(t)})^\top\nabla_\theta\log\pi_\theta^{(t)}(a\,|\,s)\right]\right|$$

where $\diamond = r$ *or* g*.*

Proof The smoothness of the log-linear policy in conjunction with an application of Taylor expansion to $\log \pi_\theta^{(t)}(a\,|\,s)$ yields

$$\log \frac{\pi_\theta^{(t)}(a\,|\,s)}{\pi_\theta^{(t+1)}(a\,|\,s)} \,+\, \left(\theta^{(t+1)} - \theta^{(t)}\right)^\top \nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s) \;\leq\; \frac{\beta}{2}\left\|\theta^{(t+1)} - \theta^{(t)}\right\|^2 \tag{34}$$

where $\theta^{(t+1)} - \theta^{(t)} = \eta_1 w^{(t)}/(1-\gamma)$. Fixing π and ρ, we use d to denote d_ρ^π to obtain

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

where (a) is because of (34), we use the update $w^{(t)} = w_r^{(t)} + \lambda^{(t)}w_g^{(t)}$ for a given $\lambda^{(t)}$ in (b), (c) is due to $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$, and in (d) we apply the performance difference lemma (see Remark 8), the definitions of $\mathrm{err}_r^{(t)}(\pi)$ and $\mathrm{err}_g^{(t)}(\pi)$, and $\|w_\diamond^{(t)}\| \leq W$. Rearrangement of the above inequality yields

$$
\begin{aligned}
& V_r^{\pi}(\rho) \,-\, V_r^{(t)}(\rho) \\
\leq\;& \frac{1}{\eta_1}\mathbb{E}_{s\sim d}\left[D_{\rm KL}\left(\pi(\cdot\,|\,s)\,\|\,\pi_\theta^{(t)}(\cdot\,|\,s)\right) \,-\, D_{\rm KL}\left(\pi(\cdot\,|\,s)\,\|\,\pi_\theta^{(t+1)}(\cdot\,|\,s)\right)\right] \\
& +\, \frac{1}{1-\gamma}\,{\rm err}_r^{(t)}(\pi) \,+\, \frac{2}{(1-\gamma)^2\xi}\,{\rm err}_g^{(t)}(\pi) \,+\, \beta\frac{\eta_1 W^2}{(1-\gamma)^2} \,+\, \beta\frac{4\eta_1 W^2}{(1-\gamma)^4\xi^2} \\
& -\, \lambda^{(t)}\big(\, V_g^{\pi}(\rho) - V_g^{(t)}(\rho)\,\big)
\end{aligned}
$$

where we utilize $0 \leq \lambda^{(t)} \leq 2/((1-\gamma)\xi)$ from the dual update in (26).

Averaging the inequality above over $t = 0, 1, \ldots, T-1$ yields

$$
\begin{aligned}
& \frac{1}{T}\sum_{t\,=\,0}^{T-1}\big(\, V_r^{\pi}(\rho) \,-\, V_r^{(t)}(\rho)\,\big) \\
\leq\;& \frac{1}{\eta_1 T}\sum_{t\,=\,0}^{T-1}\mathbb{E}_{s\sim d}\left[D_{\rm KL}\big(\pi(\cdot\,|\,s)\,\|\,\pi_\theta^{(t)}(\cdot\,|\,s)\big) \,-\, D_{\rm KL}\left(\pi(\cdot\,|\,s)\,\|\,\pi_\theta^{(t+1)}(\cdot\,|\,s)\right)\right] \\
& +\, \frac{1}{(1-\gamma)T}\sum_{t\,=\,0}^{T-1}{\rm err}_r^{(t)}(\pi) \,+\, \frac{2}{(1-\gamma)^2\xi T}\sum_{t\,=\,0}^{T-1}{\rm err}_g^{(t)}(\pi) \,+\, \beta\frac{\eta_1 W^2}{(1-\gamma)^2} \,+\, \beta\frac{4\eta_1 W^2}{(1-\gamma)^4\xi^2} \\
& -\, \frac{1}{T}\sum_{t\,=\,0}^{T-1}\lambda^{(t)}\big(\, V_g^{\pi}(\rho) - V_g^{(t)}(\rho)\,\big)
\end{aligned}
$$

which implies

$$
\begin{aligned}
& \frac{1}{T}\sum_{t\,=\,0}^{T-1}\big(\, V_r^{\pi}(\rho) \,-\, V_r^{(t)}(\rho)\,\big) \\
\leq\;& \frac{\log|A|}{\eta_1 T} \,+\, \frac{1}{(1-\gamma)T}\sum_{t\,=\,0}^{T-1}{\rm err}_r^{(t)}(\pi) \,+\, \frac{2}{(1-\gamma)^2\xi T}\sum_{t\,=\,0}^{T-1}{\rm err}_g^{(t)}(\pi) \\
& +\, \beta\frac{\eta_1 W^2}{(1-\gamma)^2} \,+\, \beta\frac{4\eta_1 W^2}{(1-\gamma)^4\xi^2} \,-\, \frac{1}{T}\sum_{t\,=\,0}^{T-1}\lambda^{(t)}\big(V_g^{\pi}(\rho) - V_g^{(t)}(\rho)\big).
\end{aligned}
$$

If we choose the comparison policy $\pi = \pi^\star$, then we have

$$
\begin{aligned}
& \frac{1}{T}\sum_{t\,=\,0}^{T-1}\big(\, V_r^{\star}(\rho) \,-\, V_r^{(t)}(\rho)\,\big) \,+\, \frac{1}{T}\sum_{t\,=\,0}^{T-1}\lambda^{(t)}\big(\, V_g^{\star}(\rho) \,-\, V_g^{(t)}(\rho)\,\big) \\
\leq\;& \frac{\log|A|}{\eta_1 T} \,+\, \frac{1}{(1-\gamma)T}\sum_{t\,=\,0}^{T-1}{\rm err}_r^{(t)}(\pi^\star) \,+\, \frac{2}{(1-\gamma)^2\xi T}\sum_{t\,=\,0}^{T-1}{\rm err}_g^{(t)}(\pi^\star) \\
& +\, \beta\frac{\eta_1 W^2}{(1-\gamma)^2} \,+\, \beta\frac{4\eta_1 W^2}{(1-\gamma)^4\xi^2}.
\end{aligned} \tag{35}
$$

Proving the first inequality. By the same reasoning as in (22a),

$$
\begin{aligned}
0 \;\leq\; \big(\lambda^{(T)}\big)^2 \;&=\; \sum_{t=0}^{T-1} \big((\lambda^{(t+1)})^2 \,-\, (\lambda^{(t)})^2 \big) \\
&\leq\; 2\eta_2 \sum_{t=0}^{T-1} \lambda^{(t)} \big(b \,-\, V_g^{(t)}(\rho) \big) \,+\, \eta_2^2 \sum_{t=0}^{T-1} \big(V_g^{(t)}(\rho) \,-\, b \big)^2 \\
&\overset{(a)}{\leq}\; 2\eta_2 \sum_{t=0}^{T-1} \lambda^{(t)} \big(V_g^{\star}(\rho) \,-\, V_g^{(t)}(\rho) \big) \,+\, \frac{\eta_2^2 T}{(1-\gamma)^2}
\end{aligned} \tag{36a}
$$

where (a) is because of the feasibility of $\pi^\star$: $V_g^\star(\rho) \geq b$, and $|V_g^{(t)}(\rho) - b| \leq 1/(1-\gamma)$. Hence,

$$
-\frac{1}{T} \sum_{t=0}^{T-1} \lambda^{(t)} \big(V_g^{\star}(\rho) \,-\, V_g^{(t)}(\rho) \big) \;\leq\; \frac{\eta_2}{2(1-\gamma)^2}. \tag{36b}
$$

By adding the inequality (36b) to (35) on both sides and taking $\eta_1 = \eta_2 = 1/\sqrt{T}$, we obtain the first inequality.

Proving the second inequality. Since the dual update in (26) is the same as the one in (15a), we can use the same reasoning to conclude (23). Adding the inequality (23) to (35) on both sides and using $V_g^\star(\rho) \geq b$ yield

$$
\begin{aligned}
& \frac{1}{T} \sum_{t=0}^{T-1} \big(V_r^{\star}(\rho) \,-\, V_r^{(t)}(\rho) \big) \,+\, \frac{\lambda}{T} \sum_{t=0}^{T-1} \big(b \,-\, V_g^{(t)}(\rho) \big) \\
\leq\; & \frac{\log|A|}{\eta_1 T} \,+\, \frac{1}{(1-\gamma)T} \sum_{t=0}^{T-1} \mathrm{err}_r^{(t)}(\pi^\star) \,+\, \frac{2}{(1-\gamma)^2 \xi T} \sum_{t=0}^{T-1} \mathrm{err}_g^{(t)}(\pi^\star) \\
& +\, \beta \frac{\eta_1 W^2}{(1-\gamma)^2} \,+\, \beta \frac{4\eta_1 W^2}{(1-\gamma)^4 \xi^2} \,+\, \frac{1}{2\eta_2 T} \big| \lambda^{(0)} - \lambda \big|^2 \,+\, \frac{\eta_2}{2(1-\gamma)^2}.
\end{aligned} \tag{37}
$$

Taking $\lambda = \frac{2}{(1-\gamma)\xi}$ when $\sum_{t=0}^{T-1} \big(b - V_g^{(t)}(\rho) \big) \geq 0$ and $\lambda = 0$ otherwise, we obtain

$$
\begin{aligned}
& V_r^{\star}(\rho) \,-\, \frac{1}{T} \sum_{t=0}^{T-1} V_r^{(t)}(\rho) \,+\, \frac{2}{(1-\gamma)\xi} \left[b \,-\, \frac{1}{T} \sum_{t=0}^{T-1} V_g^{(t)}(\rho) \right]_+ \\
\leq\; & \frac{\log|A|}{\eta_1 T} \,+\, \frac{1}{(1-\gamma)T} \sum_{t=0}^{T-1} \mathrm{err}_r^{(t)}(\pi^\star) \,+\, \frac{2}{(1-\gamma)^2 \xi T} \sum_{t=0}^{T-1} \mathrm{err}_g^{(t)}(\pi^\star) \\
& +\, \beta \frac{\eta_1 W^2}{(1-\gamma)^2} \,+\, \beta \frac{4\eta_1 W^2}{(1-\gamma)^4 \xi^2 T} \,+\, \frac{2}{\eta_2 (1-\gamma)^2 \xi^2} \,+\, \frac{\eta_2}{2(1-\gamma)^2}.
\end{aligned}
$$

Since $V_r^{(t)}(\rho)$ and $V_g^{(t)}(\rho)$ are linear functions in the occupancy measure (Altman, 1999, Chapter 10), there exists a policy π' such that $V_r^{\pi'}(\rho) = \frac{1}{T} \sum_{t=0}^{T-1} V_r^{(t)}(\rho)$ and $V_g^{\pi'}(\rho) =$

$\frac{1}{T}\sum_{t=0}^{T-1} V_g^{(t)}(\rho)$. Hence,

$$
\begin{aligned}
& V_r^\star(\rho) \,-\, V_r^{\pi'}(\rho) \,+\, \frac{2}{(1-\gamma)\xi}\left[b \,-\, V_g^{\pi'}(\rho)\right]_+ \\
\leq\; & \frac{\log|A|}{\eta_1 T} \,+\, \frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\mathrm{err}_r^{(t)}(\pi^\star) \,+\, \frac{2}{(1-\gamma)^2\xi T}\sum_{t=0}^{T-1}\mathrm{err}_g^{(t)}(\pi^\star) \\
& + \beta\frac{\eta_1 W^2}{(1-\gamma)^2} \,+\, \beta\frac{4\eta_1 W^2}{(1-\gamma)^4\xi^2} \,+\, \frac{2}{\eta_2(1-\gamma)^2\xi^2 T} \,+\, \frac{\eta_2}{2(1-\gamma)^2}.
\end{aligned}
$$

From Lemma 3 (ii), we have $\lambda^\star \leq 1/((1-\gamma)\xi)$. Application of Lemma 5 with $C = 2/((1-\gamma)\xi)$ yields

$$
\begin{aligned}
\left[b \,-\, V_g^{\pi'}(\rho)\right]_+ \;\leq\; & \frac{\xi\log|A|}{\eta_1 T} \,+\, \frac{\xi}{T}\sum_{t=0}^{T-1}\mathrm{err}_r^{(t)}(\pi^\star) \,+\, \frac{2}{(1-\gamma)T}\sum_{t=0}^{T-1}\mathrm{err}_g^{(t)}(\pi^\star) \\
& + \beta\frac{\eta_1\xi W^2}{1-\gamma} \,+\, \beta\frac{4\eta_1 W^2}{(1-\gamma)^3\xi} \,+\, \frac{2}{\eta_2(1-\gamma)\xi T} \,+\, \frac{\eta_2\xi}{2(1-\gamma)}.
\end{aligned}
$$

which leads to our constraint violation bound if we further utilize $\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right) = b - V_g^{\pi'}(\rho)$ and $\eta_1 = \eta_2 = 1/\sqrt{T}$. ■

The analysis of Lemma 19 is based on the generalization of Lemma 12 to the function approximation setting using the property of policy smoothness. A crucial step is to use the original optimal policy as our comparison policy in hindsight, instead of a sub-optimal policy within policy class (Ding et al., 2020, Theorem 2). Although the strong duality may not hold because of the insufficient expressiveness of the parametrized policy class, we can characterize the regret and constraint violation bounds, up to some function approximation errors.

Proof [Proof of Theorem 16]

When $\|\phi_{s,a}\| \leq B$, for the log-linear policy class, $\log\pi_\theta(a\,|\,s)$ is β-smooth with $\beta = B^2$. By Lemma 19, it remains to consider the randomness in the sequences of $\{w_r^{(t)}, w_g^{(t)}\}$ and the error bounds for $\mathrm{err}_r^{(t)}(\pi^\star)$ and $\mathrm{err}_g^{(t)}(\pi^\star)$. Application of the triangle inequality yields

$$
\begin{aligned}
\mathrm{err}_r^{(t)}(\pi^\star) \;\leq\; & \left|\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[A_r^{(t)}(s,a) \,-\, \left(w_{r,\star}^{(t)}\right)^\top\nabla_\theta\log\pi_\theta^{(t)}(a\,|\,s)\right]\right| \\
& + \left|\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\left(w_{r,\star}^{(t)} \,-\, w_r^{(t)}\right)^\top\nabla_\theta\log\pi_\theta^{(t)}(a\,|\,s)\right]\right|.
\end{aligned} \tag{38}
$$

Application of (31) and $A_r^{(t)}(s,a) = Q_r^{(t)}(s,a) - \mathbb{E}_{a'\sim\pi_\theta^{(t)}(\cdot\,|\,s)}\left[\,Q_r^{(t)}(s,a')\,\right]$ yields

$$
\begin{aligned}
&\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[A_r^{(t)}(s,a) \,-\, \left(w_{r,\star}^{(t)}\right)^\top \nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s)\right] \\
&= \; \mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[Q_r^{(t)}(s,a) \,-\, \phi_{s,a}^\top w_{r,\star}^{(t)}\right] \\
&\quad\; - \mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a'\sim\pi_\theta^{(t)}(\cdot\,|\,s)}\left[Q_r^{(t)}(s,a') \,-\, \phi_{s,a'}^\top w_{r,\star}^{(t)}\right] \\
&\leq \; \sqrt{\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left(Q_r^{(t)}(s,a) \,-\, \phi_{s,a}^\top w_{r,\star}^{(t)}\right)^2} \qquad\qquad (39) \\
&\quad\; + \sqrt{\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a'\sim\pi_\theta^{(t)}(\cdot\,|\,s)}\left(Q_r^{(t)}(s,a') \,-\, \phi_{s,a'}^\top w_{r,\star}^{(t)}\right)^2} \\
&\leq \; 2\sqrt{|A|\,\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim \mathrm{Unif}_A}\left[\left(Q_r^{(t)}(s,a) \,-\, \phi_{s,a}^\top w_{r,\star}^{(t)}\right)^2\right]} \\
&= \; 2\sqrt{|A|\,\mathcal{E}_r^{\nu^\star}\left(w_{r,\star}^{(t)};\theta^{(t)}\right)}.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\left(w_{r,\star}^{(t)} \,-\, w_r^{(t)}\right)^\top \nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s)\right] \\
&= \; \mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\left(w_{r,\star}^{(t)} \,-\, w_r^{(t)}\right)^\top \phi_{s,a}\right] \\
&\quad\; - \mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a'\sim\pi_\theta^{(t)}(\cdot\,|\,s)}\left[\left(w_{r,\star}^{(t)} \,-\, w_r^{(t)}\right)^\top \phi_{s,a'}\right] \qquad\qquad (40) \\
&\leq \; 2\sqrt{|A|\,\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim \mathrm{Unif}_A}\left[\left(\left(w_{r,\star}^{(t)} \,-\, w_r^{(t)}\right)^\top \phi_{s,a}\right)^2\right]} \\
&= \; 2\sqrt{|A|\left\|w_{r,\star}^{(t)} \,-\, w_r^{(t)}\right\|_{\Sigma_{\nu^\star}}^2}
\end{aligned}
$$

where $\Sigma_{\nu^\star} := \mathbb{E}_{(s,a)\sim\nu^\star}\left[\,\phi_{s,a}\phi_{s,a}^\top\,\right]$. From the definition of κ, we have

$$
\left\|w_{r,\star}^{(t)} \,-\, w_r^{(t)}\right\|_{\Sigma_{\nu^\star}}^2 \;\leq\; \kappa\left\|w_{r,\star}^{(t)} \,-\, w_r^{(t)}\right\|_{\Sigma_{\nu_0}}^2 \;\leq\; \frac{\kappa}{1-\gamma}\left\|w_{r,\star}^{(t)} \,-\, w_r^{(t)}\right\|_{\Sigma_{\nu^{(t)}}}^2 \qquad (41)
$$

where we use $(1-\gamma)\nu_0 \leq \nu_{\nu_0}^{\pi^{(t)}} := \nu^{(t)}$ in the second inequality. Evaluation of the first-order optimality condition of $w_{r,\star}^{(t)} \in \mathrm{argmin}_{\|w_r\|_2\leq W}\; \mathcal{E}_r^{\nu^{(t)}}(w_r;\theta^{(t)})$ yields

$$
\left(w_r \,-\, w_{r,\star}^{(t)}\right)^\top \nabla_w \mathcal{E}_r^{\nu^{(t)}}(w_{r,\star}^{(t)};\theta^{(t)}) \;\geq\; 0, \;\; \text{for any } w_r \text{ satisfying } \|w_r\| \;\leq\; W.
$$

Thus,

$$
\begin{aligned}
& \mathcal{E}_r^{\nu^{(t)}}(w_r;\theta^{(t)}) - \mathcal{E}_r^{\nu^{(t)}}(w_{r,\star}^{(t)};\theta^{(t)}) \\
= \;\; & \mathbb{E}_{s,a\,\sim\,\nu^{(t)}}\left[\left(Q_r^{(t)}(s,a) \,-\, \phi_{s,a}^\top w_{r,\star}^{(t)} \,+\, \phi_{s,a}^\top w_{r,\star}^{(t)} \,-\, \phi_{s,a}^\top w_r\right)^2\right] \,-\, \mathcal{E}_r^{\nu^{(t)}}(w_{r,\star}^{(t)};\theta^{(t)}) \\
= \;\; & 2\left(w_{r,\star}^{(t)} \,-\, w_r\right)^\top \mathbb{E}_{s,a\,\sim\,\nu^{(t)}}\left[\left(Q_r^{(t)}(s,a) \,-\, \phi_{s,a}^\top w_{r,\star}^{(t)}\right)\phi_{s,a}\right] \\
& + \mathbb{E}_{s,a\,\sim\,\nu^{(t)}}\left[\left(\phi_{s,a}^\top w_{r,\star}^{(t)} \,-\, \phi_{s,a}^\top w_r\right)^2\right] \\
= \;\; & \left(w_r \,-\, w_{r,\star}^{(t)}\right)^\top \nabla_w \mathcal{E}_r^{\nu^{(t)}}(w_{r,\star}^{(t)};\theta^{(t)}) \,+\, \left\|w_r \,-\, w_{r,\star}^{(t)}\right\|_{\Sigma_{\nu^{(t)}}}^2 \\
\geq \;\; & \left\|w_r \,-\, w_{r,\star}^{(t)}\right\|_{\Sigma_{\nu^{(t)}}}^2.
\end{aligned}
$$

Taking $w_r = w_r^{(t)}$ in the above inequality and combining it with (40) and (41) yield

$$
\begin{aligned}
& \mathbb{E}_{s\,\sim\, d_\rho^\star}\mathbb{E}_{a\,\sim\,\pi^\star(\cdot\,|\,s)}\left[\left(w_{r,\star}^{(t)} \,-\, w_r^{(t)}\right)^\top \nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s)\right] \\
\leq \;\; & 2\sqrt{\frac{\kappa\,|A|}{1-\gamma}\left(\mathcal{E}_r^{\nu^{(t)}}(w_r^{(t)};\theta^{(t)}) \,-\, \mathcal{E}_r^{\nu^{(t)}}(w_{r,\star}^{(t)};\theta^{(t)})\right)}.
\end{aligned}
\tag{42}
$$

Substitution of (39) and (42) into the right-hand side of (38) yields

$$
\mathbb{E}\left[\mathrm{err}_r^{(t)}(\pi^\star)\right] \;\leq\; 2\sqrt{|A|\,\mathbb{E}\left[\mathcal{E}_r^{\nu^\star}(w_{r,\star}^{(t)};\theta^{(t)})\right]} \,+\, 2\sqrt{\frac{\kappa\,|A|}{1-\gamma}\mathbb{E}\left[\mathcal{E}_r^{\nu^{(t)}}(w_r^{(t)};\theta^{(t)}) - \mathcal{E}_r^{\nu^{(t)}}(w_{r,\star}^{(t)};\theta^{(t)})\right]}.
$$

By the same reasoning, we can establish a similar bound on $\mathbb{E}\big[\mathrm{err}_g^{(t)}(\pi^\star)\big]$. Finally, our desired results follow by applying Assumption 14 and Lemma 19. ■

To obtain zero constraint violation, we apply the algorithm (26) to Problem (1) with a conservative constraint: $V_g^\pi(\rho) \geq b + \delta$ for some $\delta > 0$, , as done in Corollary 13. In addition to the parameters (ϵ, δ), the errors of function approximation (e.g., ϵ_{est} and ϵ_{bias}) are required to be small.

Corollary 20 (Zero constraint violation: log-linear policy parametrization) *Let Assumption 2 hold for $\xi > 0$ and let us fix a state distribution ρ and replace the constraint of Problem (1) by $V_g^\pi(\rho) \geq \bar{b}$, where $\bar{b} := b + \delta$ for some $\delta > 0$. For $\epsilon < \xi/2$, there exists $\delta = \Theta(\epsilon)$ such that if we choose $T = \Omega(1/\epsilon^2)$, $\eta_1 = \eta_2 = 1/\sqrt{T}$, and Assumption 14 holds for*

$\epsilon_{\text{est}} = \epsilon_{\text{bias}} = O(\epsilon^2)$, *then the iterates* $\{\theta^{(t)}\}_{t=0}^{T-1}$ *generated by the algorithm (26) satisfy*

$$\text{(Optimality gap)}\;\; \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right)\right] \;=\; O(\epsilon)$$

$$\text{(Constraint violation)}\;\; \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right)\right]_+ \;\leq\; 0.$$

Proof The proof idea is similar to the one used in the proof of Theorem 16. Using the new constraint $V_g^\pi(\rho) \geq \bar{b}$, Problem (1) satisfies Assumption 2 for $\bar{\xi} := \xi - \delta$ where $\delta < \xi$, and there exists an optimal policy $\bar{\pi}^\star$. Without loss of generality, by restricting $\delta < \xi/2$, we can replace Λ by $\bar{\Lambda} := [\,0, 4/((1-\gamma)\xi)\,]$, which contains $[\,0, 2/((1-\gamma)\bar{\xi})\,]$ for any such $\bar{\xi}$. Thus, we can apply the NPG-PD algorithm (26) to this conservative problem using the projection set $\bar{\Lambda}$. It is straightforward to check that (37) holds for $V_r^{\bar{\pi}^\star}(\rho)$ and $V_g^{\bar{\pi}^\star}(\rho)$. Thus, the second inequality proof in Lemma 19 in conjunction with $\epsilon_{\text{est}} = \epsilon_{\text{bias}} = O(\epsilon^2)$ proves that after $T = \Omega(1/\epsilon^2)$ iterations,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^{\bar{\pi}^\star}(\rho) - V_r^{(t)}(\rho)\right)\right] \;=\; O(\epsilon) \tag{43}$$

where the expectation $\mathbb{E}$ is taken over the randomness of approximate algorithm that is used to solve (32). Let $q^\star$ and $\bar{q}^\star$ be the occupancy measures induced by policies $\pi^\star$ and $\bar{\pi}^\star$, respectively. In the occupancy measure space, Problem (1) becomes a linear program, and thus, $V_r^{\pi^\star}(\rho) = \langle r, q^\star\rangle$ and $V_r^{\bar{\pi}^\star}(\rho) = \langle r, \bar{q}^\star\rangle$. By the continuity of optimal objective function in convex optimization (Terazono and Matani, 2015), $|V_r^{\pi^\star}(\rho) - V_r^{\bar{\pi}^\star}(\rho)| \leq 2\epsilon/((1-\gamma)\xi)$ for $\delta = \epsilon$. Therefore, we can replace $V_r^{\bar{\pi}^\star}(\rho)$ in (43) by $V_r^\star(\rho)$ to bound the optimality gap by the same desired accuracy ϵ up to some problem-dependent constant.

To establish the bound on the constraint violation, the key change begins with (37). Since we use $\bar{b} = b + \delta$ and $V_r^{\bar{\pi}^\star}(\rho)$, the right-hand side of (37) contains an extra term $2\epsilon/((1-\gamma)\xi) - \lambda\delta$. Similarly, there are two options for selecting λ: $\lambda = 4/((1-\gamma)\xi)$ when $\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right) \geq 0$ and $\lambda = 0$ otherwise. In the first case, if we set $\delta = \epsilon$ and $\epsilon_{\text{est}} = \epsilon_{\text{bias}} = O(\epsilon^2)$, then the extra term $-2\epsilon/((1-\gamma)\xi)$ cancels the rate $O(1/\sqrt{T})$ for $T = \Omega(1/\epsilon^2)$ and the function approximation errors ϵ_{est} and ϵ_{bias}, concluding zero constraint violation according to Lemma 5. On the other hand, the second case is exactly the zero constraint violation. ∎

Remark 21 (Zero constraint violation: log-linear policy parametrization) *As done in Corollary 20, we can refine the constraint violation in Corollary 18 to be zero. Given a small desired accuracy* $\epsilon > 0$, *there exists* $\delta = \Theta(\epsilon)$ *such that if* $T = \Omega(1/\epsilon^2)$, *and* $\epsilon_{\text{est}} = \epsilon_{\text{approx}} = O(\epsilon^2)$, *then*

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

We note that $\epsilon_{\text{approx}} = O(\epsilon^2)$ *is more difficult to achieve than* $\epsilon_{\text{bias}} = O(\epsilon^2)$ *in practice.*

5.3 General smooth policy class

For a general class of smooth policies (Zhang et al., 2020b; Agarwal et al., 2021), we now establish the convergence of algorithm (26) with an approximate gradient update:

$$
\begin{aligned}
w^{(t)} \;&=\; w_r^{(t)} \,+\, \lambda^{(t)}\, w_g^{(t)} \\
w_\diamond^{(t)} \;&\approx\; \operatorname*{argmin}_{\|w_\diamond\|_2\,\leq\, W} \; E_\diamond^{\nu^{(t)}}\big(w_\diamond;\theta^{(t)}\big)
\end{aligned}
\tag{44}
$$

where $\diamond$ denotes r or g and an exact minimizer is given by $w_{\diamond,\star}^{(t)} \in \operatorname{argmin}_{\|w_\diamond\|_2 \leq W} E_\diamond^{\nu^{(t)}}(w_\diamond;\theta^{(t)})$. We treat $w_\diamond^{(t)}$ as random since it is typically obtained from a sample-based algorithm.

Assumption 22 (Policy smoothness) *For all $s \in S$ and $a \in A$, $\log\pi_\theta(a\,|\,s)$ is a β-smooth function of θ:*

$$
\|\nabla_\theta \log\pi_\theta(a\,|\,s) \,-\, \nabla_\theta \log\pi_{\theta'}(a\,|\,s)\| \;\leq\; \beta \left\|\theta \,-\, \theta'\right\| \quad \textit{for all} \;\; \theta,\, \theta' \,\in\, \mathbb{R}^d.
$$

Since both the tabular softmax and log-linear policies are smooth (Agarwal et al., 2021, Remark 28 and Appendix D), Assumption 22 covers a broader function class relative to the softmax policy parametrization (7).

Given a state-action distribution $\nu^{(t)}$, we introduce the estimation error as

$$
E_{\diamond,\text{est}}^{(t)} \;:=\; \mathbb{E}\left[\, E_\diamond^{\nu^{(t)}}\big(w_\diamond^{(t)};\theta^{(t)}\big) \,-\, E_\diamond^{\nu^{(t)}}\big(w_{\diamond,\star}^{(t)};\theta^{(t)}\big) \,\big|\, \theta^{(t)} \,\right].
$$

Furthermore, given a state distribution ρ and an optimal policy $\pi^\star$, we define a state-action distribution $\nu^\star(s,a) := d_\rho^{\pi^\star}(s)\pi^\star(a\,|\,s)$ as a comparator and introduce the transfer error

$$
E_{\diamond,\text{bias}}^{(t)} \;:=\; \mathbb{E}\left[\, E_\diamond^{\nu^\star}\big(w_{\diamond,\star}^{(t)};\theta^{(t)}\big) \,\right].
$$

For any state-action distribution ν, we define a Fisher information-like matrix induced by π_θ as

$$
\Sigma_\nu^\theta \;:=\; \mathbb{E}_{(s,a)\,\sim\,\nu}\left[\,\nabla_\theta \log\pi_\theta(a\,|\,s)\,(\nabla_\theta \log\pi_\theta(a\,|\,s))^\top\,\right]
$$

and use $\Sigma_\nu^{(t)}$ to denote $\Sigma_\nu^{\theta^{(t)}}$.

Assumption 23 (Estimation/transfer errors and relative condition number) *The estimation and transfer errors as well as the expected relative condition number are bounded, i.e., $E_{\diamond,\text{est}}^{(t)} \leq \epsilon_{\text{est}}$ and $E_{\diamond,\text{bias}}^{(t)} \leq \epsilon_{\text{bias}}$, for $\diamond = r$ or g, and*

$$
\mathbb{E}\left[\,\sup_{w\,\in\,\mathbb{R}^d} \frac{w^\top \Sigma_{\nu^\star}^{(t)} w}{w^\top \Sigma_{\nu_0}^{(t)} w}\,\right] \;\leq\; \kappa.
$$

We next provide convergence guarantees for the algorithm (26) in Theorem 24 using the approximate update (44). Even though we set $\theta^{(0)} = 0$ and $\lambda^{(0)} = 0$ in the proof of Theorem 24, convergence can be established for arbitrary initial conditions.

Theorem 24 (Convergence and optimality: general policy parametrization) *Let Assumptions 2 and 22 hold and let us fix a state distribution ρ, a state-action distribution ν_0, and $T > 0$. If the iterates $\{(\theta^{(t)}, \lambda^{(t)})\}_{t=0}^{T-1}$ generated by the algorithm (26) using (44) with $\eta_1 = \eta_2 = 1/\sqrt{T}$ satisfy Assumption 23 and $\|w_\diamond^{(t)}\| \le W$, then*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right)\right] \le \frac{C_3}{(1-\gamma)^4}\frac{1}{\sqrt{T}} + \frac{1+2/\xi}{(1-\gamma)^2}\left(\sqrt{\epsilon_{\text{bias}}} + \sqrt{\frac{\kappa\,\epsilon_{\text{est}}}{1-\gamma}}\right)$$

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right)\right]_+ \le \frac{C_4}{(1-\gamma)^3}\frac{1}{\sqrt{T}} + \frac{2+\xi}{1-\gamma}\left(\sqrt{\epsilon_{\text{bias}}} + \sqrt{\frac{\kappa\,\epsilon_{\text{est}}}{1-\gamma}}\right)$$

where $C_3 := 1 + \log|A| + 5\beta W^2/\xi^2$ and $C_4 := (1 + \log|A| + \beta W^2)\xi + (2 + 4\beta W^2)/\xi$.

Proof Since Lemma 19 holds for any smooth policy class that satisfies Assumption 22, it remains to bound $\text{err}_\diamond^{(t)}(\pi^\star)$ for $\diamond = r$ or g. We next separately bound each term on the right-hand side of (38). For the first term,

$$\begin{aligned}
&\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[A_r^{(t)}(s,a) - \left(w_{r,\star}^{(t)}\right)^\top \nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s)\right] \\
&\le \sqrt{\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left(A_r^{(t)}(s,a) - (w_{r,\star}^{(t)})^\top \nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s)\right)^2} \qquad (45)\\
&= \sqrt{E_r^{\nu^\star}\left(w_{r,\star}^{(t)};\theta^{(t)}\right)}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\left(w_{r,\star}^{(t)} - w_r^{(t)}\right)^\top \nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s)\right] \\
&\le \sqrt{\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\left(\left(w_{r,\star}^{(t)} - w_r^{(t)}\right)^\top \nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s)\right)^2\right]} \qquad (46a)\\
&= \sqrt{\left\|w_{r,\star}^{(t)} - w_r^{(t)}\right\|^2_{\Sigma_{\nu^\star}^{(t)}}}.
\end{aligned}$$

Let $\kappa^{(t)} := \left\|\left(\Sigma_{\nu_0}^{(t)}\right)^{-1/2}\Sigma_{\nu^\star}^{(t)}\left(\Sigma_{\nu_0}^{(t)}\right)^{-1/2}\right\|_2$ be the relative condition number at time t. Thus,

$$\begin{aligned}
\left\|w_{r,\star}^{(t)} - w_r^{(t)}\right\|^2_{\Sigma_{\nu^\star}^{(t)}} &\le \left\|\left(\Sigma_{\nu_0}^{(t)}\right)^{-1/2}\Sigma_{\nu^\star}^{(t)}\left(\Sigma_{\nu_0}^{(t)}\right)^{-1/2}\right\| \left\|w_{r,\star}^{(t)} - w_r^{(t)}\right\|^2_{\Sigma_{\nu_0}^{(t)}} \\
&\overset{(a)}{\le} \frac{\kappa^{(t)}}{1-\gamma}\left\|w_{r,\star}^{(t)} - w_r^{(t)}\right\|^2_{\Sigma_{\nu^{(t)}}} \\
&\overset{(b)}{\le} \frac{\kappa^{(t)}}{1-\gamma}\left(E_r^{\nu^{(t)}}\left(w_r^{(t)};\theta^{(t)}\right) - E_r^{\nu^{(t)}}\left(w_{r,\star}^{(t)};\theta^{(t)}\right)\right)
\end{aligned}$$

where we use $(1-\gamma)\nu_0 \leq \nu_{\nu_0}^{\pi^{(t)}} := \nu^{(t)}$ in (a), and we get (b) by the same reasoning as bounding (41). Taking an expectation over the randomness in $w_r^{(t)}$ and $w_{r,\star}^{(t)}$ for the inequality above from both sides yields

$$\begin{aligned}
\mathbb{E}\left[\left\|w_{r,\star}^{(t)} - w_r^{(t)}\right\|^2_{\Sigma_{\nu^\star}^{(t)}}\right] \;&\leq\; \mathbb{E}\left[\frac{\kappa^{(t)}}{1-\gamma}\mathbb{E}\left[E_r^{\nu^{(t)}}(w_r^{(t)};\theta^{(t)}) - E_r^{\nu^{(t)}}(w_{r,\star}^{(t)};\theta^{(t)})\,|\,\theta^{(t)}\right]\right] \\
&\leq\; \mathbb{E}\left[\frac{\kappa^{(t)}}{1-\gamma}\right]\epsilon_{\text{est}} \qquad (46b)\\
&\leq\; \frac{\kappa\,\epsilon_{\text{est}}}{1-\gamma}
\end{aligned}$$

where the last two inequalities are due to Assumption 23.

Substitution of (45) and (46) to the right-hand side of (38) yields an upper bound on $\mathbb{E}\big[\,\text{err}_r^{(t)}(\pi^\star)\,\big]$. By the same reasoning, we can establish a similar bound on $\mathbb{E}\big[\,\text{err}_g^{(t)}(\pi^\star)\,\big]$. Finally, application of these upper bounds to Lemma 19 yields the desired result. ■

We refine the constraint violation in Theorem 24 to be zero by employing the same reasoning as in the proof of Corollary 20. We state it below as Corollary 25 and leave out the proof to avoid repetition.

Corollary 25 (Zero constraint violation: general policy parametrization) *Let Assumptions 2 and 22 hold, let us fix a state distribution ρ and a state-action distribution ν_0, and replace the constraint of Problem (1) by $V_g^\pi(\rho) \geq \bar{b}$, where $\bar{b} := b+\delta$ for some $\delta > 0$. For $\epsilon < \xi/2$, there exists $\delta = \Theta(\epsilon)$ such that if we choose $T = \Omega(1/\epsilon^2)$, $\eta_1 = \eta_2 = 1/\sqrt{T}$, and Assumption 23 holds for $\epsilon_{\text{est}} = \epsilon_{\text{bias}} = O(\epsilon^2)$, then the iterates $\{\theta^{(t)}\}_{t=0}^{T-1}$ generated by the algorithm (26) using (44) satisfy*

$$\begin{aligned}
\text{(Optimality gap)}\;\; \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\big(V_r^\star(\rho) - V_r^{(t)}(\rho)\big)\right] \;&=\; O(\epsilon)\\
\text{(Constraint violation)}\;\; \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\big(b - V_g^{(t)}(\rho)\big)\right]_+ \;&\leq\; 0.
\end{aligned}$$

6 Sample-based NPG-PD algorithms

We now leverage the convergence results established in Theorems 16 and 24 to design two model-free algorithms that utilize sample-based estimates. In particular, we propose a sample-based extension of the NPG-PD algorithm (26) with function approximation and $\Lambda = [\,0, 2/((1-\gamma)\xi)\,]$ as follows

$$\begin{aligned}
\theta^{(t+1)} \;&=\; \theta^{(t)} + \frac{\eta_1}{1-\gamma}\,\hat{w}^{(t)} \\
\lambda^{(t+1)} \;&=\; \mathcal{P}_\Lambda\left(\lambda^{(t)} - \eta_2\big(\hat{V}_g^{(t)}(\rho) - b\big)\right)
\end{aligned} \qquad (47)$$

where $\hat{w}^{(t)}$ and $\hat{V}_g^{(t)}(\rho)$ are the sample-based estimates of the gradient and the value function, respectively. At each time t, we can access a constrained MDP environment by executing a policy π with terminating probability $1-\gamma$. For the minimization problem in (44), we can run the stochastic gradient descent (SGD) for K rounds, $w_{\diamond,k+1} = \mathcal{P}_{\|w_{\diamond,k}\|\leq W}(w_{\diamond,k} - \alpha_k\, G_{\diamond,k})$, where α_k is the stepsize. Here, $G_{\diamond,k}$ is a sample-based estimate of the population gradient $\nabla_\theta E_\diamond^{\nu^{(t)}}(w_\diamond;\theta^{(t)})$:

$$G_{\diamond,k} \;=\; 2\left((w_{\diamond,k})^\top \nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s) \;-\; \hat{A}_\diamond^{(t)}(s,a)\right) \nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s)$$

where $\hat{A}_\diamond^{(t)}(s,a) := \hat{Q}_\diamond^{(t)}(s,a) - \hat{V}_\diamond^{(t)}(s)$, $\hat{Q}_\diamond^{(t)}(s,a)$ and $\hat{V}_\diamond^{(t)}(s)$ are undiscounted sums that are collected in Algorithm 2. In addition, we estimate $\hat{V}_g^{(t)}(\rho)$ using an undiscounted sum in Algorithm 3. As shown in Appendix D, $G_{\diamond,k}$, $\hat{A}_\diamond^{(t)}(s,a)$, and $\hat{V}_g^{(t)}(\rho)$ are unbiased estimates and we approximate the gradient using the average of the SGD iterates $\hat{w}^{(t)} = 2(K(K+1))^{-1}\sum_{k=1}^{K}(k+1)(w_{r,k} + \lambda^{(t)} w_{g,k})$, which is an approximate solution to least-squares regression (Lacoste-Julien et al., 2012).

To establish the sample complexity of Algorithm 1, we assume that the score function $\nabla_\theta \log \pi_\theta(a\,|\,s)$ has a bounded norm and the policy parametrization π_θ has a non-degenerate Fisher information matrix (Zhang et al., 2020b; Agarwal et al., 2021; Liu et al., 2020a).

Assumption 26 (Lipschitz policy) *For $0 \leq t < T$, the policy $\pi^{(t)}$ satisfies*

$$\left\|\nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s)\right\| \;\leq\; L_\pi, \quad \text{where } L_\pi > 0.$$

Assumption 27 (Fisher-non-degenerate policy) *There exists a $\sigma_F > 0$ such that*

$$\Sigma_\nu^\theta \;\succcurlyeq\; \sigma_F I$$

for all ν and $\theta \in \mathbb{R}^d$, where I is the identity matrix in $\mathbb{R}^{d\times d}$.

Assumption 27 holds for the Gaussian policy class when the parameterized mean has a full row-rank Jacobian and the covariance matrix is fixed (Fatkhullin et al., 2023), a policy class in the full rank exponential family (Ding et al., 2022b), and certain neural policies (Liu et al., 2020a). We introduce it to tighten the sample complexity analysis, although this assumption does not necessarily hold for the tabular softmax policy (Fatkhullin et al., 2023).

In Theorem 28, we establish the sample complexity of Algorithm 1.

Theorem 28 (Sample complexity: general policy parametrization) *Let Assumptions 2, 22, 26, and 27 hold and let us fix a state distribution ρ, a state-action distribution ν_0, and $T > 0$. If the iterates $\{(\theta^{(t)},\lambda^{(t)})\}_{t=0}^{T-1}$ are generated by the sample-based NPG-PD method described in Algorithm 1 with $\eta_1 = \eta_2 = 1/\sqrt{T}$ and $\alpha_k = 2/(\sigma_F(k+1))$, in which K rounds of trajectory samples are used at each time t, then*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) \;-\; V_r^{(t)}(\rho)\right)\right] \;\leq\; \frac{C_5}{(1-\gamma)^4}\frac{1}{\sqrt{T}} + \frac{1+2/\xi}{(1-\gamma)^3}\left(\sqrt{\epsilon_{\text{bias}}} + \sqrt{\frac{2\,\kappa\, G^2}{\sigma_F(K+1)}}\right)$$

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b \;-\; V_g^{(t)}(\rho)\right)\right]_+ \;\leq\; \frac{C_6}{(1-\gamma)^3}\frac{1}{\sqrt{T}} + \frac{2+\xi}{(1-\gamma)^2}\left(\sqrt{\epsilon_{\text{bias}}} + \sqrt{\frac{2\,\kappa\, G^2}{\sigma_F(K+1)}}\right)$$

Algorithm 1 Sample-based NPG-PD algorithm with general policy parametrization

Initialization: Learning rates η_1 and η_2, number of SGD iterations K, SGD learning rate $\alpha_k = \frac{2}{\sigma_F(k+1)}$ for $k \geq 0$.
Initialize $\theta^{(0)} = 0$, $\lambda^{(0)} = 0$.
for $t = 0, \ldots, T-1$ **do**
  Initialize $w_{r,0} = w_{g,0} = 0$.
  for $k = 0, 1, \ldots, K-1$ **do**
    Estimate $\hat{A}_r(s,a)$ and $\hat{A}_g(s,a)$ for some $(s,a) \sim \nu^{(t)}$, using Algorithm 2 with policy $\pi_\theta^{(t)}$.
    Take a step of SGD:

$$
\begin{aligned}
w_{r,k+1} &= \mathcal{P}_{\|w_r\| \leq W}\Big(w_{r,k} - 2\alpha_k\big((w_{r,k})^\top \nabla_\theta \log \pi_\theta^{(t)}(s,a) - \hat{A}_r^{(t)}(s,a)\big)\nabla_\theta \log \pi_\theta^{(t)}(s,a)\Big) \\
w_{g,k+1} &= \mathcal{P}_{\|w_g\| \leq W}\Big(w_{g,k} - 2\alpha_k\big((w_{g,k})^\top \nabla_\theta \log \pi_\theta^{(t)}(s,a) - \hat{A}_g^{(t)}(s,a)\big)\nabla_\theta \log \pi_\theta^{(t)}(s,a)\Big).
\end{aligned}
$$

  end for
  Set $\hat{w}^{(t)} = \hat{w}_r^{(t)} + \lambda^{(t)}\hat{w}_g^{(t)}$, where

$$
\hat{w}_r^{(t)} \;=\; \frac{2}{K(K+1)}\sum_{k=0}^{K-1}(k+1)w_{r,k} \;\text{ and }\; \hat{w}_g^{(t)} \;=\; \frac{2}{K(K+1)}\sum_{k=0}^{K-1}(k+1)w_{g,k}.
$$

  Estimate $\hat{V}_g^{(t)}(\rho)$ using Algorithm 3 with policy $\pi_\theta^{(t)}$.
  Natural policy gradient primal-dual update:

$$
\begin{aligned}
\theta^{(t+1)} &= \theta^{(t)} + \eta_1\, \hat{w}^{(t)} \\
\lambda^{(t+1)} &= \mathcal{P}_{[\,0,\, 2/((1-\gamma)\xi)\,]}\Big(\lambda^{(t)} - \eta_2\big(\hat{V}_g^{(t)}(\rho) - b\,\big)\Big).
\end{aligned}
$$

end for

Algorithm 2 A-Unbiased estimate ($\mathcal{A}_\diamond^{\text{est}}$, $\diamond = r$ or g)

Input: Initial state-action distribution ν_0, policy π, discount factor γ.
Sample $(s_0, a_0) \sim \nu_0$, execute the policy π with probability γ at each step h; otherwise, accept (s_h, a_h) as the sample.
Start with (s_h, a_h), execute the policy π with the termination probability $1-\gamma$. Once terminated, add all rewards/utilities from step h onward as $\hat{Q}_\diamond^\pi(s_h, a_h)$ for $\diamond = r$ or g, respectively.
Start with s_h, sample $a_h' \sim \pi(\cdot\,|\,s_h)$, and execute the policy π with the termination probability $1-\gamma$. Once terminated, add all rewards/utilities from step h onward as $\hat{V}_\diamond^\pi(s_h)$ for $\diamond = r$ or g, respectively.
Output: (s_h, a_h) and $\hat{A}_\diamond^\pi(s_h, a_h) := \hat{Q}_\diamond^\pi(s_h, a_h) - \hat{V}_\diamond^\pi(s_h)$, $\diamond = r$ or g.

Algorithm 3 V-Unbiased estimate ($\mathcal{V}_g^{\text{est}}$)

1: **Input**: Initial state distribution ρ, policy π, discount factor γ.
2: Sample $s_0 \sim \rho$, execute the policy π with the termination probability $1-\gamma$. Once terminated, add all utilities up as $\hat{V}_g^\pi(\rho)$.
3: **Output**: $\hat{V}_g^\pi(\rho)$.

where $C_5 := 2 + \log|A| + 5\beta W^2/\xi^2$, $C_6 := (2+\log|A| + \beta W^2)\xi + (2+4\beta W^2)/\xi$, *and* $G^2 := 4(W^2 L_\pi^2 + 2/(1-\gamma)^2)L_\pi^2$.

In Theorem 28, the sampling effect appears as an error of rate $1/\sqrt{K}$, where K is the number of sampled trajectories. This rate follows the standard SGD result (Lacoste-Julien et al., 2012) and it can be increased to $1/K^{1/4}$ under less restrictive assumptions on the policy class (Shamir and Zhang, 2013). When $\epsilon_{\text{bias}} = 0$, it takes $O(1/\epsilon^4)$ sampled trajectories for Algorithm 1 to output an ϵ-optimal policy. The proof of Theorem 28 in Appendix E follows the proof of Theorem 24 except that we use sample-based estimates of gradients in the primal update and sample-based value functions in the dual update. Compared to (Ding et al., 2020, Theorem 3), the improved sample complexity from $O(1/\epsilon^8)$ to $O(1/\epsilon^4)$ is owed to a new regret-type primal-dual analysis in Section 5.2.

Algorithm 4 is utilized for the log-linear policy parametrization. For the feature $\phi_{s,a}$ that has bounded norm $\|\phi_{s,a}\| \leq B$, the sample-based gradient in SGD has the second-order moment bound $G^2 := 4(W^2B^2 + 2/(1-\gamma)^2)B^2$. In Theorem 29, we establish the sample complexity for Algorithm 4; see Appendix F for the proof.

Theorem 29 (Sample complexity: log-linear policy parametrization) *Let Assumption 2 hold and let us fix a state distribution ρ and a state-action distribution ν_0. If the iterates $\{(\theta^{(t)}, \lambda^{(t)})\}_{t=0}^{T-1}$ are generated by the sample-based NPG-PD method described in Algorithm 4 with $\|\phi_{s,a}\| \leq B$, $\eta_1 = \eta_2 = 1/\sqrt{T}$, and $\alpha_k = 2/(\sigma_F(k+1))$, in which K rounds of trajectory samples are used at each time t, and there exists $\sigma_F > 0$ such that $\mathbb{E}_{(s,a)\sim\nu^{(t)}}\left[\phi_{s,a}\phi_{s,a}^\top\right] \succcurlyeq \sigma_F I$, then*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right)\right] \leq \frac{C_5}{(1-\gamma)^4}\frac{1}{\sqrt{T}} + \frac{2+4/\xi}{(1-\gamma)^3}\left(\sqrt{|A|\,\epsilon_{\text{bias}}} + \sqrt{\frac{2\kappa\,|A|\,G^2}{\sigma_F(K+1)}}\right)$$

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b - V_g^{(t)}(\rho)\right)\right]_+ \leq \frac{C_6}{(1-\gamma)^3}\frac{1}{\sqrt{T}} + \frac{4+2\xi}{(1-\gamma)^2}\left(\sqrt{|A|\,\epsilon_{\text{bias}}} + \sqrt{\frac{2\kappa\,|A|\,G^2}{\sigma_F(K+1)}}\right)$$

where $C_5 := 2+\log|A| + 5\beta W^2/\xi^2$ *and* $C_6 := (2+\log|A|+\beta W^2)\xi + (2+4\beta W^2)/\xi$.

When we specialize the log-linear policy to be the softmax policy, Algorithm 4 becomes a sample-based implementation of the NPG-PD method (15) that utilizes the state-action value functions. In this case, $\epsilon_{\text{bias}} = 0$ and $B = 1$ in Theorem 29. When there are no sampling effects, i.e., as $K \to \infty$, our rate $(1/\sqrt{T}, 1/\sqrt{T})$ matches the rate in Theorem 10. It takes $O(1/\epsilon^4)$ sampled trajectories for Algorithm 4 to output an ϵ-optimal policy.

Algorithm 4 Sample-based NPG-PD algorithm with log-linear policy parametrization

Input: Learning rates η_1 and η_2, number of SGD iterations K, SGD learning rate $\alpha_k = \frac{2}{\sigma_F(k+1)}$ for $k \geq 0$.
Initialize $\theta^{(0)} = 0$, $\lambda^{(0)} = 0$.
for $t = 0, \ldots, T-1$ **do**
  Initialize $w_{r,0} = w_{g,0} = 0$.
  for $k = 0, 1, \ldots, K-1$ **do**
    Estimate $\hat{Q}_r^{(t)}(s,a)$ and $\hat{Q}_g^{(t)}(s,a)$ for some $(s,a) \sim \nu^{(t)}$, using Algorithm 5 with log-linear policy $\pi_\theta^{(t)}$.
    Take a step of SGD:

$$
\begin{aligned}
w_{r,k+1} &= \mathcal{P}_{\|w_r\| \leq W} \Big(w_{r,k} \, - \, 2\alpha_k \left(\phi_{s,a}^\top w_{r,k} - \hat{Q}_r^{(t)}(s,a) \right) \phi_{s,a} \Big) \\
w_{g,k+1} &= \mathcal{P}_{\|w_g\| \leq W} \Big(w_{g,k} \, - \, 2\alpha_k \left(\phi_{s,a}^\top w_{g,k} - \hat{Q}_g^{(t)}(s,a) \right) \phi_{s,a} \Big).
\end{aligned}
$$

  end for
  Set $\hat{w}^{(t)} = \hat{w}_r^{(t)} + \lambda^{(t)} \hat{w}_g^{(t)}$, where

$$
\hat{w}_r^{(t)} \; = \; \frac{2}{K(K+1)} \sum_{k=0}^{K-1} (k+1) w_{r,k} \;\; \text{and} \;\; \hat{w}_g^{(t)} \; = \; \frac{2}{K(K+1)} \sum_{k=0}^{K-1} (k+1) w_{g,k}.
$$

  Estimate $\hat{V}_g^{(t)}(\rho)$ using Algorithm 3 with log-linear policy $\pi_\theta^{(t)}$.
  Natural policy gradient primal-dual update:

$$
\begin{aligned}
\theta^{(t+1)} &= \theta^{(t)} + \frac{\eta_1}{1-\gamma} \hat{w}^{(t)} \\
\lambda^{(t+1)} &= \mathcal{P}_{[\,0,\, 2/((1-\gamma)\xi)\,]} \Big(\lambda^{(t)} \, - \, \eta_2 \left(\hat{V}_g^{(t)}(\rho) \, - \, b \right) \Big).
\end{aligned}
\tag{48}
$$

end for

Algorithm 5 Q-Unbiased estimate ($\mathcal{Q}_\diamond^{\text{est}}$, $\diamond = r$ or g)

Input: Initial state-action distribution ν_0, policy π, discount factor γ.
Sample $(s_0, a_0) \sim \nu_0$, execute the policy π with probability γ at each step h; otherwise, accept (s_h, a_h) as the sample.
Start with (s_h, a_h), execute the policy π with the termination probability $1-\gamma$. Once terminated, add all rewards/utilities from step h onward as $\hat{Q}_\diamond^\pi(s_h, a_h)$ for $\diamond = r$ or g, respectively.
Output: (s_h, a_h) and $\hat{Q}_\diamond^\pi(s_h, a_h)$, $\diamond = r$ or g.

Remark 30 (Zero constraint violation for sample-based NPG-PD algorithms) *As done in Corollary 20, we can refine the constraint violation in Theorem 29 to be zero. Given a small desired accuracy* $\epsilon > 0$*, there exists* $\delta = \Theta(\epsilon)$ *such that if* $T = K = \Omega(1/\epsilon^2)$*, and* $\epsilon_{\text{bias}} = O(\epsilon^2)$*, then*

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

Similarly, we can strengthen Theorem 28 to achieve zero constraint violation.

7 Computational experiments

We utilize a set of robotic tasks to demonstrate the effectiveness of our sample-based NPG-PD method described in Algorithm 1. In our computational experiments, the robotic agents are trained to move along a straight line or in a plane with speed limits for safety (Zhang et al., 2020c). We compare the performance of our NPG-PD algorithm with two classes of representative state-of-the-art methods: (i) two classical primal-dual policy search methods: Trust Region Policy Optimization based Lagrangian (TRPOLag) method and Proximal Policy Optimization based Lagrangian (PPOLag) method (Ray et al., 2019); (ii) two methods that utilize the state-of-the-art policy optimization techniques: Constrained Update Projection (CUP) approach (Yang et al., 2022) and First Order Constrained Optimization in Policy Space (FOCOPS) algorithm (Zhang et al., 2020c). We conduct computational experiments in the OmniSafe framework (Ji et al., 2024) and implement robotic environments using the OpenAI Gym (Brockman et al., 2016) for the MuJoCo simulators (Todorov et al., 2012).

We train six MuJoCo robotic agents to walk: Ant-v1, Humanoid-v1, HalfCheetah-v1, Walker2d-v1, Hopper-v1, and Swimmer-v1, while constraining the moving speed to be under a given threshold. Figure 2 shows that, in the first two tasks, our NPG-PD algorithm uniformly outperforms other four methods by reaching higher rewards while maintaining similar constraint satisfaction costs. This superior performance of NPG-PD is also demonstrated in HalfCHeetah-v1 and Walker2d-v1 tasks in Figure 3; in particular, we note that NPG-PD achieves a performance similar to that of PPOLag and that they both outperform the other three methods in Walker2d-v1 task. On the other hand, PPOLag does not perform well in Hopper-v1 in Figure 4. For the last two tasks, Figure 4 shows a competitive performance of NPG-PD with two state-of-the-art methods: FOCOPS and CUP. Even though early oscillatory behavior slows down convergence of NPG-PD in Hopper-v1, it achieves higher rewards than CUP and FOCOPS. This demonstrates that NPG-PD not only converges faster than classical Lagrangian-based primal-dual methods but also matches the performance of state-of-the-art policy optimization methods.

8 Concluding remarks

We have proposed a Natural Policy Gradient Primal-Dual (NPG-PD) algorithm for solving the optimal control problems in constrained MDPs. Our algorithm utilized natural policy gradient ascent to update the primal variable and projected subgradient descent to update the dual variable. Although the underlying maximization involves a nonconcave objective function and a nonconvex constraint set, we have established global convergence for both softmax

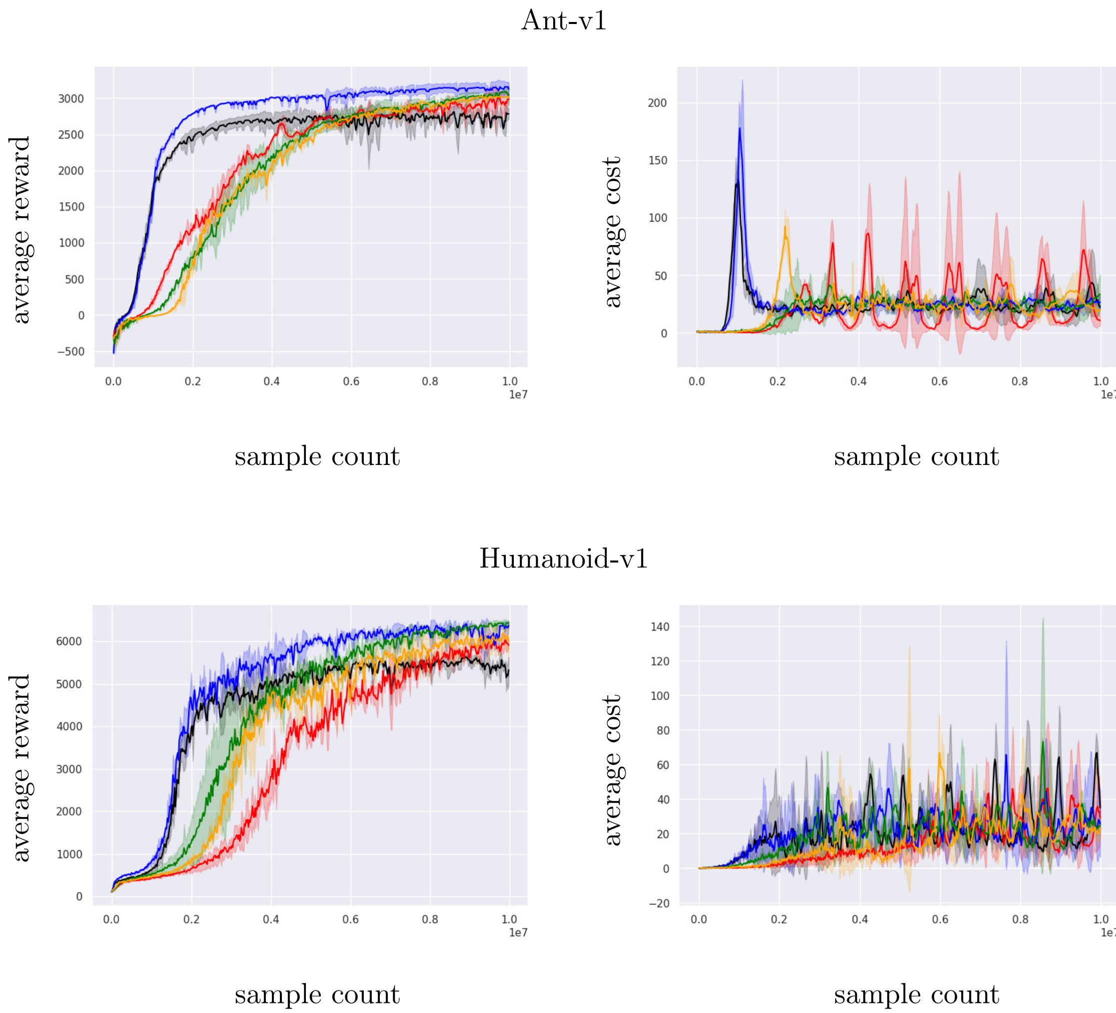

Figure 2: Learning curves of NPG-PD method (—, blue), CUP (Yang et al., 2022) (—, red), FOCOPS (Zhang et al., 2020c) (—, orange), TRPOLag (Ray et al., 2019) (—, black), and PPOLag (Ray et al., 2019) (—, green) for Ant-v1 and Humanoid-v1 robotic tasks with the speed limit 25. The vertical axes represent the average reward and the average cost (i.e., average speed). The solid lines show the means of 1000 bootstrap samples obtained over 3 random seeds and the shaded regions display the bootstrap 95% confidence intervals.

or general smooth policy parametrizations, and have provided finite-sample complexity guarantees for two model-free extensions of the NPG-PD algorithm. To the best of our knowledge, our work is the first to offer finite-time performance guarantees for policy-based primal-dual methods in the context of discounted infinite-horizon constrained MDPs.

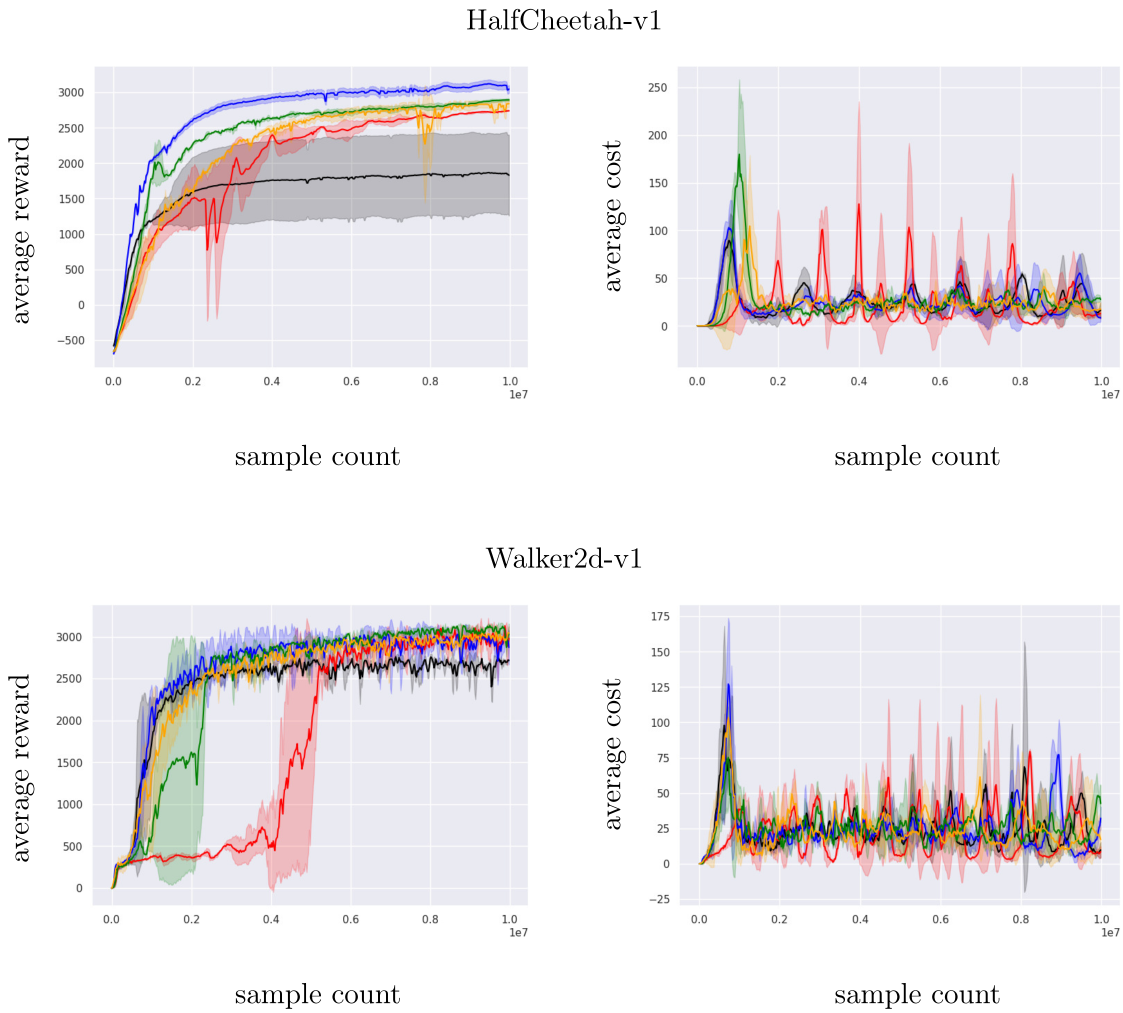

Figure 3: Learning curves of NPG-PD method (—, blue), CUP (Yang et al., 2022) (—, red), FOCOPS (Zhang et al., 2020c) (—, orange), TRPOLag (Ray et al., 2019) (—, black), and PPOLag (Ray et al., 2019) (—, green) for HalfCheetah-v1 and Walker2d-v1 robotic tasks with the speed limit 25. The vertical axes represent the average reward and the average cost (i.e., average speed). The solid lines show the means of 1000 bootstrap samples obtained over 3 random seeds and the shaded regions display the bootstrap 95% confidence intervals.

In future work, we will address the oscillatory behavior commonly observed in primal-dual methods (Stooke et al., 2020; Moskovitz et al., 2023). To ensure policy-iterate convergence, recent works have incorporated optimism and regularization into single-timescale primal-dual algorithms (Ding et al., 2023; Müller et al., 2024). It is relevant to examine whether similar techniques can achieve policy-iterate convergence in the function approximation setting, e.g.,

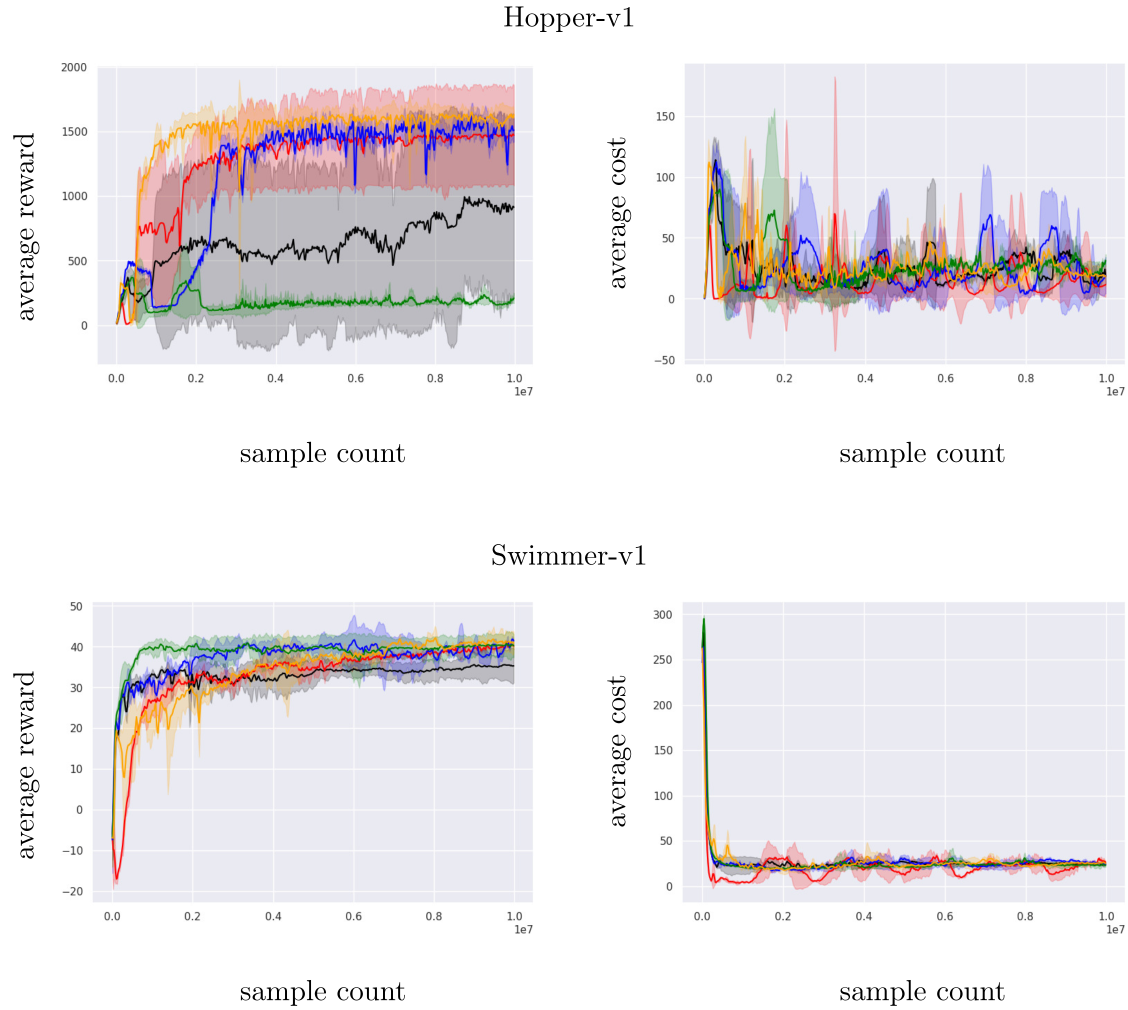

Figure 4: Learning curves of NPG-PD method (—, blue), CUP (Yang et al., 2022) (—, red), FOCOPS (Zhang et al., 2020c) (—, orange), TRPOLag (Ray et al., 2019) (—, black), and PPOLag (Ray et al., 2019) (—, green) for Hopper-v1 and Swimmer-v1 robotic tasks with the speed limit 25. The vertical axes represent the average reward and the average cost (i.e., average speed). The solid lines show the means of 1000 bootstrap samples obtained over 3 random seeds and the shaded regions display the bootstrap 95% confidence intervals.

using regularization (Montenegro et al., 2024). We also aim to extend our policy gradient primal-dual algorithms to other constrained MDP settings, e.g., those with continuous state/action spaces (Rozada et al., 2025), as well as to those with risk-sensitive and resilient constraints (Chow et al., 2017; Ding et al., 2024). Additional future directions include further improving the sample efficiency of policy gradient primal-dual algorithms (Mondal and

Aggarwal, 2024), examining the robustness against model uncertainties (Zhang et al., 2024), and developing constrained policy optimization methods that work with offline datasets (Hong et al., 2024; Wei et al., 2024).

Acknowledgments and Disclosure of Funding

Research of D. Ding and M. R. Jovanović was supported by the National Science Foundation under Awards ECCS-1708906 and ECCS-1809833. Research of K. Zhang and T. Başar was supported by the US Army Research Laboratory (ARL) Cooperative Agreement W911NF-17-2-0196 and by the Office of Naval Research (ONR) MURI Grant N00014-16-1-2710. K. Zhang also acknowledges the support from the Army Research Laboratory (ARL) Grant W911NF-24-1-0085, the NSF CAREER Award ECCS-2443704, and the AFOSR YIP Award FA9550-25-1-0258. T. Başar also acknowledges the support provided by the Army Research Office (ARO) under Grant W911NF-24-1-0085.

Appendix A. Proof of Lemma 6

We prove Lemma 6 by providing a constrained MDP example as shown in Figure 1. States s_3, s_4, and s_5 are three terminal states with zero reward and utility. We consider a 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

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

We consider the following two policies $\pi^{(1)}$ and $\pi^{(2)}$ using the softmax parametrization (7):

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

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

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 a 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 a 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.

Appendix B. Proof of Theorem 7

Let us first recall the notion of occupancy measure (Altman, 1999). An occupancy measure q^π of a policy π is defined as a set of distributions generated by executing π:

$$q_{s,a}^\pi \;=\; \sum_{t=0}^{\infty} \gamma^t \, P^\pi(\, s_t = s, a_t = a \,|\, s_0 \sim \rho\,) \tag{49}$$

for all $s \in S$ and $a \in A$, where $P^\pi\,(s_t = s, a_t = a \,|\, s_0 \sim \rho)$ is the probability of visiting a state-action pair (s,a) under the policy π for an initial state s_0. For brevity, we put all $q_{s,a}^\pi$ together as $q^\pi \in \mathbb{R}^{|S||A|}$ and $q_a^\pi = [\,q_{1,a}^\pi, \cdots, q_{|S|,a}^\pi\,]^\top$. For an action a, we collect all transition probabilities $P(s' \,|\, s, a)$ for $s', s \in S$ to have shorthand notation $P_a \in \mathbb{R}^{|S|\times|S|}$. The occupancy measure q^π has to satisfy a set of linear constraints given by $\mathcal{Q} := \{q^\pi \in \mathbb{R}^{|S||A|} \,|\, \sum_{a\,\in\,A}(I - \gamma P_a^\top) q_a^\pi = \rho \text{ and } q^\pi \geq 0\}$. With a slight abuse of notation, we write $r \in [0,1]^{|S||A|}$ and $g \in [0,1]^{|S||A|}$. Thus, the value functions V_r^π, V_g^π: $S \to \mathbb{R}$ under the initial state distribution ρ are linear in q^π:

$$V_r^\pi(\rho) \;=\; \langle q^\pi, r\rangle \;:=\; F_r(q^\pi) \;\;\text{and}\;\; V_g^\pi(\rho) \;=\; \langle q^\pi, g\rangle \;:=\; F_g(q^\pi).$$

We are now in a position to consider the primal problem (5) as a linear program:

$$\underset{q^\pi\,\in\,\mathcal{Q}}{\text{maximize}} \;\; F_r(q^\pi) \;\;\text{subject to}\;\; F_g(q^\pi) \;\geq\; b \tag{50}$$

where the maximization is over all occupancy measures $q^\pi \in \mathcal{Q}$. Once we compute a solution q^π, the associated policy solution π can be recovered via

$$\pi(a\,|\,s) \;=\; \frac{q_{s,a}^\pi}{\sum_{a\,\in\,A} q_{s,a}^\pi} \;\;\text{for all } s \in S, a \in A. \tag{51}$$

Abstractly, we let π^q: $\mathcal{Q} \to \Delta_A^{|S|}$ be a mapping from an occupancy measure q^π to a policy π. Similarly, as defined by (49) we let q^π: $\Delta_A^{|S|} \to \mathcal{Q}$ be a mapping from a policy π to an occupancy measure q^π. Clearly, $q^\pi = (\pi^q)^{-1}$.

Despite the nonconvexity essence of (5) in policy space, the reformulation (50) reveals the underlying convexity in occupancy measure q^π. In Lemma 31, we exploit this convexity to show the average policy improvement over T steps.

Lemma 31 (Bounded average performance) *Let assumptions in Theorem 7 hold. Then, the iterates $\{(\theta^{(t)}, \lambda^{(t)})\}_{t=0}^{T-1}$ generated by the PG-PD method (13) satisfy*

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(F_r(q^{\theta^\star}) - F_r(q^{\theta^{(t)}})\right) + \frac{1}{T}\sum_{t=0}^{T-1}\lambda^{(t)}\left(F_g(q^{\theta^\star}) - F_g(q^{\theta^{(t)}})\right) \;\leq\; \frac{D_\theta L_\theta}{T^{1/4}} \tag{52}$$

where $D_\theta = \frac{8|S|}{(1-\gamma)^2}\|d_\rho^{\pi^\star}/\rho\|_\infty^2$ and $L_\theta = \frac{2|A|(1+2/\xi)}{(1-\gamma)^4}$.

Proof From the dual update in (13), we have $0 \leq \lambda^{(t)} \leq 2/((1-\gamma)\xi)$. From the smooth property of the value functions under the direct policy parametrization (Agarwal et al., 2021, Lemma D.3), we have

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

If we fix $\lambda^{(t)} \geq 0$, then

$$\begin{aligned} & \left| (F_r + \lambda^{(t)} F_g)(q^\theta) \,-\, (F_r + \lambda^{(t)} F_g)(q^{\theta^{(t)}}) \,-\, \left\langle \nabla_\theta F_r(q^{\theta^{(t)}}) + \lambda^{(t)} \nabla_\theta F_g(q^{\theta^{(t)}}), \theta - \theta^{(t)} \right\rangle \right| \\ \leq\;& \frac{L_\theta}{2} \left\| \theta - \theta^{(t)} \right\|^2. \end{aligned}$$

Thus,

$$\begin{aligned} (F_r + \lambda^{(t)} F_g)(q^\theta) \;&\geq\; (F_r + \lambda^{(t)} F_g)(q^{\theta^{(t)}}) \,+\, \left\langle \nabla_\theta F_r(q^{\theta^{(t)}}) + \lambda^{(t)} \nabla_\theta F_g(q^{\theta^{(t)}}), \theta - \theta^{(t)} \right\rangle \\ &\quad -\, \frac{L_\theta}{2} \left\| \theta - \theta^{(t)} \right\|^2 \\ &\geq\; (F_r + \lambda^{(t)} F_g)(q^\theta) \,-\, L_\theta \left\| \theta - \theta^{(t)} \right\|^2. \end{aligned} \tag{53}$$

We note that the primal update in (13) is equivalent to

$$\begin{aligned} \theta^{(t+1)} \;=\; \underset{\theta \,\in\, \Theta}{\operatorname{argmax}} \Bigg\{ & V_r^{\theta^{(t)}}(\rho) \,+\, \lambda^{(t)} V_g^{\theta^{(t)}}(\rho) \\ & + \left\langle \nabla_\theta V_r^{\theta^{(t)}}(\rho) + \lambda^{(t)} \nabla_\theta V_g^{\theta^{(t)}}(\rho), \theta - \theta^{(t)} \right\rangle \,-\, \frac{1}{2\eta_1} \left\| \theta - \theta^{(t)} \right\|^2 \Bigg\}. \end{aligned}$$

By taking $\eta_1 = 1/L_\theta$ and $\theta = \theta^{(t+1)}$ in (53), we have

$$\begin{aligned} & (F_r + \lambda^{(t)} F_g)(q^{\theta^{(t+1)}}) \\ \geq\;& \underset{\theta \,\in\, \Theta}{\text{maximize}} \Bigg\{ (F_r + \lambda^{(t)} F_g)(q^{\theta^{(t)}}) \\ & \quad + \left\langle \nabla_\theta F_r(q^{\theta^{(t)}}) \,+\, \lambda^{(t)} \nabla_\theta F_g(q^{\theta^{(t)}}), \theta - \theta^{(t)} \right\rangle \,-\, \frac{L_\theta}{2} \left\| \theta - \theta^{(t)} \right\|^2 \Bigg\} \qquad (54) \\ \geq\;& \underset{\theta \,\in\, \Theta}{\text{maximize}} \left\{ (F_r + \lambda^{(t)} F_g)(q^\theta) \,-\, L_\theta \left\| \theta - \theta^{(t)} \right\|^2 \right\} \\ \geq\;& \underset{\alpha \,\in\, [0,1]}{\text{maximize}} \left\{ (F_r + \lambda^{(t)} F_g)(q^{\theta_\alpha}) \,-\, L_\theta \left\| \theta_\alpha - \theta^{(t)} \right\|^2 \right\} \end{aligned}$$

where $\theta_\alpha := \pi^q(\alpha q^{\theta^\star} + (1-\alpha) q^{\theta^{(t)}})$, we apply (53) for the second inequality, and the last inequality is due to $\pi^q \circ q^\pi = \mathrm{id}_{SA}$ and linearity of q^θ in θ. Since F_r and F_g are linear in q^θ, we have

$$(F_r + \lambda^{(t)} F_g)(q^{\theta_\alpha}) \;=\; \alpha (F_r + \lambda^{(t)} F_g)(q^{\theta^\star}) \,+\, (1-\alpha)(F_r + \lambda^{(t)} F_g)(q^{\theta^{(t)}}). \tag{55}$$

By the definition of π^q,

$$(\pi^q(q) - \pi^q(q'))_{sa} \;=\; \frac{1}{\sum_{a\,\in\,A} q_{sa}}(q_{sa} - q'_{sa}) \,+\, \frac{\sum_{a\,\in\,A} q'_{sa} - \sum_{a\,\in\,A} q_{sa}}{\sum_{a\,\in\,A} q_{sa} \sum_{a\,\in\,A} q'_{sa}} q'_{sa}$$

which, together with $\|x+y\|^2 \le 2\|x\|^2 + 2\|y\|^2$, gives

$$\begin{aligned}
&\|\pi^q(q) - \pi^q(q')\|^2 \\
\le\;& 2 \sum_{s\,\in\,S} \sum_{a\,\in\,A} \frac{(q_{sa} - q'_{sa})^2}{(\sum_{a\,\in\,A} q_{sa})^2} \,+\, 2 \sum_{s\,\in\,S} \sum_{a\,\in\,A} \left(\frac{\sum_{a\,\in\,A} q'_{sa} - \sum_{a\,\in\,A} q_{sa}}{\sum_{a\,\in\,A} q_{sa} \sum_{a\,\in\,A} q'_{sa}} \right)^2 (q'_{sa})^2 \\
\le\;& 2 \sum_{s\,\in\,S} \frac{1}{(\sum_{a\,\in\,A} q_{sa})^2} \left(\sum_{a\,\in\,A} (q_{sa} - q'_{sa})^2 \,+\, \left(\sum_{a\,\in\,A} q'_{sa} - \sum_{a\,\in\,A} q_{sa} \right)^2 \right).
\end{aligned}$$

Therefore,

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

in which the upper bound further can be relaxed into

$$\begin{aligned}
& \sum_{s\,\in\,S} \frac{4\alpha^2}{\left(\sum_{a\,\in\,A} q_{sa}^{\theta^{(t)}}\right)^2} \left(\left(\sum_{a\,\in\,A} q_{sa}^{\theta^\star} \right)^2 \,+\, \left(\sum_{a\,\in\,A} q_{sa}^{\theta^{(t)}} \right)^2 \right) \\
=\;& 4\alpha^2 \sum_{s\,\in\,S} \frac{\left(d_\rho^{\pi^\star}(s)\right)^2 + \left(d_\rho^{\pi^{(t)}}(s)\right)^2}{\left(d_\rho^{\pi^{(t)}}(s)\right)^2} \\
\le\;& 4\alpha^2 |S| \,+\, 4\alpha^2 |S| \left\| \frac{d_\rho^{\pi^\star}}{d_\rho^{\pi^{(t)}}} \right\|_\infty^2 \\
\le\;& 4\alpha^2 |S| \left(1 \,+\, \frac{1}{(1-\gamma)^2} \left\| \frac{d_\rho^{\pi^\star}}{\rho} \right\|_\infty^2 \right) \\
\le\;& \alpha^2 D_\theta
\end{aligned} \tag{56}$$

where we apply $d_\rho^{\pi^{(t)}} \geq (1-\gamma)\rho$ componentwise in the second inequality.

We now apply (55) and (56) to (54),

$$\begin{aligned}
& (F_r + \lambda^{(t)} F_g)(q^{\theta^\star}) \,-\, (F_r + \lambda^{(t)} F_g)(q^{\theta^{(t+1)}}) \\
\leq \;\; & \underset{\alpha\,\in\,[0,1]}{\text{minimize}} \left\{ L_\theta \left\| \theta_\alpha - \theta^{(t)} \right\|^2 \,+\, (F_r + \lambda^{(t)} F_g)(q^{\theta^\star}) \,-\, (F_r + \lambda^{(t)} F_g)(q^{\theta_\alpha}) \right\} \\
\leq \;\; & \underset{\alpha\,\in\,[0,1]}{\text{minimize}} \left\{ \alpha^2 D_\theta L_\theta \,+\, (1-\alpha)\big((F_r + \lambda^{(t)} F_g)(q^{\theta^\star}) \,-\, (F_r + \lambda^{(t)} F_g)(q^{\theta^{(t)}}) \big) \right\}
\end{aligned}$$

which further implies

$$\begin{aligned}
& (F_r + \lambda^{(t+1)} F_g)(q^{\theta^\star}) \,-\, (F_r + \lambda^{(t+1)} F_g)(q^{\theta^{(t+1)}}) \\
\leq \;\; & \underset{\alpha\,\in\,[0,1]}{\text{minimize}} \left\{ \alpha^2 D_\theta L_\theta \,+\, (1-\alpha)\big((F_r + \lambda^{(t)} F_g)(q^{\theta^\star}) \,-\, (F_r + \lambda^{(t)} F_g)(q^{\theta^{(t)}}) \big) \right\} \qquad (57) \\
& - \, (\lambda^{(t)} - \lambda^{(t+1)})(F_g(q^{\theta^\star}) - F_g(q^{\theta^{(t+1)}})).
\end{aligned}$$

We check the right-hand side of the inequality (57). By the dual update in (13), it is easy to see that $-(\lambda^{(t)} - \lambda^{(t+1)})(F_g(q^{\theta^\star}) - F_g(q^{\theta^{(t+1)}})) \leq |\lambda^{(t)} - \lambda^{(t+1)}|/(1-\gamma) \leq \eta_2/(1-\gamma)^2$. We can solve the minimization problem in (57) by taking $\alpha = 0$ if $\alpha^{(t)} < 0$; $\alpha = 1$ if $\alpha^{(t)} > 0$; $\alpha = \alpha^{(t)}$ if $\alpha^{(t)} \in [0,1]$, where

$$\alpha^{(t)} \;:=\; \frac{(F_r + \lambda^{(t)} F_g)(q^{\theta^\star}) \,-\, (F_r + \lambda^{(t)} F_g)(q^{\theta^{(t)}})}{2 D_\theta L_\theta}.$$

We next discuss three cases.

(i) When $\alpha^{(t)} < 0$, we set $\alpha = 0$ for (57),

$$(F_r + \lambda^{(t+1)} F_g)(q^{\theta^\star}) \,-\, (F_r + \lambda^{(t+1)} F_g)(q^{\theta^{(t+1)}}) \;\leq\; \frac{D_\theta L_\theta}{2\sqrt{T}}; \qquad (58)$$

(ii) When $\alpha^{(t)} > 1$, we set $\alpha = 1$ that leads to $(F_r + \lambda^{(t+1)} F_g)(q^{\theta^\star}) - (F_r + \lambda^{(t+1)} F_g)(q^{\theta^{(t+1)}}) \leq \frac{3}{2} D_\theta L_\theta$, i.e., $\alpha^{(t+1)} \leq 3/4$. Thus, this case reduces to the next case.

(iii) When $0 \leq \alpha^{(t)} \leq 1$, we can express (57) as

$$\begin{aligned}
& (F_r + \lambda^{(t+1)} F_g)(q^{\theta^\star}) - (F_r + \lambda^{(t+1)} F_g)(q^{\theta^{(t+1)}}) \\
\leq \;\; & \left(1 - \tfrac{(F_r + \lambda^{(t)} F_g)(q^{\theta^\star}) - (F_r + \lambda^{(t)} F_g)(q^{\theta^{(t)}})}{4 D_\theta L_\theta} \right) \times \left((F_r + \lambda^{(t)} F_g)(q^{\theta^\star}) - (F_r + \lambda^{(t)} F_g)(q^{\theta^{(t)}}) \right) \\
& + \frac{D_\theta L_\theta}{2\sqrt{T}}
\end{aligned}$$

or equivalently

$$\alpha^{(t+1)} \;\leq\; \left(1-\frac{\alpha^{(t)}}{2}\right)\alpha^{(t)} \;+\; \frac{1}{4\sqrt{T}}. \tag{59}$$

By choosing $\lambda^{(0)} = 0$ and $\theta^{(0)}$ such that $V_r^{\theta^{(0)}}(\rho) \geq V_r^{\theta^\star}(\rho)$, we know that $\alpha^{(0)} \leq 0$. Thus, $\alpha^{(1)} \leq 1/(4\sqrt{T})$. By (58), the case $\alpha^{(1)} \leq 0$ is trivial. Without loss of generality, we assume that $0 \leq \alpha^{(t)} \leq 1/T^{1/4} \leq 1$. By induction over t for (59),

$$\alpha^{(t+1)} \;\leq\; \left(1-\frac{\alpha^{(t)}}{2}\right)\alpha^{(t)} \;+\; \frac{1}{4\sqrt{T}} \;\leq\; \frac{1}{T^{1/4}}. \tag{60}$$

By combining (58) and (60), and averaging over $t = 0, 1, \cdots, T-1$, we get the desired bound. ∎

Proof [Proof of Theorem 7]
Bounding the optimality gap. By the dual update (13) and $\lambda^{(0)} = 0$, it is convenient to bound $(\lambda^{(T)})^2$ by

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

where the inequality is due to the feasibility of the optimal policy $\pi^\star$ or the associated occupancy measure $q^\star = q^{\theta^\star}$: $F_g(q^\star) \geq b$, and $|F_g(q^{\theta^{(t)}}) - b| \leq 1/(1-\gamma)$. The above inequality further implies

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

By substituting the above inequality into (52) in Lemma 31, we obtain the desired optimality gap bound, where we take $\eta_2 = (1-\gamma)^2 D_\theta L_\theta/(2\sqrt{T})$.
Bounding the constraint violation. From the dual update in (13), we have for any $\lambda \in [\,0, 2/((1-\gamma)\xi)\,]$

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

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

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

which further implies

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

We note that $F_g(q^{\theta^\star}) \geq b$. By adding the inequality above to (52) in Lemma 31 from both sides, we have

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

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

$$F_r(q^{\theta^\star}) - F_r(q') + \frac{2}{(1-\gamma)\xi}\left[b - F_g(q')\right]_+ \leq \frac{D_\theta L_\theta}{T^{1/4}} + \frac{1}{2\eta_2(1-\gamma)^2\xi^2 T} + \frac{\eta_2}{2(1-\gamma)^2}$$

where there exists q' such that $F_r(q') := \frac{1}{T}\sum_{t=0}^{T-1} F_r(q^{\theta^{(t)}})$ and $F_g(q') := \frac{1}{T}\sum_{t=0}^{T-1} F_g(q^{\theta^{(t)}})$ by the definition of occupancy measure.

From Lemma 3 (ii), we have $\lambda^\star \leq 1/((1-\gamma)\xi)$. Application of Lemma 5 with $C = 2/((1-\gamma)\xi)$ yields

$$\left[b - F_g(q')\right]_+ \leq \frac{(1-\gamma)\xi D_\theta L_\theta}{T^{1/4}} + \frac{1}{2\eta_2(1-\gamma)\xi T} + \frac{\eta_2 \xi}{2(1-\gamma)}$$

which readily leads to the desired constraint violation bound by noting that

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(b - F_g(q^{\theta^{(t)}})\right) = b - F_g(q')$$

and taking $\eta_2 = (1-\gamma)^2 D_\theta L_\theta/(2\sqrt{T})$ and $\|d_\rho^{\pi^\star}/\rho\|_\infty^2 \geq (1-\gamma)^2$.

■

Appendix C. Proof of Lemma 9

The dual update follows Lemma 3. Since $\lambda^\star \leq (V_r^\star(\rho) - V_r^{\bar\pi}(\rho))/\xi$ with $0 \leq V_r^\star,\, V_r^{\bar\pi} \leq 1/(1-\gamma)$, we take a projection interval $\Lambda = [\,0, 2/((1-\gamma)\xi)\,]$ such that the upper bound $2/((1-\gamma)\xi)$ satisfies $2/((1-\gamma)\xi) \geq 2\lambda^\star$.

We now verify the primal update. We expand the primal update in (14) into

$$\theta^{(t+1)} \;=\; \theta^{(t)} \,+\, \eta_1 F_\rho^\dagger(\theta^{(t)}) \nabla_\theta V_r^{\theta^{(t)}}(\rho) \,+\, \eta_1 \lambda^{(t)} F_\rho^\dagger(\theta^{(t)}) \nabla_\theta V_g^{\theta^{(t)}}(\rho). \tag{61}$$

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

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

Using the Moore-Penrose inverse, an optimal solution reads

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

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 NPG direction $F_\rho^\dagger(\theta)\nabla_\theta V_r^{\pi_\theta,\lambda}(\rho)$.

On the other hand, it is easy to verify that $A_r^{\pi_\theta}$ is a minimizer of $E_r(w)$. The softmax parametrization (7) implies that

$$\frac{\partial \log \pi_\theta(a\,|\,s)}{\partial \theta_{s',a'}} \;=\; \mathbb{I}\{s = s'\}\left(\mathbb{I}\{a = a'\} - \pi_\theta(a'\,|\,s)\right) \tag{62}$$

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

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

The above equality together with the fact: $\sum_{a\,\in\, A}\pi_\theta(a\,|\,s)A_r^{\pi_\theta,\lambda}(s,a) = 0$, show 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 (62) yields

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

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:

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

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 by, up to some state-dependent offset

$$F_\rho^\dagger(\theta)\nabla_\theta V_r^{\pi_\theta}(\rho) \;=\; \frac{A_r^{\pi_\theta}}{1-\gamma} + u \tag{63}$$

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

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

$$F_\rho^\dagger(\theta)\nabla_\theta V_g^{\pi_\theta}(\rho) \;=\; \frac{A_g^{\pi_\theta}}{1-\gamma} + v \tag{64}$$

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

Substituting (63) and (64) into the primal update (61) 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 $c_s = d_s = 0$.

Appendix D. Sample-based NPG-PD algorithm with function approximation

We describe a sample-based NPG-PD algorithm with function approximation in Algorithm 1. We calculate the computational complexity of Algorithm 1 as follows: each round has expected length $2/(1-\gamma)$, yielding the expected number of total samples $4KT/(1-\gamma)$; 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 $O(dKT + KT/(1-\gamma))$.

The following unbiased estimates that are useful in our analysis.

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

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

By a similar agument as above,

$$\begin{aligned}
\mathbb{E}\left[\hat{Q}_r^{(t)}(s,a)\right] &= \mathbb{E}\left[\sum_{k=0}^{K'-1} r(s_k,a_k) \,\middle|\, \theta^{(t)}, s_0 = s, a_0 = a\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{\infty} \mathbb{I}\{K'-1 \geq k \geq 0\} r(s_k,a_k) \,\middle|\, \theta^{(t)}, s_0 = s, a_0 = a\right] \\
&= \sum_{k=0}^{\infty} \mathbb{E}\left[\mathbb{E}_{K'}\left[\mathbb{I}\{K'-1 \geq k \geq 0\}\right] r(s_k,a_k) \,\middle|\, \theta^{(t)}, s_0 = s, a_0 = a\right] \\
&= \sum_{k=0}^{\infty} \mathbb{E}\left[\gamma^k r(s_k,a_k) \,\middle|\, \theta^{(t)}, s_0 = s, a_0 = a\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(s_k,a_k) \,\middle|\, \theta^{(t)}, s_0 = s, a_0 = a\right] \\
&= Q_r^{(t)}(s,a).
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\hat{A}_r^{(t)}(s,a)\right] = \mathbb{E}\left[\hat{Q}_r^{(t)}(s,a)\right] - \mathbb{E}\left[\hat{V}_r^{(t)}(s)\right] = Q_r^{(t)}(s,a) - V_r^{(t)}(s) = A_r^{(t)}(s,a).$$

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

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

where (a) is due to $0 \leq g(x_k,a_k) \leq 1$ and $V_g^{(t)}(s) \geq 0$ and (b) is clear from $K' \sim \text{Geo}(1-\gamma)$. Similarly, we have the variance bound $\text{Var}\left[\hat{Q}_r^{(t)}(s,a)\right] \leq \frac{1}{(1-\gamma)^2}$.

By the sampling scheme of Algorithm 2, we can show that $G_{r,k}$ is an unbiased estimate of the population gradient $\nabla_\theta E_r^{\nu^{(t)}}(w_r;\theta^{(t)})$:

$$
\begin{aligned}
\mathbb{E}_{(s,a)\sim d^{(t)}}\left[G_{\diamond,k}\right] &= 2\mathbb{E}\left[\left(w_{r,k}^\top \nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s) \,-\, \hat{A}_r^{(t)}(s,a)\right)\nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s)\right] \\
&= 2\mathbb{E}\left[\left(w_{r,k}^\top \nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s) \,-\, \mathbb{E}\left[\hat{A}_r^{(t)}(s,a)\,\middle|\, s,a\right]\right)\nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s)\right] \\
&= 2\mathbb{E}\left[\left(w_{r,k}^\top \nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s) \,-\, A_r^{(t)}(s,a)\right)\nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s)\right] \\
&= \nabla_{w_r} E_r^{\nu^{(t)}}(w_r;\theta^{(t)}).
\end{aligned}
$$

Appendix E. Proof of Theorem 28

We first adapt Lemma 19 to the sample-based case as follows.

Lemma 32 (Sample-based regret/violation lemma) *Let Assumption 2 hold and let us fix a state distribution ρ and $T>0$. Assume that $\log\pi_\theta(a\,|\,s)$ is β-smooth in θ for any (s,a). If the iterates $\{(\theta^{(t)},\lambda^{(t)})\}_{t=0}^{T-1}$ are generated by Algorithm 1 with $\theta^{(0)}=0$, $\lambda^{(0)}=0$, $\eta_1=\eta_2=1/\sqrt{T}$, and $\|\hat{w}_r^{(t)}\|$, $\|\hat{w}_g^{(t)}\|\le W$, then*

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) \,-\, V_r^{(t)}(\rho)\right)\right] &\le \frac{C_5}{(1-\gamma)^5}\frac{1}{\sqrt{T}} + \sum_{t=0}^{T-1}\frac{\mathbb{E}\left[\widehat{\mathrm{err}}_r^{(t)}(\pi^\star)\right]}{(1-\gamma)T} + \sum_{t=0}^{T-1}\frac{2\mathbb{E}\left[\widehat{\mathrm{err}}_g^{(t)}(\pi^\star)\right]}{(1-\gamma)^2\xi T} \\
\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(b \,-\, V_g^{(t)}(\rho)\right)\right]_+ &\le \frac{C_6}{(1-\gamma)^4}\frac{1}{\sqrt{T}} + \sum_{t=0}^{T-1}\frac{\xi\mathbb{E}\left[\widehat{\mathrm{err}}_r^{(t)}(\pi^\star)\right]}{T} + \sum_{t=0}^{T-1}\frac{2\mathbb{E}\left[\widehat{\mathrm{err}}_g^{(t)}(\pi^\star)\right]}{(1-\gamma)T}
\end{aligned}
$$

where $C_5 = 2+\log|A|+5\beta W^2/\xi^2$, $C_6=(2+\log|A|+\beta W^2)\xi+(2+4\beta W^2)/\xi$, and

$$
\widehat{\mathrm{err}}_\diamond^{(t)}(\pi) \;:=\; \left|\mathbb{E}_{s\sim d_\rho^\pi}\mathbb{E}_{a\sim\pi(\cdot\,|\,s)}\left[A_\diamond^{(t)}(s,a) \,-\, \left(\hat{w}_\diamond^{(t)}\right)^\top\nabla_\theta\log\pi_\theta^{(t)}(a\,|\,s)\right]\right|, \quad \textit{where } \diamond = r \textit{ or } g.
$$

Proof The smoothness of log-linear policy in conjunction with an application of Taylor's theorem to $\log\pi_\theta^{(t)}(a\,|\,s)$ yield

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

where $\theta^{(t+1)} - \theta^{(t)} = \frac{\eta_1}{1-\gamma}\hat{w}^{(t)}$. 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_{\rm KL}\left(\pi^\star(\cdot\,|\,s)\,\|\,\pi_\theta^{(t)}(\cdot\,|\,s)\right) \,-\, D_{\rm KL}\left(\pi^\star(\cdot\,|\,s)\,\|\,\pi_\theta^{(t+1)}(\cdot\,|\,s)\right)\right] \\
= \;\; & -\,\mathbb{E}_{s\sim d^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\log\frac{\pi_\theta^{(t)}(a\,|\,s)}{\pi_\theta^{(t+1)}(a\,|\,s)}\right] \\
\geq \;\; & \eta_1\mathbb{E}_{s\sim d^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\left(\hat{w}^{(t)}\right)^\top\nabla_\theta\log\pi_\theta^{(t)}(a\,|\,s)\right] \,-\, \beta\frac{\eta_1^2}{2(1-\gamma)^2}\left\|\hat{w}^{(t)}\right\|^2 \\
= \;\; & \eta_1\mathbb{E}_{s\sim d^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\left(\hat{w}_r^{(t)}\right)^\top\nabla_\theta\log\pi_\theta^{(t)}(a\,|\,s)\right] \\
& +\,\eta_1\lambda^{(t)}\mathbb{E}_{s\sim d^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\left(\hat{w}_g^{(t)}\right)^\top\nabla_\theta\log\pi_\theta^{(t)}(a\,|\,s)\right] \,-\, \beta\frac{\eta_1^2}{2(1-\gamma)^2}\left\|\hat{w}^{(t)}\right\|^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) \\
& +\,\eta_1\mathbb{E}_{s\sim d^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\left(\hat{w}_r^{(t)}+\lambda^{(t)}\hat{w}_g^{(t)}\right)^\top\nabla_\theta\log\pi_\theta^{(t)}(a\,|\,s)-\big(A_r^{(t)}(s,a)+\lambda^{(t)}A_g^{(t)}(s,a)\big)\right] \\
& -\,\beta\frac{\eta_1^2}{(1-\gamma)^2}\left(\left\|\hat{w}_r^{(t)}\right\|^2+\big(\lambda^{(t)}\big)^2\left\|\hat{w}_g^{(t)}\right\|^2\right) \\
\geq \;\; & \eta_1(1-\gamma)\big(V_r^\star(\rho)-V_r^{(t)}(\rho)\big) \,+\, \eta_1(1-\gamma)\lambda^{(t)}\big(V_g^\star(\rho)\,-\,V_g^{(t)}(\rho)\big) \\
& -\,\eta_1\widehat{\rm err}_r^{(t)}(\pi^\star) \,-\, \eta_1\lambda^{(t)}\widehat{\rm err}_g^{(t)}(\pi^\star) \,-\, \beta\frac{\eta_1^2\,W^2}{(1-\gamma)^2} \,-\, \beta\frac{\eta_1^2\,W^2}{(1-\gamma)^2}\big(\lambda^{(t)}\big)^2
\end{aligned}
$$

where $\hat{w}^{(t)} = \hat{w}_r^{(t)} + \lambda^{(t)}\hat{w}_g^{(t)}$ for a given $\lambda^{(t)}$, in the last inequality we apply the performance difference lemma, notation of $\widehat{\rm err}_r^{(t)}(\pi^\star)$ and $\widehat{\rm err}_g^{(t)}(\pi^\star)$, and $\|\hat{w}_r^{(t)}\|$, $\|\hat{w}_g^{(t)}\| \leq W$.

Rearranging the inequality above leads to

$$
\begin{aligned}
& V_r^\star(\rho) \,-\, V_r^{(t)}(\rho) \\
\leq \;\; & \frac{1}{1-\gamma}\frac{1}{\eta_1}\mathbb{E}_{s\sim d^\star}\left[D_{\rm KL}\left(\pi^\star(\cdot\,|\,s)\,\|\,\pi_\theta^{(t)}(\cdot\,|\,s)\right) \,-\, D_{\rm KL}\left(\pi^\star(\cdot\,|\,s)\,\|\,\pi_\theta^{(t+1)}(\cdot\,|\,s)\right)\right] \\
& +\,\frac{1}{1-\gamma}\widehat{\rm err}_r^{(t)}(\pi^\star) \,+\, \frac{2}{(1-\gamma)^2\xi}\widehat{\rm err}_g^{(t)}(\pi^\star) \,+\, \beta\frac{\eta_1 W^2}{(1-\gamma)^3} \,+\, \beta\frac{4\eta_1 W^2}{(1-\gamma)^5\xi^2} \\
& -\,\lambda^{(t)}\big(V_g^\star(\rho)\,-\,V_g^{(t)}(\rho)\big)
\end{aligned}
$$

where we utilize $0 \leq \lambda^{(t)} \leq 2/((1-\gamma)\xi)$ from the dual update of Algorithm 1.

Therefore,

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right)\\
\leq\;& \frac{1}{(1-\gamma)\eta_1 T}\sum_{t=0}^{T-1}\mathbb{E}_{s\sim d^\star}\left[D_{\mathrm{KL}}\left(\pi^\star(\cdot\,|\,s)\,\|\,\pi_\theta^{(t)}(\cdot\,|\,s)\right) - D_{\mathrm{KL}}(\pi^\star\left(\cdot\,|\,s)\,\|\,\pi_\theta^{(t+1)}(\cdot\,|\,s)\right)\right]\\
&+\frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\widehat{\mathrm{err}}_r^{(t)}(\pi^\star) + \frac{2}{(1-\gamma)^2\xi T}\sum_{t=0}^{T-1}\widehat{\mathrm{err}}_g^{(t)}(\pi^\star) + \beta\frac{\eta_1 W^2}{(1-\gamma)^3} + \beta\frac{4\eta_1 W^2}{(1-\gamma)^5\xi^2}\\
&-\frac{1}{T}\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho) - V_g^{(t)}(\rho)\right)\\
\leq\;& \frac{\log|A|}{(1-\gamma)\eta_1 T} + \frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\widehat{\mathrm{err}}_r^{(t)}(\pi^\star) + \frac{2}{(1-\gamma)^2\xi T}\sum_{t=0}^{T-1}\widehat{\mathrm{err}}_g^{(t)}(\pi^\star)\\
&+\beta\frac{\eta_1 W^2}{(1-\gamma)^3} + \beta\frac{4\eta_1 W^2}{(1-\gamma)^5\xi^2} + \frac{1}{T}\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^\pi(\rho) - V_g^{(t)}(\rho)\right)
\end{aligned}
$$

where in the last inequality we take a telescoping sum of the first sum and drop a non-positive term. Taking the expectation over the randomness in sampling on both sides of the inequality above yields

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(V_r^\star(\rho) - V_r^{(t)}(\rho)\right)\right] + \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\lambda^{(t)}\left(V_g^\star(\rho) - V_g^{(t)}(\rho)\right)\right]\\
\leq\;& \frac{\log|A|}{(1-\gamma)\eta_1 T} + \frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\mathbb{E}\left[\widehat{\mathrm{err}}_r^{(t)}(\pi^\star)\right] + \frac{2}{(1-\gamma)^2\xi T}\sum_{t=0}^{T-1}\mathbb{E}\left[\widehat{\mathrm{err}}_g^{(t)}(\pi^\star)\right]\\
&+\beta\frac{\eta_1 W^2}{(1-\gamma)^3} + \beta\frac{4\eta_1 W^2}{(1-\gamma)^5\xi^2}.
\end{aligned} \tag{65}
$$

Proving the first inequality. From the dual update in Algorithm 1, we have

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

where the second 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 $\hat{V}_g^{(t)}(\rho)$ is an estimate that is independent of $\lambda^{(t)}$ given the past history, $\lambda^{(t)}$ is independent of $V_g^{(t)}(\rho)-\hat{V}_g^{(t)}(\rho)$ at time t and thus $\mathbb{E}\big[\,\lambda^{(t)}\big(V_g^{(t)}(\rho)-\hat{V}_g^{(t)}(\rho)\big)\,\big]=0$ due to the fact $\mathbb{E}\big[\,\hat{V}_g^{(t)}(\rho)\,\big]=V_g^{(t)}(\rho)$; see it in Appendix D. Therefore,

$$-\,\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\lambda^{(t)}\big(\,V_g^{\star}(\rho)\,-\,V_g^{(t)}(\rho)\,\big)\right]\;\leq\;\mathbb{E}\left[\frac{\eta_2}{2T}\sum_{t=0}^{T-1}\big(\,\hat{V}_g^{(t)}(\rho)\,-\,b\,\big)^2\right]\;\leq\;\frac{2\eta_2}{(1-\gamma)^2}\tag{66}$$

where in the second inequality we drop a non-positive term and use the fact

$$\mathbb{E}\left[\,\big(\hat{V}_g^{(t)}(\rho)\big)^2\,\right]\;=\;\mathrm{Var}\left[\,\hat{V}_g^{(t)}(s)\,\right]\,+\,\left(\mathbb{E}\left[\,\hat{V}_g^{(t)}(s)\,\right]\right)^2\;\leq\;\frac{2}{(1-\gamma)^2}$$

where the inequality is due to that $\mathrm{Var}\big[\,\hat{V}_g^{(t)}(s)\,\big]\leq 1/(1-\gamma)^2$; see it in Appendix D, and $\mathbb{E}\left[\,\hat{V}_g^{(t)}(\rho)\,\right]=V_g^{(t)}(\rho)$, where $0\leq V_g^{(t)}(s)\leq 1/(1-\gamma)$.

Adding the inequality (66) to (65) on both sides and taking $\eta_1=\eta_2=1/\sqrt{T}$ yield the first inequality.

Proving the second inequality. From the dual update in Algorithm 1, we have for any $\lambda\in\Lambda:=\big[\,0,2/((1-\gamma)\xi)\,\big]$

$$\begin{aligned}
&\mathbb{E}\left[|\lambda^{(t+1)}-\lambda|^2\right]\\
&=\;\;\mathbb{E}\left[\,\left|\mathcal{P}_\Lambda\left(\lambda^{(t)}\,-\,\eta_2\big(\,\hat{V}_g^{(t)}(\rho)\,-\,b\,\big)\right)-\mathcal{P}_\Lambda(\lambda)\right|^2\,\right]\\
&\overset{(a)}{\leq}\;\;\mathbb{E}\left[\,\left|\lambda^{(t)}\,-\,\eta_2\big(\,\hat{V}_g^{(t)}(\rho)\,-\,b\,\big)\,-\,\lambda\right|^2\,\right]\\
&=\;\;\mathbb{E}\left[\,\left|\lambda^{(t)}\,-\,\lambda\right|^2\,\right]\,-\,2\eta_2\,\mathbb{E}\left[\,\big(\,\hat{V}_g^{(t)}(\rho)\,-\,b\,\big)\,\big(\lambda^{(t)}\,-\,\lambda\big)\,\right]\,+\,\eta_2^2\,\mathbb{E}\left[\,\big(\,\hat{V}_g^{(t)}(\rho)\,-\,b\,\big)^2\,\right]\\
&\overset{(b)}{\leq}\;\;\mathbb{E}\left[\,\left|\lambda^{(t)}\,-\,\lambda\right|^2\,\right]\,-\,2\eta_2\,\mathbb{E}\left[\,\big(\,\hat{V}_g^{(t)}(\rho)\,-\,b\,\big)\big(\,\lambda^{(t)}\,-\,\lambda\,\big)\,\right]\,+\,\frac{3\eta_2^2}{(1-\gamma)^2}
\end{aligned}$$

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

$$\begin{aligned}
0\;\;&\leq\;\;\frac{1}{T}\,\mathbb{E}\left[\,|\lambda^{(T)}\,-\,\lambda|^2\,\right]\\
&\leq\;\;\frac{1}{T}\,\mathbb{E}\left[\,\left|\lambda^{(0)}\,-\,\lambda\right|^2\,\right]\,-\,\frac{2\eta_2}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\,\big(\,\hat{V}_g^{(t)}(\rho)\,-\,b\,\big)\big(\,\lambda^{(t)}\,-\,\lambda\,\big)\,\right]\,+\,\frac{3\eta_2^2}{(1-\gamma)^2}
\end{aligned}$$

which further implies that

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

where we use $\mathbb{E}\big[\,\hat{V}_g^{(t)}(\rho)\,\big] = V_g^{(t)}(\rho)$ and $\lambda^{(t)}$ is independent of $\hat{V}_g^{(t)}(\rho)$ given the past history. We now add the above inequality into (65) on both sides and utilize $V_g^\star(\rho) \geq b$:

$$
\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\big(\,V_r^\star(\rho) \,-\, V_r^{(t)}(\rho)\,\big)\right] + \lambda\,\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\big(\,b \,-\, V_g^{(t)}(\rho)\,\big)\right]\\
&\leq\;\; \frac{\log|A|}{(1-\gamma)\eta_1 T} + \frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\mathbb{E}\left[\mathrm{err}_r^{(t)}(\pi^\star)\right] + \frac{2}{(1-\gamma)^2\xi T}\sum_{t=0}^{T-1}\mathbb{E}\left[\mathrm{err}_g^{(t)}(\pi^\star)\right]\\
&\quad + \beta\frac{\eta_1 W^2}{(1-\gamma)^3} + \beta\frac{4\eta_1 W^2}{(1-\gamma)^5\xi^2} + \frac{1}{2\eta_2 T}\mathbb{E}\left[\left|\lambda^{(0)}-\lambda\right|^2\right] + \frac{2\eta_2}{(1-\gamma)^2}.
\end{aligned}
$$

By taking $\lambda = \frac{2}{(1-\gamma)\xi}$ when $\sum_{t=0}^{T-1}\big(b - V_g^{(t)}(\rho)\big) \geq 0$; otherwise $\lambda = 0$, we reach

$$
\begin{aligned}
&\mathbb{E}\left[V_r^\star(\rho) \,-\, \frac{1}{T}\sum_{t=0}^{T-1}V_r^{(t)}(\rho)\right] + \frac{2}{(1-\gamma)\xi}\mathbb{E}\left[b \,-\, \frac{1}{T}\sum_{t=0}^{T-1}V_g^{(t)}(\rho)\right]_+\\
&\leq\;\; \frac{\log|A|}{(1-\gamma)\eta_1 T} + \frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\mathbb{E}\left[\mathrm{err}_r^{(t)}(\pi^\star)\right] + \frac{2}{(1-\gamma)^2\xi T}\sum_{t=0}^{T-1}\mathbb{E}\left[\mathrm{err}_g^{(t)}(\pi^\star)\right]\\
&\quad + \beta\frac{\eta_1 W^2}{(1-\gamma)^3} + \beta\frac{4\eta_1 W^2}{(1-\gamma)^5\xi^2} + \frac{2}{\eta_2(1-\gamma)^2\xi^2 T} + \frac{2\eta_2}{(1-\gamma)^2}.
\end{aligned}
$$

Since $V_r^{(t)}(\rho)$ and $V_g^{(t)}(\rho)$ are linear functions in the occupancy measure (Altman, 1999, Chapter 10), there exists a policy π' such that $V_r^{\pi'}(\rho) = \frac{1}{T}\sum_{t=0}^{T-1}V_r^{(t)}(\rho)$ and $V_g^{\pi'}(\rho) = \frac{1}{T}\sum_{t=0}^{T-1}V_g^{(t)}(\rho)$. Hence,

$$
\begin{aligned}
&\mathbb{E}\left[V_r^\star(\rho) \,-\, V_r^{\pi'}(\rho)\right] + \frac{2}{(1-\gamma)\xi}\mathbb{E}\left[b \,-\, V_g^{\pi'}(\rho)\right]_+\\
&\leq\;\; \frac{\log|A|}{(1-\gamma)\eta_1 T} + \frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\mathbb{E}\left[\mathrm{err}_r^{(t)}(\pi^\star)\right] + \frac{2}{(1-\gamma)^2\xi T}\sum_{t=0}^{T-1}\mathbb{E}\left[\mathrm{err}_g^{(t)}(\pi^\star)\right]\\
&\quad + \beta\frac{\eta_1 W^2}{(1-\gamma)^3} + \beta\frac{4\eta_1 W^2}{(1-\gamma)^5\xi^2} + \frac{2}{\eta_2(1-\gamma)^2\xi^2 T} + \frac{2\eta_2}{(1-\gamma)^2}.
\end{aligned}
$$

From Lemma 3 (ii), we have $\lambda^\star \leq 1/((1-\gamma)\xi)$. Application of Lemma 5 with $C = 2/((1-\gamma)\xi)$ yields

$$
\begin{aligned}
\mathbb{E}\left[b \,-\, V_g^{\pi'}(\rho)\right]_+ \;\;\leq\;\; & \frac{\xi\log|A|}{\eta_1 T} + \frac{\xi}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\mathrm{err}_r^{(t)}(\pi^\star)\right] + \frac{2}{(1-\gamma)T}\sum_{t=0}^{T-1}\mathbb{E}\left[\mathrm{err}_g^{(t)}(\pi^\star)\right]\\
& + \beta\frac{\eta_1\xi W^2}{(1-\gamma)^2} + \beta\frac{4\eta_1 W^2}{(1-\gamma)^4\xi} + \frac{2}{\eta_2(1-\gamma)\xi T} + \frac{2\eta_2\,\xi}{(1-\gamma)}.
\end{aligned}
$$

which leads to our constraint violation bound if we utilize $\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\big(b - V_g^{(t)}(\rho)\big)\right] = \mathbb{E}\left[b - V_g^{\pi'}(\rho)\right]$ and taking $\eta_1 = \eta_2 = 1/\sqrt{T}$. ■

Proof [Proof of Theorem 28]

By Lemma 32, we only need to consider the randomness in sequences of $\hat{w}^{(t)}$ and bound $\mathbb{E}\left[\operatorname{err}_\diamond^{(t)}(\pi^\star)\right]$ for $\diamond = r$ or g. Application of the triangle inequality yields

$$\begin{aligned}\widehat{\operatorname{err}}_r^{(t)}(\pi^\star) \;\; & \leq \;\; \left|\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[A_r^{(t)}(s,a) \,-\, \left(w_{r,\star}^{(t)}\right)^\top\nabla_\theta\log\pi_\theta^{(t)}(a\,|\,s)\right]\right| \\ & \quad + \left|\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\left(w_{r,\star}^{(t)} \,-\, \hat{w}_r^{(t)}\right)^\top\nabla_\theta\log\pi_\theta^{(t)}(a\,|\,s)\right]\right|\end{aligned} \tag{67}$$

where $w_{r,\star}^{(t)} \in \operatorname{argmin}_{\|w_r\|_2\leq W} E_r^{\nu^{(t)}}(w_r;\theta^{(t)})$. We next bound each term on the right-hand side of (67), separately. For the first term,

$$\begin{aligned}& \mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[A_r^{(t)}(s,a) \,-\, \left(w_{r,\star}^{(t)}\right)^\top\nabla_\theta\log\pi_\theta^{(t)}(a\,|\,s)\right] \\ \leq \;\; & \sqrt{\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left(A_r^{(t)}(s,a) \,-\, \left(w_{r,\star}^{(t)}\right)^\top\nabla_\theta\log\pi_\theta^{(t)}(a\,|\,s)\right)^2} \\ = \;\; & \sqrt{E_r^{\nu^\star}\left(w_{r,\star}^{(t)};\theta^{(t)}\right)}.\end{aligned} \tag{68}$$

Similarly,

$$\begin{aligned}& \mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\left(w_{r,\star}^{(t)} \,-\, \hat{w}_r^{(t)}\right)^\top\nabla_\theta\log\pi_\theta^{(t)}(a\,|\,s)\right] \\ \leq \;\; & \sqrt{\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\left(\left(w_{r,\star}^{(t)} \,-\, \hat{w}_r^{(t)}\right)^\top\nabla_\theta\log\pi_\theta^{(t)}(a\,|\,s)\right)^2\right]} \\ = \;\; & \sqrt{\left\|w_{r,\star}^{(t)} \,-\, \hat{w}_r^{(t)}\right\|^2_{\Sigma_{\nu^\star}^{(t)}}}.\end{aligned} \tag{69}$$

We let $\kappa^{(t)} := \left\|\left(\Sigma_{\nu_0}^{(t)}\right)^{-1/2}\Sigma_{\nu^\star}^{(t)}\left(\Sigma_{\nu_0}^{(t)}\right)^{-1/2}\right\|_2$ be the relative condition number at time t. Thus,

$$\begin{aligned}\left\|w_{r,\star}^{(t)} \,-\, \hat{w}_r^{(t)}\right\|^2_{\Sigma_{\nu^\star}^{(t)}} \;\; & \leq \;\; \left\|\left(\Sigma_{\nu_0}^{(t)}\right)^{-1/2}\Sigma_{\nu^\star}^{(t)}\left(\Sigma_{\nu_0}^{(t)}\right)^{-1/2}\right\|\left\|w_{r,\star}^{(t)} \,-\, \hat{w}_r^{(t)}\right\|^2_{\Sigma_{\nu_0}^{(t)}} \\ & \overset{(a)}{\leq} \;\; \frac{\kappa^{(t)}}{1-\gamma}\left\|w_{r,\star}^{(t)} \,-\, \hat{w}_r^{(t)}\right\|^2_{\Sigma_{\nu^{(t)}}} \\ & \overset{(b)}{\leq} \;\; \frac{\kappa^{(t)}}{1-\gamma}\left(E_r^{\nu^{(t)}}\left(\hat{w}_r^{(t)};\theta^{(t)}\right) \,-\, E_r^{\nu^{(t)}}\left(w_{r,\star}^{(t)};\theta^{(t)}\right)\right)\end{aligned} \tag{70}$$

where we use $(1-\gamma)\nu_0 \leq \nu_{\nu_0}^{\pi^{(t)}} := \nu^{(t)}$ in (a), and we get (b) due to that the first-order optimality condition for $w_{r,\star}^{(t)}$:

$$\left(w_r \,-\, w_{r,\star}^{(t)}\right)^\top\nabla_w E_r^{\nu^{(t)}}\left(w_{r,\star}^{(t)};\theta^{(t)}\right) \;\geq\; 0, \;\; \text{for any } w_r \text{ satisfying } \|w_r\| \,\leq\, W$$

further implies that

$$
\begin{aligned}
& E_r^{\nu^{(t)}}\big(w_r;\theta^{(t)}\big) \,-\, E_r^{\nu^{(t)}}\big(w_{r,\star}^{(t)};\theta^{(t)}\big) \\
&= \; \mathbb{E}_{(s,a)\sim\nu^{(t)}}\left[\left(A_r^{(t)}(s,a) \,-\, \phi_{s,a}^\top w_{r,\star}^{(t)} \,+\, \phi_{s,a}^\top w_{r,\star}^{(t)} \,-\, \phi_{s,a}^\top w_r\right)^2\right] \,-\, E_r^{\nu^{(t)}}\big(w_{r,\star}^{(t)};\theta^{(t)}\big) \\
&= \; 2\left(w_{r,\star}^{(t)} \,-\, w_r\right)^\top \mathbb{E}_{s,a\sim\nu^{(t)}}\left[\left(A_r^{(t)}(s,a) \,-\, \phi_{s,a}^\top w_{r,\star}^{(t)}\right)\phi_{s,a}\right] \\
&\quad\; + \mathbb{E}_{(s,a)\sim\nu^{(t)}}\left[\left(\phi_{s,a}^\top w_{r,\star}^{(t)} \,-\, \phi_{s,a}^\top w_r\right)^2\right] \\
&= \; \left(w_r \,-\, w_{r,\star}^{(t)}\right)^\top \nabla_w E_r^{\nu^{(t)}}(w_{r,\star}^{(t)};\theta^{(t)}) \,+\, \left\|w_r \,-\, w_{r,\star}^{(t)}\right\|_{\Sigma_{\nu^{(t)}}}^2 \\
&\geq \; \left\|w_r \,-\, w_{r,\star}^{(t)}\right\|_{\Sigma_{\nu^{(t)}}}^2 .
\end{aligned}
$$

Taking an expectation over (70) from both sides yields

$$
\begin{aligned}
\mathbb{E}\left[\left\|w_{r,\star}^{(t)} \,-\, w_r^{(t)}\right\|_{\Sigma_{\nu^\star}^{(t)}}^2\right] \;&\leq\; \mathbb{E}\left[\frac{\kappa^{(t)}}{1-\gamma}\,\mathbb{E}\left[E_r^{\nu^{(t)}}\big(\hat{w}_r^{(t)};\theta^{(t)}\big) \,-\, E_r^{\nu^{(t)}}\big(w_{r,\star}^{(t)};\theta^{(t)}\big)\,\Big|\,\theta^{(t)}\right]\right] \\
&\overset{(a)}{\leq}\; \mathbb{E}\left[\frac{\kappa^{(t)}}{1-\gamma}\,\frac{2G^2}{\sigma_F(K+1)}\right] \\
&\overset{(b)}{\leq}\; \frac{2\kappa\, G^2}{\sigma_F(1-\gamma)(K+1)}
\end{aligned}
\tag{71}
$$

where (a) is due to the standard SGD result (Lacoste-Julien et al., 2012): for $\alpha_k = 2/(\sigma_F(k+1))$,

$$
E_{r,\text{est}}^{(t)} \;=\; \mathbb{E}\left[E_r^{\nu^{(t)}}\big(\hat{w}_r^{(t)};\theta^{(t)}\big) \,-\, E_r^{\nu^{(t)}}\big(w_{r,\star}^{(t)};\theta^{(t)}\big)\right] \;\leq\; \frac{2G^2}{\sigma_F(K+1)}
$$

and (b) follows from Assumption 23. Here, it is straightforward to check the second-order moment of stochastic gradient $G_{\diamond,k}$ using Assumption 26:

$$
\mathbb{E}\left[\|G_{\diamond,k}\|^2\right] \;\leq\; 4L_\pi^2\left(W^2L_\pi^2 \,+\, \frac{2}{(1-\gamma)^2}\right) \;=\; G^2.
$$

Substitution of (69) and (71) into the right-hand side of (67) yields an upper bound on $\mathbb{E}\big[\,\text{err}_r^{(t)}(\pi^\star)\,\big]$. By the same reasoning, we can establish a similar bound on $\mathbb{E}\big[\,\text{err}_g^{(t)}(\pi^\star)\,\big]$. Finally, application of these upper bounds to Lemma 32 leads to our desired results. ■

Appendix F. Proof of Theorem 29

By $\|\phi_{s,a}\| \leq B$, for the log-linear policy class, $\log \pi_\theta(a\,|\,s)$ is β-smooth with $\beta = B^2$. By Lemma 32, we only need to consider the randomness in the sequence of $\hat{w}^{(t)}$ and the error bounds for $\mathbb{E}\big[\,\widehat{\text{err}}_r^{(t)}(\pi^\star)\,\big]$ and $\mathbb{E}\big[\,\widehat{\text{err}}_g^{(t)}(\pi^\star)\,\big]$. We first use (67) and consider the following cases. By (31) and $A_r^{(t)}(s,a) = Q_r^{(t)}(s,a) - \mathbb{E}_{a'\sim\pi_\theta^{(t)}(\cdot\,|\,s)}\Big[\,Q_r^{(t)}(s,a')\,\Big]$,

$$
\begin{aligned}
& \mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\,A_r^{(t)}(s,a) \;-\; \left(w_{r,\star}^{(t)}\right)^\top \nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s)\,\right] \\
&= \;\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\,Q_r^{(t)}(s,a) \;-\; \phi_{s,a}^\top w_{r,\star}^{(t)}\,\right] \\
&\quad\; - \mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a'\sim\pi_\theta^{(t)}(\cdot\,|\,s)}\left[\,Q_r^{(t)}(s,a') \;-\; \phi_{s,a'}^\top w_{r,\star}^{(t)}\,\right] \\
&\leq \;\sqrt{\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\left(Q_r^{(t)}(s,a) \;-\; \phi_{s,a}^\top w_{r,\star}^{(t)}\right)^2\right]} \\
&\quad\; + \sqrt{\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a'\sim\pi_\theta^{(t)}(\cdot\,|\,s)}\left[\left(Q_r^{(t)}(s,a') \;-\; \phi_{s,a'}^\top w_{r,\star}^{(t)}\right)^2\right]} \\
&\leq \;2\sqrt{|A|\,\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim \text{Unif}_A}\left[\left(Q_r^{(t)}(s,a) \;-\; \phi_{s,a}^\top w_{r,\star}^{(t)}\right)^2\right]} \\
&= \;2\sqrt{|A|\,\mathcal{E}_r^{\nu^\star}\big(w_{r,\star}^{(t)};\theta^{(t)}\big)}.
\end{aligned}
\tag{72}
$$

Similarly,

$$
\begin{aligned}
& \mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\,\left(w_{r,\star}^{(t)} \;-\; \hat{w}_r^{(t)}\right)^\top \nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s)\,\right] \\
&= \;\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\,\left(w_{r,\star}^{(t)} \;-\; \hat{w}_r^{(t)}\right)^\top \phi_{s,a}\,\right] \\
&\quad\; - \mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a'\sim\pi_\theta^{(t)}(\cdot\,|\,s)}\left[\,\left(w_{r,\star}^{(t)} \;-\; \hat{w}_r^{(t)}\right)^\top \phi_{s,a'}\,\right] \\
&\leq \;2\sqrt{|A|\mathbb{E}_{s\sim d_\rho^\star}\mathbb{E}_{a\sim \text{Unif}_A}\left[\left(\left(w_{r,\star}^{(t)} \;-\; \hat{w}_r^{(t)}\right)^\top \phi_{s,a}\right)^2\right]} \\
&= \;2\sqrt{|A|\,\left\|w_{r,\star}^{(t)} \;-\; \hat{w}_r^{(t)}\right\|_{\Sigma_{\nu^\star}}^2}
\end{aligned}
\tag{73}
$$

where $\Sigma_{\nu^\star} := \mathbb{E}_{(s,a)\sim\nu^\star}\big[\,\phi_{s,a}\phi_{s,a}^\top\,\big]$. By the definition of κ,

$$
\left\|w_{r,\star}^{(t)} \;-\; \hat{w}_r^{(t)}\right\|_{\Sigma_{\nu^\star}}^2 \;\leq\; \kappa\left\|w_{r,\star}^{(t)} \;-\; \hat{w}_r^{(t)}\right\|_{\Sigma_{\nu_0}}^2 \;\leq\; \frac{\kappa}{1-\gamma}\left\|w_{r,\star}^{(t)} \;-\; \hat{w}_r^{(t)}\right\|_{\Sigma_{\nu^{(t)}}}^2
\tag{74}
$$

where we use $(1-\gamma)\nu_0 \leq \nu_{\nu_0}^{\pi^{(t)}} := \nu^{(t)}$ in the second inequality. We note that $w_{r,\star}^{(t)} \in \text{argmin}_{\|w_r\|_2\leq W}\ \mathcal{E}_r^{\nu^{(t)}}(w_r;\theta^{(t)})$. Application of the first-order optimality condition for $w_{r,\star}^{(t)}$

yields

$$\left(w_r \,-\, w_{r,\star}^{(t)}\right)^\top \nabla_\theta \mathcal{E}_r^{\nu^{(t)}}\big(w_{r,\star}^{(t)};\theta^{(t)}\big) \;\geq\; 0, \;\;\text{for any } w_r \text{ satisfying } \|w_r\| \,\leq\, W.$$

Thus,

$$\begin{aligned}
& \mathcal{E}_r^{\nu^{(t)}}\big(w_r;\theta^{(t)}\big) \,-\, \mathcal{E}_r^{\nu^{(t)}}\big(w_{r,\star}^{(t)};\theta^{(t)}\big) \\
= \;\; & \mathbb{E}_{s,a\,\sim\,\nu^{(t)}}\left[\left(Q_r^{(t)}(s,a) \,-\, \phi_{s,a}^\top w_{r,\star}^{(t)} \,+\, \phi_{s,a}^\top w_{r,\star}^{(t)} \,-\, \phi_{s,a}^\top w_r\right)^2\right] \,-\, \mathcal{E}_r^{\nu^{(t)}}\big(w_{r,\star}^{(t)};\theta^{(t)}\big) \\
= \;\; & 2\left(w_{r,\star}^{(t)} \,-\, w_r\right)^\top \mathbb{E}_{s,a\,\sim\,\nu^{(t)}}\left[\left(Q_r^{(t)}(s,a) \,-\, \phi_{s,a}^\top w_{r,\star}^{(t)}\right)\phi_{s,a}\right] \\
& + \mathbb{E}_{s,a\,\sim\,\nu^{(t)}}\left[\left(\phi_{s,a}^\top w_{r,\star}^{(t)} \,-\, \phi_{s,a}^\top w_r\right)^2\right] \\
= \;\; & \left(w_r \,-\, w_{r,\star}^{(t)}\right)^\top \nabla_\theta \mathcal{E}_r^{\nu^{(t)}}\big(w_{r,\star}^{(t)};\theta^{(t)}\big) \,+\, \left\|w_r \,-\, w_{r,\star}^{(t)}\right\|_{\Sigma_{\nu^{(t)}}}^2 \\
\geq \;\; & \left\|w_r \,-\, w_{r,\star}^{(t)}\right\|_{\Sigma_{\nu^{(t)}}}^2.
\end{aligned}$$

Taking $w_r = \hat{w}_r^{(t)}$ in the inequality above and combining it with (74) and (73) yield

$$\begin{aligned}
& \mathbb{E}_{s\,\sim\,d_\rho^\star}\mathbb{E}_{a\,\sim\,\pi^\star(\cdot\,|\,s)}\left[\left(w_{r,\star}^{(t)} \,-\, \hat{w}_r^{(t)}\right)^\top \nabla_\theta \log \pi_\theta^{(t)}(a\,|\,s)\right] \\
\leq \;\; & 2\sqrt{\frac{|A|\kappa}{1-\gamma}\left(\mathcal{E}_r^{\nu^{(t)}}\big(\hat{w}_r^{(t)};\theta^{(t)}\big) \,-\, \mathcal{E}_r^{\nu^{(t)}}\big(w_{r,\star}^{(t)};\theta^{(t)}\big)\right)}.
\end{aligned} \tag{75}$$

We now substitute (72) and (75) into the right-hand side of (67) as follows

$$\begin{aligned}
\mathbb{E}\left[\operatorname{err}_r^{(t)}(\pi^\star)\right] \;\;\leq\;\; & 2\sqrt{|A|\,\mathbb{E}\left[\mathcal{E}_r^{d^\star}\big(w_{r,\star}^{(t)};\theta^{(t)}\big)\right]} \,+\, 2\sqrt{\frac{|A|\kappa}{1-\gamma}\mathbb{E}\left[\mathcal{E}_r^{\nu^{(t)}}\big(\hat{w}_r^{(t)};\theta^{(t)}\big) - \mathcal{E}_r^{\nu^{(t)}}\big(w_{r,\star}^{(t)};\theta^{(t)}\big)\right]} \\
\leq\;\; & 2\sqrt{|A|\,\mathbb{E}\left[\mathcal{E}_r^{d^\star}\big(w_{r,\star}^{(t)};\theta^{(t)}\big)\right]} \,+\, 2\sqrt{\frac{|A|\kappa}{1-\gamma}\,\frac{2G^2}{\sigma_F(K+1)}}
\end{aligned}$$

where the second inequality is due to the standard SGD result (Lacoste-Julien et al., 2012): for $\alpha_k = 2/(\sigma_F(k+1))$,

$$\mathcal{E}_{r,\text{est}}^{(t)} \;=\; \mathbb{E}\left[\mathcal{E}_r^{\nu^{(t)}}\big(\hat{w}_r^{(t)};\theta^{(t)}\big) \,-\, \mathcal{E}_r^{\nu^{(t)}}\big(w_{r,\star}^{(t)};\theta^{(t)}\big)\right] \;\leq\; \frac{2G^2}{\sigma_F(K+1)}.$$

By the same reasoning, we can find a similar bound on $\mathbb{E}\big[\operatorname{err}_g^{(t)}(\pi^\star)\big]$. Finally, our desired results follow by applying Assumption 14 and Lemma 32.

References

Felisa Vázquez Abad and Vikram Krishnamurthy. Policy gradient stochastic approximation algorithms for adaptive control of constrained time varying Markov decision processes. In *Proceedings of the 42nd IEEE Conference on Decision and Control*, volume 3, pages 2823–2828, 2003.

Felisa Vázquez Abad, Vikram Krishnamurthy, Katerine Martin, and Irina Baltcheva. Self learning control of constrained Markov chains – a gradient approach. In *Proceedings of the 41st IEEE Conference on Decision and Control*, volume 2, 2002.

Naoki Abe, Prem Melville, Cezar Pendus, Chandan K Reddy, David L Jensen, Vince P Thomas, James J Bennett, Gary F Anderson, Brent R Cooley, Melissa Kowalczyk, et al. Optimizing debt collections using constrained reinforcement learning. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 75–84, 2010.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 22–31, 2017.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Eitan Altman. *Constrained Markov Decision Processes*, volume 7. CRC Press, 1999.

Sanae Amani, Christos Thrampoulidis, and Lin F Yang. Safe reinforcement learning with linear function approximation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 243–253, 2021.

Kenneth J Arrow. *Studies in Linear and Non-linear Programming.* Stanford University Press, 1958.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Proceedings of the 27th Annual Conference on Advances in Neural Information Processing Systems*, pages 773–781, 2013.

Qinbo Bai, Amrit Singh Bedi, Mridul Agarwal, Alec Koppel, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, volume 36, pages 3682–3689, 2022.

Qinbo Bai, Vaneet Aggarwal, and Ather Gattami. Provably sample-efficient model-free algorithm for MDPs with peak constraints. *Journal of Machine Learning Research*, 24(60): 1–25, 2023a.

Qinbo Bai, Amrit Singh Bedi, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via conservative natural policy gradient primal-dual algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6737–6744, 2023b.

Amir Beck. *First-order Methods in Optimization*, volume 25. SIAM, 2017.

Dimitri P Bertsekas. *Nonlinear Programming: Second Edition*. Athena Scientific, 2008.

Dimitri P Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 2014.

Jalaj Bhandari and Daniel Russo. On the linear convergence of policy gradient methods for finite MDPs. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pages 2386–2394, 2021.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *Operations Research*, 2024.

Shalabh 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.

Vivek S Borkar. An actor-critic algorithm for constrained Markov decision processes. *Systems & Control Letters*, 54(3):207–213, 2005.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym. *arXiv preprint arXiv:1606.01540*, 2016.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70 (4):2563–2578, 2022.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A Lyapunov-based approach to safe reinforcement learning. In *Proceedings of the 32nd Annual Conference on Neural Information Processing Systems*, volume 31, 2018.

Dongsheng Ding and Mihailo R Jovanović. Policy gradient primal-dual mirror descent for constrained MDPs with large state spaces. In *Proceedings of the 61st Conference on Decision and Control*, pages 4892–4897, 2022.

Dongsheng Ding, Kaiqing Zhang, Tamer Başar, and Mihailo R Jovanović. Natural policy gradient primal-dual method for constrained Markov decision processes. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, volume 33, pages 8378–8390, 2020.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanović. Provably efficient safe exploration via primal-dual policy optimization. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, pages 3304–3312, 2021.

Dongsheng Ding, Kaiqing Zhang, Tamer Başar, and Mihailo R Jovanović. Convergence and optimality of policy gradient primal-dual method for constrained markov decision processes. In *Proceedings of the 2022 American Control Conference (ACC)*, pages 2851–2856, 2022a.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Alejandro Ribeiro. Last-iterate convergent policy gradient primal-dual methods for constrained MDPs. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, volume 36, pages 66138–66200, 2023.

Dongsheng Ding, Zhengyan Huan, and Alejandro Ribeiro. Resilient constrained reinforcement learning. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pages 3412–3420, 2024.

Yuhao Ding, Junzi Zhang, and Javad Lavaei. On the global optimum convergence of momentum-based policy gradient. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 1910–1934, 2022b.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained MDPs. *arXiv preprint arXiv:2003.02189*, 2020.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online Markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for fisher-non-degenerate policies. In *Proceedings of the 40th International Conference on Machine Learning*, pages 9827–9869, 2023.

Maryam Fazel, Rong Ge, Sham M Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1467–1476, 2018.

Jaime F Fisac, Anayo K Akametalu, Melanie N Zeilinger, Shahab Kaynama, Jeremy Gillula, and Claire J Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, 2018.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Kihyuk Hong, Yuhang Li, and Ambuj Tewari. A primal-dual-critic algorithm for offline constrained reinforcement learning. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pages 280–288, 2024.

Arushi Jain, Sharan Vaswani, Reza Babanezhad, Csaba Szepesvari, and Doina Precup. Towards painless policy optimization for constrained MDPs. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, pages 895–905, 2022.

Jiaming Ji, Jiayi Zhou, Borong Zhang, Juntao Dai, Xuehai Pan, Ruiyang Sun, Weidong Huang, Yiran Geng, Mickel Liu, and Yaodong Yang. Omnisafe: An infrastructure for accelerating safe reinforcement learning research. *Journal of Machine Learning Research*, 25(285):1–6, 2024.

Sham M Kakade. A natural policy gradient. In *Proceedings of the 16th Annual Conference on Neural Information Processing Systems*, pages 1531–1538, 2002.

Sham M Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning*, volume 2, pages 267–274, 2002.

Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On linear and super-linear convergence of natural policy gradient algorithm. *Systems & Control Letters*, 164:105214, 2022.

Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

Tianjiao Li, Ziwei Guan, Shaofeng Zou, Tengyu Xu, Yingbin Liang, and Guanghui Lan. Faster algorithm and sharper analysis for constrained Markov decision process. *Operations Research Letters*, 54:107107, 2024.

Qingkai Liang, Fanyu Que, and Eytan Modiano. Accelerated primal-dual policy optimization for safe reinforcement learning. *arXiv preprint arXiv:1802.06480*, 2018.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Tianyi Lin, Chi Jin, and Michael I Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6083–6093, 2020.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. In *Proceedings of the 33rd Annual Conference on Advances in Neural Information Processing Systems*, pages 10564–10575, 2019.

Tao Liu, Ruida Zhou, Dileep Kalathil, P R Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained MDPs. In *Proceedings of the 35th Annual Conference on Neural Information Processing Systems*, volume 34, pages 17183–17193, 2021a.

Tao Liu, Ruida Zhou, Dileep Kalathil, P R Kumar, and Chao Tian. Policy optimization for constrained MDPs with provable fast global convergence. *arXiv preprint arXiv:2111.00552*, 2021b.

Yanli Liu, Kaiqing Zhang, Tamer Başar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, volume 33, pages 7624–7636, 2020a.

Yongshuai Liu, Jiaxin Ding, and Xin Liu. IPO: Interior-point policy optimization under constraints. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, volume 34, pages 4940–4947, 2020b.

David G Luenberger and Yinyu Ye. *Linear and Nonlinear Programming*, volume 2. Springer, 1984.

Mehrdad Mahdavi, Rong Jin, and Tianbao Yang. Trading regret for efficiency: online convex optimization with long term constraints. *Journal of Machine Learning Research*, 13(Sep): 2503–2528, 2012.

Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter L Bartlett, and Martin J Wainwright. Derivative-free methods for policy optimization: Guarantees for linear-quadratic systems. *Journal of Machine Learning Research*, 51:1–51, 2020.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6820–6829, 2020.

Sobhan Miryoosefi and Chi Jin. A simple reward-free approach to constrained reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, pages 15666–15698, 2022.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1928–1937, 2016.

Hesameddin Mohammadi, Armin Zare, Mahdi Soltanolkotabi, and Mihailo R Jovanović. Global exponential convergence of gradient methods over the nonconvex landscape of the linear quadratic regulator. In *Proceedings of the 58th IEEE Conference on Decision and Control*, pages 7474–7479, Nice, France, 2019.

Hesameddin Mohammadi, Mahdi Soltanolkotabi, and Mihailo R Jovanović. Learning the model-free linear quadratic regulator via random search. In *Proceedings of Machine Learning Research, 2nd Annual Conference on Learning for Dynamics and Control*, volume 120, pages 1–9, Berkeley, CA, 2020.

Hesameddin Mohammadi, Mahdi Soltanolkotabi, and Mihailo R Jovanović. On the linear convergence of random search for discrete-time LQR. *IEEE Control Systems Letters*, 5(3): 989–994, July 2021.

Hesameddin Mohammadi, Armin Zare, Mahdi Soltanolkotabi, and Mihailo R Jovanović. Convergence and sample complexity of gradient methods for the model-free linear-quadratic regulator problem. *IEEE Transactions on Automatic Control*, 67(5):2435–2450, May 2022.

Washim U Mondal and Vaneet Aggarwal. Sample-efficient constrained reinforcement learning with general parameterization. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*, volume 37, pages 68380–68405, 2024.

Alessandro Montenegro, Marco Mussi, Matteo Papini, and Alberto Maria Metelli. Last-iterate global convergence of policy gradients for constrained reinforcement learning. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*, pages 126363–126416, 2024.

Ted Moskovitz, Brendan O'Donoghue, Vivek Veeriah, Sebastian Flennerhag, Satinder Singh, and Tom Zahavy. Reload: Reinforcement learning with optimistic ascent-descent for last-iterate convergence in constrained MDPs. In *Proceedings of the 40th International Conference on Machine Learning*, pages 25303–25336, 2023.

Adrian Müller, Pragnya Alatur, Volkan Cevher, Giorgia Ramponi, and Niao He. Truly no-regret learning in constrained MDPs. In *Proceedings of the 41st International Conference on Machine Learning*, pages 36605–36653, 2024.

Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Proceedings of the 33rd Annual Conference on Advances in Neural Information Processing Systems*, pages 14905–14916, 2019.

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

Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. In *Proceedings of the 33rd Annual Conference on Advances in Neural Information Processing Systems*, pages 7553–7563, 2019.

Santiago Paternain, Miguel Calvo-Fullana, Luiz FO Chamon, and Alejandro Ribeiro. Safe policies for reinforcement learning via primal-dual methods. *IEEE Transactions on Automatic Control*, 2022.

A.B. Piunovskiy. *Optimal Control of Random Sequences in Problems with Constraints*. Springer, 1997.

Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual optimization: Stochastically constrained Markov decision processes with adversarial losses and unknown transitions. In *Proceedings of the 34th Annual Conference on Advances in Neural Information Processing Systems*, 2020.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 7(1):2, 2019.

Sergio Rozada, Dongsheng Ding, Antonio G Marques, and Alejandro Ribeiro. Deterministic policy gradient primal-dual methods for continuous-space constrained MDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20200–20208, 2025.

Harsh Satija, Philip Amortila, and Joelle Pineau. Constrained Markov decision processes via backward value functions. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8502–8511, 2020.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1889–1897, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning*, pages 71–79, 2013.

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, volume 34, pages 5668–5675, 2020.

Rahul Singh, Abhishek Gupta, and Ness B Shroff. Learning in constrained Markov decision processes. *IEEE Transactions on Control of Network Systems*, 10(1):441–453, 2022.

Thomas Spooner and Rahul Savani. A natural actor-critic algorithm with downside risk constraints. *arXiv preprint arXiv:2007.04203*, 2020.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID Lagrangian methods. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9133–9143, 2020.

Yanan Sui, Joel Burdick, Yisong Yue, et al. Stagewise safe Bayesian optimization with Gaussian processes. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4781–4789, 2018.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 14th Annual Conference on Advances in Neural Information Processing Systems*, pages 1057–1063, 2000.

Yasushi Terazono and Ayumu Matani. Continuity of optimal solution functions and their conditions on objective functions. *SIAM Journal on Optimization*, 25(4):2050–2060, 2015.

Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012.

Eiji Uchibe and Kenji Doya. Constrained reinforcement learning from intrinsic and extrinsic rewards. In *Proceedings of the International Conference on Development and Learning*, pages 163–168, 2007.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

Honghao Wei, Xin Liu, and Lei Ying. A provably-efficient model-free algorithm for infinite-horizon average-reward constrained Markov decision processes. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, volume 36, pages 3868–3876, 2022.

Honghao Wei, Xiyue Peng, Arnob Ghosh, and Xin Liu. Adversarially trained weighted actor-critic for safe offline reinforcement learning. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*, pages 52806–52835, 2024.

Xiaohan Wei, Hao Yu, and Michael J Neely. Online primal-dual mirror descent under stochastic constraints. *Proceedings of the International Conference on Measurement and Modeling of Computer Systems*, 2020.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. CRPO: A new approach for safe reinforcement learning with convergence guarantee. In *Proceedings of the 38th International Conference on Machine Learning*, pages 11480–11491, 2021.

Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. In *Proceedings of the 34th Annual Conference on Advances in Neural Information Processing Systems*, 2020a.

Long Yang, Jiaming Ji, Juntao Dai, Linrui Zhang, Binbin Zhou, Pengfei Li, Yaodong Yang, and Gang Pan. Constrained update projection approach to safe policy optimization. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems*, volume 35, pages 9111–9124, 2022.

Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J Ramadge. Projection-based constrained policy optimization. In *Proceedings of the 8th International Conference on Learning Representations*, 2020b.

Donghao Ying, Yuhao Ding, and Javad Lavaei. A dual approach to constrained markov decision processes with entropy regularization. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 1887–1909, 2022.

Hao Yu, Michael Neely, and Xiaohan Wei. Online convex optimization with stochastic constraints. In *Proceedings of the 31st Annual Conference on Advances in Neural Information Processing Systems*, pages 1428–1438, 2017.

Ming Yu, Zhuoran Yang, Mladen Kolar, and Zhaoran Wang. Convergent policy optimization for safe reinforcement learning. In *Proceedings of the 33rd Annual Conference on Advances in Neural Information Processing Systems*, pages 3121–3133, 2019.

Tiancheng Yu, Yi Tian, Jingzhao Zhang, and Suvrit Sra. Provably efficient algorithms for multi-objective competitive RL. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Jianjun Yuan and Andrew Lamperski. Online convex optimization for cumulative constraints. In *Proceedings of the 32nd Annual Conference on Advances in Neural Information Processing Systems*, pages 6137–6146, 2018.

Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, volume 33, pages 4572–4583, 2020a.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020b.

Xiangyuan Zhang, Kaiqing Zhang, Erik Miehling, and Tamer Başar. Non-cooperative inverse reinforcement learning. In *Proceedings of the 33rd Annual Conference on Advances in Neural Information Processing Systems*, pages 9482–9493, 2019.

Yiming Zhang, Quan Vuong, and Keith Ross. First order constrained optimization in policy space. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems*, volume 33, 2020c.

Zhengfei Zhang, Kishan Panaganti, Laixi Shi, Yanan Sui, Adam Wierman, and Yisong Yue. Distributionally robust constrained reinforcement learning under strong duality. In *Proceedings of the Reinforcement Learning Conference*, 2024.

Liyuan Zheng and Lillian J Ratliff. Constrained upper confidence reinforcement learning. In *Proceedings of the 2nd Annual Conference on Learning for Dynamics and Control*, pages 620–629, 2020.