2021 61st IEEE Conference on Decision and Control (CDC)
December 6-9, 2022. Cancún, Mexico

Policy gradient primal-dual mirror descent for constrained MDPs with large state spaces

Dongsheng Ding and Mihailo R. Jovanović

***Abstract*— We study constrained sequential decision-making problems modeled by constrained Markov decision processes with potentially infinite state spaces. We propose a Bregman distance-based direct policy search method – policy gradient primal-dual mirror descent – which includes the natural policy primal-dual method and the projected policy primal-dual method as two special cases. When the exact gradient is known, we prove dimension-free global convergence with a sublinear rate in both optimality gap and constraint violation. When the exact gradient is not available, we instantiate our algorithm in the linear function approximation setting and establish sample complexity guarantees. The introduction of the Bregman-distance regularizers enjoys the dimension-free property with applicability to large-scale spaces, the first of its kind in the constrained RL literature.**

I. Introduction

The constrained Markov decision process (constrained MDP) [1] has become a critical environment model in reinforcement learning (RL) [2], [3]. A popular class of RL methods for constrained MDPs built on the policy gradient (PG) method [4] search policies via gradient descent/ascent or primal/dual type updates (e.g., [5], [6]). More appealing are their generality in PG methods [7]–[9] and effectiveness in using Lagrange method to handle constraints [10], [11]. However, PG method and theory for constrained MDPs with large state spaces is relatively less established from an optimization perspective [12]–[14].

Our contribution: We propose a Bregman distance-based direct policy search method for constrained MDPs with potentially infinite state spaces – policy gradient primal-dual mirror descent – which includes the natural policy primal-dual method and the projected policy primal-dual method as two special cases. When the exact gradient is known, we exploit the structural properties of value functions to prove dimension-free global convergence with sublinear rate $O(1/\sqrt{T})$, regarding the average optimality gap and constraint violation, where T is the number of total iterations. When the exact gradient is not available, we present a sample-based policy gradient primal-dual mirror descent using the linear function approximation and establish sample complexity guarantees. The introduction of the Bregman-distance regularizers enjoys the dimension-free property with applicability to large-scale spaces, the first of its kind in the constrained RL literature.

Financial support from the National Science Foundation under awards ECCS-1708906 and ECCS-1809833.

D. Ding and M. R. Jovanović are with the Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089. E-mails: dongshed@usc.edu, mihailo@usc.edu.

Related work: There has been considerable interest in the development of policy gradient primal-dual methods for constrained MDPs [5], [6], [15]–[17]. The classical performance for these algorithms is the asymptotic convergence to a local stationary point. Several recent works show that policy gradient primal-dual methods [7], [18]–[23] enjoy non-asymptotic convergence that is more preferable in practice. Although these methods are intrinsically related to the mirror descent analysis [24], the classical mirror descent method based general Bregman-distance regularizers [25] have not been utilized. To fill in this gap, we propose a policy gradient primal-dual mirror descent method based on Bregman-distance regularizers that covers the natural policy primal-dual method [7], [18], [20] and a new projected policy primal-dual method. Moreover, our unified analysis does not assume a finite state space and any policy parametrization. Our work is also related to recent works that have significantly advanced learning constrained MDPs with large state spaces using the function approximation [7], [19], [26]–[28]. In contrast, our linear function approximation is more general than the linear constrained MDP assumption [19], [27], [28].

Paper organization: The rest of the paper is organized as follows. We provide background material in Section II, present our method and convergence theory in Section III, establish a model-free method and its sample complexity in Section IV. We conclude the paper in Section V.

II. Preliminaries

We consider a discounted constrained Markov decision process $(S, A, P, r, g, b, \gamma, \rho)$, where S is the state space, A is the action space, P is the transition probability measure which specifies the transition probability $P(s' \,|\, s, a)$ from state s to state s' under action a, r, g: $S \times A \to [0, 1]$ are the reward/utility functions, b is a constraint offset, $\gamma \in [0, 1)$ is the discount factor, and ρ is the initial state distribution.

A stochastic policy is a function π: $S \to \Delta_A$ that determines the action chosen by the agent based on the current state $a_t \sim \pi(\cdot \,|\, s_t)$ at time t, where Δ_A is a probability simplex on A. Let the set of all policies be Π. 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 $s_{t+1} \sim P(\cdot \,|\, s_t, a_t)$.

Let the symbol $\diamond$ be r or g. Given a policy π, the value function $V_\diamond^\pi$: $S \to \mathbb{R}$ is defined as the following expected value of total discounted rewards or utilities

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

where expectation is taken over the randomness of the trajectory τ induced by π. Starting from a state-action pair (s,a), we introduce the state-action value functions $Q_\diamond^\pi(s,a)$: $S\times A\to\mathbb{R}$, and advantage functions $A_\diamond^\pi$: $S\times A\to\mathbb{R}$,

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

The fact that $r,\ g\in[0,1]$ yields $0\le V_\diamond^\pi(s)\le\frac{1}{1-\gamma}$. The expectation over ρ is $V_\diamond^\pi(\rho):=\mathbb{E}_{s_0\sim\rho}[V_\diamond^\pi(s_0)]$. The relation between $V_\diamond^\pi$ and $Q_\diamond^\pi$ is stipulated by Bellman equations [24] and $V_\diamond^\pi(s)=\langle Q_\diamond^\pi(s,\cdot),\pi(\cdot|s)\rangle$. Let the discounted visitation distribution $d_{s_0}^\pi$ and its expectation be

$$d_{s_0}^\pi(s)\;=\;(1-\gamma)\sum_{t=0}^{\infty}\gamma^t P^\pi(s_t=s\,|\,s_0)$$

and $d_\rho^\pi\;=\;\mathbb{E}_{s_0\sim\rho}[d_{s_0}^\pi(s)]$. It is useful to introduce the distribution mismatch coefficient $\kappa:=\sup_\pi\left\|d_\rho^\pi/\rho\right\|_\infty$ that captures exploration difficulty in PG methods [24].

Let $b\in(0,1/(1-\gamma)]$ be the constraint offset. We consider a constrained policy optimization problem that maximizes the reward value function subject to a constraint on the utility value function [1],

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

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

A. Useful Problem Properties

The max-min problem associated with (1) is given by

$$\underset{\pi\,\in\,\Pi}{\text{maximize}}\ \underset{\lambda\,\ge\,0}{\text{minimize}}\ V_L^{\pi,\lambda}(\rho)$$

where $V_L^{\pi,\lambda}(\rho)\;=\;V_r^\pi(\rho)\;+\;\lambda\,(V_g^\pi(\rho)\;-\;b)$ is the Lagrangian. The dual function is defined as $V_D^\lambda(\rho)\;=\;\text{maximize}_\pi\,V_L^{\pi,\lambda}(\rho)$. Let the optimal solution to Problem (1) be $\pi^\star$ and the optimal dual variable be $\lambda^\star\;=\;\arg\min_{\lambda\,\ge\,0}V_D^\lambda(\rho)$. We use shorthand $V_r^{\pi^\star}(\rho)=V_r^\star(\rho)$ and $V_D^{\lambda^\star}(\rho)=V_D^\star(\rho)$ whenever it is clear from the context.

Let $[x]_+=\max(x,0)$. Despite non-convexity [7], strong duality, boundedness of the optimal dual variable, and constraint violation hold; see their proofs in [7], [11].

Lemma 1 (Strong Duality & Bounded $\lambda^\star$): Let Assumption 1 hold. Then, $V_r^\star(\rho)\;=\;V_D^\star(\rho)$, and $0\;\le\;\lambda^\star\;\le\;\left(V_r^\star(\rho)-V_r^{\bar\pi}(\rho)\right)/\xi$.

Lemma 2 (Constraint Violation): Let Assumption 1 hold. If there exists a policy $\pi\in\Pi$ and $C\ge2\lambda^\star$ such that $V_r^\star(\rho)-V_r^\pi(\rho)+C[b-V_g^\pi(\rho)]_+\le\delta$, then $[b-V_g^\pi(\rho)]_+\;\le\;\frac{2\delta}{C}$.

III. Method and Theory

We present a direct policy search method for constrained MDPs and establish convergence when the gradient is exact.

A. Policy Gradient Primal-Dual Mirror Descent

Our Algorithm 1 has two updates. The first one is a policy mirror descent step that solves a proximal policy optimization sub-problem in (2). The second update executes a gradient descent type step for the dual variable in (3).

Algorithm 1 Policy Gradient Primal-Dual Mirror Descent

1: **Initialization**: Stepsizes α and η, number of iterations T, $\pi^0(a\,|\,s)=1/|A|$ for all (s,a), and $\lambda^0=0$.
2: **for** $t=0,1,\dots,T-1$ **do**
3: Define policy $\pi^{t+1}(\cdot\,|\,s)$ for $s\in S$,

$$\begin{aligned}\pi^{t+1}(\cdot\,|\,s):=\underset{\pi(\cdot\,|s)\,\in\,\Delta_A}{\arg\max}\ &\alpha\left\langle Q_r^t(s,\cdot)+\lambda^t Q_g^t(s,\cdot),\pi(\cdot\,|\,s)\right\rangle\\ &-D\big(\pi(\cdot\,|\,s),\pi^t(\cdot\,|\,s)\big)\,.\end{aligned}\tag{2}$$

4: Dual update,

$$\lambda^{t+1}\;=\;\mathcal{P}_\Lambda\left(\lambda^t\;-\;\eta\,(V_g^t(\rho)-b)\right).\tag{3}$$

5: **end for**

In line 3, we form a proximal policy optimization problem as follows. By performance difference lemma [24], we express the Lagrangian $V_L^{\pi,\lambda}(\rho)$ with fixed λ^t at time t as

$$\begin{aligned}&V_L^{\pi,\lambda^t}(\rho)\;=\;V_L^{\pi^t,\lambda^t}(\rho)\\ &+\frac{1}{1-\gamma}\mathbb{E}_{s\,\sim\,d_\rho^\pi}\left[\langle Q_r^{\pi^t}(s,\cdot)+\lambda^t Q_g^{\pi^t}(s,\cdot),\pi(\cdot\,|\,s)-\pi^t(\cdot\,|\,s)\rangle\right].\end{aligned}$$

The policy gradient direction is given by $d_\rho^t(s)Q_L^t(s,a)$ for any (s,a), where $Q_L^t(s,a):=Q_r^t(s,a)+\lambda^t Q_g^t(s,a)$ is the Lagrangian-like function in which we suppress notation π^t in value functions. Mirror descent step with stepsize α reads,

$$\begin{aligned}\pi^{t+1}(\cdot\,|\,s)\;=\;\underset{\pi(\cdot\,|s)\,\in\,\Delta_A}{\arg\max}\ \Big(&\alpha\left\langle d_\rho^t(s)\,Q_L^t(s,\cdot),\pi(\cdot\,|\,s)\right\rangle\\ &-d_\rho^t(s)D\big(\pi(\cdot\,|\,s),\pi^t(\cdot\,|\,s)\big)\Big)\end{aligned}\tag{4}$$

where $D(\pi(\cdot\,|\,s),\pi'(\cdot\,|\,s))$ is the Bregman distance. For any $p,\ p'\in\Delta_A$, the Bregman distance between p and p' is $D(p,p'):=h(p)-h(p')-\langle\nabla h(p'),p-p'\rangle$, where h is strictly convex and continuously differentiable on the interior of Δ_A. By removing $d_\rho^t(s)$, Update (4) is equivalent to (2).

In line 4, we do projected sub-gradient descent of $V_L^{\pi^t,\lambda}(\rho)$ at policy π^t under the same state distribution ρ. By Lemma 1, it suffices to restrict dual iterates in a bounded interval Λ that contains the optimal $\lambda^\star$, e.g., $\Lambda=[\,0,2/((1-\gamma)\xi)\,]$.

Remark 1: The optimality condition for (2) yields two useful cases: (i) when $h(p)=\frac12\|p\|^2$, $D(p,p)=\frac12\|p-p'\|^2$,

$$\pi^{t+1}(\cdot\,|\,s)\;=\;\mathcal{P}_{\Delta_A}\left(\pi^t(\cdot\,|\,s)\;+\;\alpha\left(Q_r^t(s,\cdot)\;+\;\lambda^t Q_g^t(s,\cdot)\right)\right)$$

where $\mathcal{P}_{\Delta_A}(p)\;:=\;\arg\min_{p'\,\in\,\Delta_A}\|p-p'\|$. This update works as the projected Q-descent [14]; (ii) when $h(p)=\sum_{a\,\in\,A}p_a\log p_a$, $D(p,p')=\sum_{a\,\in\,A}p_a\log\frac{p_a}{p'_a}$,

$$\pi^{t+1}(\cdot\,|\,s)\;\propto\;\pi^t(\cdot\,|\,s)\,\mathrm{e}^{\alpha\left(Q_r^t(s,\cdot)+\lambda^t Q_g^t(s,\cdot)\right)}$$

which recovers the multiplicative weight update [7], [18],

[19], [29]. Therefore, Update (2) extends [7], [18] to general Bregman distance-regularizer that subsumes the Euclidean distance case. We note that explicit policy updates above define new policies over $s \in \mathcal{S}$ recursively, and a finite state space is not required at all. It is similar as the unconstrained policy optimization with function approximation [30]–[32].

B. Convergence Analysis

For brevity, we suppress notation π^t in value functions and use shorthand $D(\pi^{t+1}, \pi^t)$ for $D(\pi^{t+1}(\cdot \,|\, s), \pi^t(\cdot \,|\, s))$. The optimality condition for (2) yields Lemma 3 in Appendix A.

Lemma 3: In Algorithm 1, for any $\pi(\cdot \,|\, s) \in \Delta_A$,

$$\alpha \left\langle Q_r^t(s,\cdot) + \lambda^t Q_g^t(s,\cdot), (\pi - \pi^{t+1})(\cdot \,|\, s) \right\rangle + D\big(\pi^{t+1}, \pi^t\big) \leq D(\pi, \pi^t) - D\big(\pi, \pi^{t+1}\big).$$

The performance difference lemma [24] sets up one-step policy performance in Lemma 4; see Appendix B for proof.

Lemma 4: In Algorithm 1, for any $s \in S$,

$$\big(V_r^{t+1}(s) - V_r^t(s)\big) + \lambda^t \big(V_g^{t+1}(s) - V_g^t(s)\big) \geq \frac{1}{\alpha(1-\gamma)} \mathbb{E}_{s' \sim d_s^{t+1}} \big[D\big(\pi^{t+1}, \pi^t\big) + D\big(\pi^t, \pi^{t+1}\big) \big].$$

A key step is to establish the following average performance bound in Lemma 5; see Appendix C for proof.

Lemma 5: In Algorithm 1, for any $T > 0$,

$$\frac{1}{T}\sum_{t=0}^{T-1} \big(V_r^\star(\rho) - V_r^t(\rho)\big) + \frac{1}{T}\sum_{t=0}^{T-1} \lambda^t \big(V_g^\star(\rho) - V_g^t(\rho)\big) \leq \frac{1}{(1-\gamma)^2 T} + \frac{2\eta}{(1-\gamma)^3} + \frac{D_0}{\alpha(1-\gamma)T} \tag{5}$$

where $D_0 := \mathbb{E}_{s \sim d_\rho^\star} \big[D(\pi^\star(\cdot \,|\, s), \pi^0(\cdot \,|\, s)) \big]$.

Lemma 5 gives an upper bound on the average performance on a combination of two gaps, $V_r^\star(\rho) - V_r^t(\rho)$ and $V_g^\star(\rho) - V_g^t(\rho)$. However, this bound does not necessarily imply convergence in the optimality gap, $V_r^\star(\rho) - V_r^t(\rho)$, and the feasibility gap or constraint violation, $b - V_g^t(\rho)$. We next exploit the dual update (3) to bound them. We summarize our bounds in Theorem 6; see Appendix D for proof.

Theorem 6 (Dimension-Free Bound): Let Assumption 1 hold. Suppose $\Lambda = [0, 2/((1-\gamma)\xi)]$. For any $T > 0$, if $\alpha = D_0$ and $\eta = (1-\gamma)/(2\sqrt{T})$ in Algorithm 1, then

$$\frac{1}{T}\sum_{t=0}^{T-1} \big(V_r^\star(\rho) - V_r^t(\rho)\big) \leq \frac{4}{(1-\gamma)^2\sqrt{T}} \tag{6a}$$

$$\left[b - \frac{1}{T}\sum_{t=0}^{T-1} V_g^t(\rho) \right]_+ \leq \frac{4(\xi + 1/\xi)}{(1-\gamma)^2\sqrt{T}}. \tag{6b}$$

In Theorem 6, the average optimality gap/constraint violation decay to zero in $O(1/\sqrt{T})$, where T is the number of total iterations, if we choose stepsizes α and η appropriately. The rate $O(1/\sqrt{T})$ often refers to the sublinear rate in stochastic convex optimization [33], although our constrained problem is non-convex. This dimension-free bound has no dependence on the size of state/action space and the distribution mismatch coefficient, which covers algorithms using KL distance [18] or softmax policy [7] as two special cases.

Theorem 6 demonstrates $O(1/\epsilon^2)$ iteration complexity for yielding an ϵ-optimal policy: we select a policy π^{out} uniformly over iterates $\pi^{(1)}, \ldots, \pi^{(T)}$,

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

IV. Function Approximation

We remove the exact gradient assumption and instantiate Algorithm 1 using function approximation as Algorithm 2.

Assumption 2 (Linear Value Function): There are known feature maps $\phi_\diamond$: $S \times A \to \mathbb{R}^d$ such that for any $(s,a) \in S \times A$ and $\pi \in \Delta_A$, $Q_\diamond^\pi(s,a) = \langle \phi_\diamond(s,a), w_\diamond^\pi \rangle$, where $w_\diamond^\pi \in \mathbb{R}^d$; there also is a known feature map φ_g: $S \to \mathbb{R}^d$ such that for any $s \in S$ and $\pi \in \Delta_A$, $V_g^\pi(s) = \langle \varphi_g(s), u_g^\pi \rangle$, where $u_g^\pi \in \mathbb{R}^d$. Moreover, $\|\phi_r\|$, $\|\phi_g\|$, $\|\varphi_g\| \leq 1$ for all $(s,a) \in S \times A$, and $\|w_r^\pi\|$, $\|w_g^\pi\|$, $\|u_g^\pi\| \leq W$ for all π.

Assumption 2 adopts the linear Q assumption [24]. It is more general than the linear structure in [19], [27], [28].

Algorithm 2 has two stages. In the first stage, we rollout K sample trajectories by executing the policy π^t with h steps and continuing a unifom policy $\text{Unif}_A := \frac{1}{|A|}$ with h' steps, where h and h' follow a geometric distribution $\text{Geo}(1-\gamma)$. In each round k, by collecting rewards from step h to $h+h'-1$, we can justify $\mathbb{E}[R^k] = Q_r^t(s^k, a^k)$ as in [34], where $s^k \sim d_\rho^t$ and $a^k \sim \text{Unif}_A$. This estimation applies to Q_g^t and V_g^t.

Let $\text{LR}(\{(x^k, y^k)\}_{k=1}^K) \approx \arg\min_{\|w\| \leq W} \sum_{k=1}^K (y_k - \langle x^k, w \rangle)^2$ be a near-optimal solution to the empirical linear regression with data set $\{(x^k, y^k)\}_{k=1}^K$, and $\nu^t := d_\rho^t \circ \text{Unif}_A$. In the second stage, we approximate $Q_\diamond^t(\cdot,\cdot)$ and $V_r^t(\rho)$ in line 9 by solving least-square problems in line 8 using K samples. After obtaining estimates $\hat{Q}_\diamond^t(\cdot,\cdot)$ and $\hat{V}_r^t(\rho)$, we perform the policy and dual updates as Algorithm 1. The approximation error is measured by the losses,

$$L_\diamond^t(w_\diamond) := \mathbb{E}_{(s,a) \sim \nu^t} \left[\big(Q_\diamond^t(s,a) - \langle \phi_\diamond(s,a), w_\diamond \rangle\big)^2 \right]$$

$$E_g^t(u_g) := \mathbb{E}_{s \sim \rho} \left[\big(V_g^t(s) - \langle \varphi_g(s), u_g \rangle\big)^2 \right].$$

Assumption 3 (Bounded Statistical Error): For the iterations $\{\hat{w}_r^t, \hat{w}_g^t, \hat{u}_g^t\}_{t=0}^{T-1}$ that are generated by Algorithm 2,

$$\mathbb{E}\left[L_r^t(\hat{w}_r^t)\right], \mathbb{E}\left[L_g^t(\hat{w}_g^t)\right], \mathbb{E}\left[E_g^t(\hat{u}_g^t)\right] \leq \epsilon_{\text{stat}}$$

where expectation is over randomness in $(\hat{w}_r^t, \hat{w}_g^t, \hat{u}_g^t)$.

Theorem 7 (Agnostic Learning): Let Assumption 1 hold. Suppose $\Lambda = [0, 2/((1-\gamma)\xi)]$. For any $T > 0$, if we use $\alpha = \log|A|$ and $\eta = (1-\gamma)/(2\sqrt{T})$ in Algorithm 2, then

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \big(V_r^\star(\rho) - V_r^t(\rho)\big)\right] \lesssim \frac{W^2+1}{(1-\gamma)^2\sqrt{T}} + \frac{\sqrt{\kappa|A|\epsilon_{\text{stat}}}}{(1-\gamma)^{7/2}\xi} \tag{7a}$$

$$\mathbb{E}\left[b - \frac{1}{T}\sum_{t=0}^{T-1} V_g^t(\rho) \right]_+ \lesssim \frac{W^2 + 1/\xi^2}{(1-\gamma)^3\sqrt{T}} + \frac{\sqrt{\kappa|A|\epsilon_{\text{stat}}}}{(1-\gamma)^{5/2}}. \tag{7b}$$

where $\lesssim$ denotes $\leq$ up to an absolute constant.

We sketch the proof of Theorem 7 in Appendix E. Due to Assumption 2, function approximation error appears in Theorem 7 as an additional effect $\sqrt{\epsilon_{\text{stat}}}$. When we apply $\epsilon_{\text{stat}} = O(1/K)$ for K SGD steps [35], the sample complexity reads $TK = O(1/\epsilon^4)$ sample trajectories for obtaining an ϵ-optimal policy. When there is no approximation error: $K \to \infty$, the rate matches the one in Theorem 6.

Algorithm 2 Policy Gradient Primal-Dual Mirror Descent with Linear Function Approximation

1: **Initialization**: Stepsizes α and η, number of iterations T, $\pi^0(a \,|\, s) = 1/|A|$ for all (s, a), $\lambda^0 = 0$, and ρ.
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: **for** round $k = 1, \ldots, K$ **do**
4: Sample $h \sim \text{Geo}(1-\gamma)$ and $h' \sim \text{Geo}(1-\gamma)$.
5: Draw $s_0 \sim \rho$ and collect a trajectory $\{s_0, a_0, r_0, g_0, \ldots, s_H, a_H, r_H, g_H\}$ by executing π^t for first h steps and Unif_A for next h' steps.
6: Define $\bar{s}^k = s_0$, $s^k = s_h$, $a^k = a_h$, $\bar{R}^k = \sum_{i=0}^{h-1} r_i$,

$$R^k \;=\; \sum_{i=h}^{h+h'-1} r_i, \;\text{ and }\; G^k \;=\; \sum_{i=h}^{h+h'-1} g_i.$$

7: **end for**
8: Compute $\hat{w}_r^t$, $\hat{w}_g^t$, and $\hat{u}_g^t$ as

$$\begin{aligned}\hat{w}_r^t \;&=\; \text{LR}(\{(\phi_r(s^k, a^k), R^k)\}_{k=1}^K)\\ \hat{w}_g^t \;&=\; \text{LR}(\{(\phi_g(s^k, a^k), G^k)\}_{k=1}^K)\\ \hat{u}_g^t \;&=\; \text{LR}(\{(\varphi_g(\bar{s}^k), \bar{R}^k)\}_{k=1}^K).\end{aligned}$$

9: Define value functions

$$\hat{Q}_\diamond^t(\cdot,\cdot) := \langle \phi_\diamond(\cdot,\cdot), \hat{w}_\diamond^t \rangle \;\text{ and }\; \hat{V}_g^t(\cdot) := \langle \varphi_g(\cdot), \hat{u}_g^t \rangle.$$

10: Define policy $\pi^{t+1}(\cdot \,|\, s)$ for $s \in S$,

$$\pi^{t+1}(\cdot \,|\, s) = \underset{\pi(\cdot\,|s) \,\in\, \Delta_A}{\arg\max}\; \alpha \big\langle \hat{Q}_r^t(s,\cdot) + \lambda^t \hat{Q}_g^t(s,\cdot), \pi(\cdot \,|\, s) \big\rangle - D\big(\pi(\cdot \,|\, s), \pi^t(\cdot \,|\, s)\big). \tag{8}$$

11: Dual update,

$$\lambda^{t+1} \;=\; \mathcal{P}_\Lambda\big(\lambda^t \,-\, \eta\,(\,\hat{V}_g^t(\rho) - b\,)\big). \tag{9}$$

12: **end for**

V. Conclusion

We have studied a policy gradient primal-dual mirror descent for solving constrained MDPs with potentially infinite state spaces. We prove that the average optimality gap and constraint violation decay to zero in a sublinear rate when the exact gradient is known. When the exact gradient is not available, we present a sample-based algorithm and establish the sample complexity bound. Future directions include improving the rate for single-time scale primal-dual method and tightening sample complexity.

References

[1] E. Altman, *Constrained Markov Decision Processes*. CRC Press, 1999, vol. 7.
[2] G. Dulac-Arnold, D. Mankowitz, and T. Hester, "Challenges of real-world reinforcement learning," *arXiv preprint arXiv:1904.12901*, 2019.
[3] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 1437–1480, 2015.
[4] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An introduction*. MIT press, 2018.
[5] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," in *Proceedings of the International Conference on Learning Representations*, 2019.
[6] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proceedings of the International Conference on Machine Learning*, vol. 70, 2017, pp. 22–31.
[7] D. Ding, K. Zhang, T. Başar, and M. R. Jovanović, "Natural policy gradient primal-dual method for constrained Markov decision processes," in *Proceedings of the Advances in Neural Information Processing Systems*, 2020, pp. 8378–8390.
[8] D. Ding, K. Zhang, T. Başar, and M. R. Jovanović, "Convergence and optimality of policy gradient primal-dual method for constrained Markov decision processes," in *Proceedings of th American Control Conference*, 2022, pp. 2851–2856.
[9] D. Ding, K. Zhang, J. Duan, T. Başar, and M. R. Jovanović, "Convergence and sample complexity of natural policy gradient primal-dual methods for constrained MDPs," *arXiv preprint arXiv:2206.02346*, 2022.
[10] S. Paternain, M. Calvo-Fullana, L. F. Chamon, and A. Ribeiro, "Safe policies for reinforcement learning via primal-dual methods," *IEEE Trans. Autom. Control*, 2022.
[11] S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro, "Constrained reinforcement learning has zero duality gap," in *Proceedings of the Advances in Neural Information Processing Systems*, 2019, pp. 7553–7563.
[12] G. Lan, "Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes," *Math. Program.*, pp. 1–48, 2022.
[13] W. Zhan, S. Cen, B. Huang, Y. Chen, J. D. Lee, and Y. Chi, "Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence," *arXiv preprint arXiv:2105.11066*, 2021.
[14] L. Xiao, "On the convergence rates of policy gradient methods," *arXiv preprint arXiv:2201.07443*, 2022.
[15] Q. Liang, F. Que, and E. Modiano, "Accelerated primal-dual policy optimization for safe reinforcement learning," *arXiv preprint arXiv:1802.06480*, 2018.
[16] A. Stooke, J. Achiam, and P. Abbeel, "Responsive safety in reinforcement learning by PID Lagrangian methods," in *Proceedings of the International Conference on Machine Learning*, 2020, pp. 9133–9143.
[17] M. Yu, Z. Yang, M. Kolar, and Z. Wang, "Convergent policy optimization for safe reinforcement learning," in *Proceedings of the Advances in Neural Information Processing Systems*, 2019, pp. 3121–3133.
[18] Y. Chen, J. Dong, and Z. Wang, "A primal-dual approach to constrained Markov decision processes," *arXiv preprint arXiv:2101.10895*, 2021.
[19] D. Ding, X. Wei, Z. Yang, Z. Wang, and M. R. Jovanović, "Provably efficient safe exploration via primal-dual policy optimization," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, vol. 130, 2021, pp. 3304–3312.
[20] S. Zeng, T. T. Doan, and J. Romberg, "Finite-time complexity of online primal-dual natural actor-critic algorithm for constrained Markov decision processes," *arXiv preprint arXiv:2110.11383*, 2021.
[21] D. Ying, Y. Ding, and J. Lavaei, "A dual approach to constrained Markov decision processes with entropy regularization," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2022, pp. 1887–1909.
[22] T. Liu, R. Zhou, D. Kalathil, P. Kumar, and C. Tian, "Fast global convergence of policy optimization for constrained MDPs," *arXiv preprint arXiv:2111.00552*, 2021.
[23] Q. Bai, A. S. Bedi, M. Agarwal, A. Koppel, and V. Aggarwal, "Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 4, 2022, pp. 3682–3689.

[24] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "On the theory of policy gradient methods: Optimality, approximation, and distribution shift." *J. Mach. Learn. Res.*, vol. 22, no. 98, pp. 1–76, 2021.
[25] C. Blair, "Problem complexity and method efficiency in optimization," *SIAM Rev.*, vol. 27, no. 2, p. 264, 1985.
[26] T. Xu, Y. Liang, and G. Lan, "CRPO: A new approach for safe reinforcement learning with convergence guarantee," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 11 480–11 491.
[27] S. Amani, C. Thrampoulidis, and L. Yang, "Safe reinforcement learning with linear function approximation," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 243–253.
[28] S. Miryoosefi and C. Jin, "A simple reward-free approach to constrained reinforcement learning," in *Proceedings of the International Conference on Machine Learning*, 2022, pp. 15 666–15 698.
[29] Y. Efroni, S. Mannor, and M. Pirotta, "Exploration-exploitation in constrained MDPs," *arXiv preprint arXiv:2003.02189*, 2020.
[30] Q. Cai, Z. Yang, C. Jin, and Z. Wang, "Provably efficient exploration in policy optimization," in *Proceedings of the International Conference on Machine Learning*, 2020, pp. 1283–1294.
[31] H. Luo, C.-Y. Wei, and C.-W. Lee, "Policy optimization in adversarial MDPs: Improved exploration via dilated bonuses," in *Proceedings of the Advances in Neural Information Processing Systems*, 2021, pp. 22 931–22 942.
[32] D. Ding, C.-Y. Wei, K. Zhang, and M. R. Jovanović, "Independent policy gradient for large-scale Markov potential games: Sharper rates, function approximation, and game-agnostic convergence," in *Proceedings of the International Conference on Machine Learning*, 2022, pp. 5166–5220.
[33] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM J. Optim.*, vol. 19, no. 4, pp. 1574–1609, 2009.
[34] S. Paternain, "Stochastic control foundations of autonomous behavior," Ph.D. dissertation, University of Pennsylvania, 2018.
[35] K. Cohen, A. Nedić, and R. Srikant, "On projected stochastic gradient descent algorithm with weighted averaging for least squares regression," *IEEE Trans. Autom. Control*, vol. 62, no. 11, pp. 5974–5981, 2017.

APPENDIX

A. Proof of Lemma 3

By the optimality condition for (2), for any $\pi(\cdot \,|\, s)$,

$$\left\langle \alpha Q_L^t(s,\cdot) - \nabla D\big(\pi^{t+1},\pi^t\big), (\pi - \pi^{t+1})(\cdot \,|\, s)\right\rangle \;\le\; 0 \quad (10)$$

$$\nabla D\big(\pi^{t+1},\pi^t\big) := \nabla_\pi D\big(\pi(\cdot \,|\, s), \pi^t(\cdot \,|\, s)\big)\,\big|_{\pi(\cdot \,|\, s)=\pi^{t+1}(\cdot \,|\, s)}.$$

By the Bregman distance,

$$\begin{aligned} D(\pi,\pi^t) \;&=\; D\big(\pi^{t+1},\pi^t\big) \\ &+ \left\langle \nabla D\big(\pi^{t+1},\pi^t\big), (\pi-\pi^{t+1})(\cdot \,|\, s)\right\rangle \;+\; D\big(\pi,\pi^{t+1}\big). \end{aligned} \quad (11)$$

Combining (11) with (10) completes the proof.

B. Proof of Lemma 4

Applying performance difference lemma [24] to $V_\diamond^{t+1}(s) - V_\diamond^t(s)$ and adding them with $\lambda^t \ge 0$ yield,

$$\begin{aligned} &\big(V_r^{t+1}(s) - V_r^t(s)\big) + \lambda^t\left(V_g^{t+1}(s) - V_g^t(s)\right) \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^{t+1}}\left[\left\langle Q_L^t(s',\cdot), (\pi^{t+1}-\pi^t)(\cdot \,|\, s')\right\rangle\right] \\ &\ge \frac{1}{\alpha(1-\gamma)}\mathbb{E}_{s'\sim d_s^{t+1}}\left[D\big(\pi^{t+1},\pi^t\big) + D\big(\pi^t,\pi^{t+1}\big)\right] \end{aligned}$$

where we use $\pi = \pi^t$ in Lemma 3 for the inequality.

C. Proof of Lemma 5

We repeat the first step of proving Lemma 4. By $\lambda^t \ge 0$,

$$\begin{aligned} &\big(V_r^\star(s) - V_r^t(s)\big) + \lambda^t\left(V_g^\star(s) - V_g^t(s)\right) \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^\star}\left[\left\langle Q_L^t(s',\cdot), (\pi^\star-\pi^{t+1})(\cdot \,|\, s')\right\rangle\right] \\ &\quad + \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_s^\star}\left[\left\langle Q_L^t(s',\cdot), (\pi^{t+1}-\pi^t)(\cdot \,|\, s')\right\rangle\right]. \end{aligned} \quad (12)$$

If we apply Lemma 3 with $\pi = \pi^\star$, then,

$$\begin{aligned} &\text{LHS of (12)} + \frac{1}{\alpha(1-\gamma)}\mathbb{E}_{s'\sim d_\rho^\star}\left[D\big(\pi^{t+1},\pi^t\big)\right] \\ &\le \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_\rho^\star}\left[\left\langle Q_L^t(s',\cdot), (\pi^{t+1}-\pi^t)(\cdot \,|\, s')\right\rangle\right] \\ &\quad + \frac{1}{\alpha(1-\gamma)}\mathbb{E}_{s'\sim d_\rho^\star}\left[D(\pi^\star,\pi^t) - D\big(\pi^\star,\pi^{t+1}\big)\right]. \end{aligned} \quad (13)$$

On the other hand, by Lemma 3 with $\pi = \pi^t$, for any s,

$$\left\langle Q_L^t(s,\cdot), \pi^{t+1}(\cdot \,|\, s) - \pi^t(\cdot \,|\, s)\right\rangle \;\ge\; 0.$$

Thus,

$$\begin{aligned} &\mathbb{E}_{s'\sim d_\rho^\star}\left[\left\langle Q_L^t(s',\cdot), \pi^{t+1}(\cdot \,|\, s') - \pi^t(\cdot \,|\, s')\right\rangle\right] \\ &= \sum_{s'}\frac{d_\rho^\star(s')}{d_{d_\rho^\star}^{t+1}(s')}d_{d_\rho^\star}^{t+1}(s')\left[\left\langle Q_L^t(s',\cdot), (\pi^{t+1}-\pi^t)(\cdot \,|\, s')\right\rangle\right] \\ &\le \frac{1}{1-\gamma}\mathbb{E}_{s'\sim d_{d_\rho^\star}^{t+1}}\left[\left\langle Q_L^t(s',\cdot), (\pi^{t+1}-\pi^t)(\cdot \,|\, s')\right\rangle\right] \\ &= \big(V_r^{t+1}(d_\rho^\star) - V_r^t(d_\rho^\star)\big) + \lambda^t\left(V_g^{t+1}(d_\rho^\star) - V_g^t(d_\rho^\star)\right) \end{aligned} \quad (14)$$

where the inequality is due to $d_{d_\rho^\star}^{t+1} \ge (1-\gamma)d_\rho^\star$. Application of (14) to RHS of (13) yields another upper bound,

$$\begin{aligned} &\frac{1}{1-\gamma}\left(\big(V_r^{t+1}(d_\rho^\star) - V_r^t(d_\rho^\star)\big) + \lambda^t\left(V_g^{t+1}(d_\rho^\star) - V_g^t(d_\rho^\star)\right)\right) \\ &+ \frac{1}{\alpha(1-\gamma)}\mathbb{E}_{s'\sim d_\rho^\star}\left[D(\pi^\star,\pi^t) - D\big(\pi^\star,\pi^{t+1}\big)\right] \end{aligned}$$

which, when we sum it from $t=0$ to $t=T-1$ and ignore terms that do not affect the inequality direction, leads to

$$\begin{aligned} &\sum_{t=0}^{T-1}\big(V_r^\star(\rho) - V_r^t(\rho)\big) + \sum_{t=0}^{T-1}\lambda^t\left(V_g^\star(\rho) - V_g^t(\rho)\right) \\ &\le \frac{1}{1-\gamma}\left(V_r^T(d_\rho^\star) + \sum_{t=0}^{T-1}\lambda^t\left(V_g^{t+1}(d_\rho^\star) - V_g^t(d_\rho^\star)\right)\right) \\ &\quad + \frac{1}{\alpha(1-\gamma)}\mathbb{E}_{s'\sim d_\rho^\star}D\big(\pi^\star,\pi^0\big). \end{aligned} \quad (15)$$

We note that $\lambda^0 = 0$, $\lambda^T = \sum_{t=0}^{T-1}\big(\lambda^{t+1} - \lambda^t\big)$, and $V_g^t \le \frac{1}{1-\gamma}$. Thanks to (3), $|\lambda^t - \lambda^{t+1}| \le \frac{\eta}{1-\gamma}$ and $\lambda^T \le \frac{\eta T}{1-\gamma}$. Thus,

$$\begin{aligned} &\sum_{t=0}^{T-1}\lambda^t\left(V_g^{t+1}(d_\rho^\star) - V_g^t(d_\rho^\star)\right) \\ &\le \lambda^T V_g^T(d_\rho^\star) + \sum_{t=0}^{T-1}\left|\lambda^t - \lambda^{t+1}\right| V_g^{t+1}(d_\rho^\star) \;\le\; \frac{2\eta T}{(1-\gamma)^2}. \end{aligned} \quad (16)$$

Application of (16) to RHS of (15) completes the proof.

D. Proof of Theorem 6

We first bound the optimality gap. Since $\lambda^0 = 0$, $(\lambda^T)^2 = \sum_{t=0}^{T-1} \big((\lambda^{t+1})^2 - (\lambda^t)^2\big)$. By the dual update (3),

$$(\lambda^T)^2 \le 2\eta \sum_{t=0}^{T-1} \lambda^t \big(V_g^\star(\rho) - V_g^t(\rho)\big) + \frac{\eta^2 T}{(1-\gamma)^2}$$

where we use $V_g^\star(\rho) \ge b$ and $|V_g^t(\rho) - b| \le \frac{1}{1-\gamma}$. Thus,

$$-\frac{1}{T}\sum_{t=0}^{T-1} \lambda^t (V_g^\star(\rho) - V_g^t(\rho)) \;\le\; \frac{\eta}{2(1-\gamma)^2}. \tag{17}$$

If we add (17) to the inequality (5) from both sides, then

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} \big(V_r^\star(\rho) - V_r^t(\rho)\big) \\
&\le\; \frac{1}{(1-\gamma)^2}\left(\frac{1}{T} + \frac{2\eta}{1-\gamma}\right) + \frac{D_0}{\alpha(1-\gamma)T} + \frac{\eta}{2(1-\gamma)^2}
\end{aligned}$$

which gives the first bound by taking $\alpha = D_0$, and $\eta = \frac{1-\gamma}{2\sqrt{T}}$.

We next bound the constraint violation. By (3), for any $\lambda \in \Lambda = [0, 2/((1-\gamma)\xi)]$, $(\lambda^{t+1} - \lambda)^2$ is upper bounded by

$$(\lambda^t - \lambda)^2 \;-\; 2\eta(V_g^t(\rho) - b)(\lambda^t - \lambda) + \frac{\eta^2}{(1-\gamma)^2}$$

where we use the non-expansiveness of projection and $|V_g^t(\rho) - b| \le \frac{1}{1-\gamma}$. If we sum both sides of the above inequality over $t = 0, \cdots, T-1$ and divide it by T, then

$$\frac{1}{T}\sum_{t=0}^{T-1} (V_g^t(\rho) - b)(\lambda^t - \lambda) \;\le\; \frac{\lambda^2}{2\eta T} + \frac{\eta}{2(1-\gamma)^2}. \tag{18}$$

Adding (18) to the inequality (5) from both sides yields,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} \big(V_r^\star(\rho) - V_r^t(\rho)\big) + \frac{\lambda}{T}\sum_{t=0}^{T-1} \big(b - V_g^t(\rho)\big) \\
&\le\; \frac{1}{(1-\gamma)^2}\left(\frac{1}{T} + \frac{3\eta}{1-\gamma}\right) + \frac{D_0}{\alpha(1-\gamma)T} + \frac{\lambda^2}{2\eta T}.
\end{aligned} \tag{19}$$

We simplify RHS of (19) by taking $\alpha = D_0$, and $\eta = \frac{1-\gamma}{2\sqrt{T}}$,

$$\frac{2}{(1-\gamma)^2\sqrt{T}} + \frac{1}{(1-\gamma)T} + \frac{\lambda^2}{(1-\gamma)\sqrt{T}} + \frac{1}{4(1-\gamma)\sqrt{T}}.$$

We note that $V_r^t(\rho)$ and $V_g^t(\rho)$ are linear in the occupancy measure for π^t. By convexity of the occupancy measure set [1], $\frac{1}{T}\sum_{t=0}^{T-1} V_r^t(\rho)$ and $\frac{1}{T}\sum_{t=0}^{T-1} V_g^t(\rho)$ are linear in an occupancy measure induced by some policy π'. We choose $\lambda = \frac{2}{(1-\gamma)\xi}$ if $b - V_g^{\pi'}(\rho) \ge 0$; otherwise $\lambda = 0$. Hence, after some rearrangements, (19) becomes

$$\begin{aligned}
&\Big(V_r^\star(\rho) - V_r^{\pi'}(\rho)\Big) + \frac{2}{(1-\gamma)\xi}\Big[b - V_g^{\pi'}(\rho)\Big]_+ \\
&\le\; \frac{2}{(1-\gamma)^2\sqrt{T}} + \frac{2}{(1-\gamma)\sqrt{T}} + \frac{4}{(1-\gamma)^3\xi^2\sqrt{T}}.
\end{aligned}$$

Since $\frac{2}{(1-\gamma)\xi} \ge 2\lambda^\star$, application of Lemma 2 yields,

$$\Big[b - V_g^{\pi'}(\rho)\Big]_+ \;\le\; \frac{4\xi}{(1-\gamma)\sqrt{T}} + \frac{4}{(1-\gamma)^2\xi\sqrt{T}}$$

which leads to the second bound if we use $0 < \xi \le \frac{1}{1-\gamma}$.

E. Proof of Theorem 7

Since the idea is similar as proving Theorem 6, we only sketch some key steps. First, we employ techniques for proving Lemma 5 to show the average performance,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \big(V_r^\star(\rho) - V_r^t(\rho)\big) + \lambda^t \big(V_g^\star(\rho) - V_g^t(\rho)\big)\right] \\
&\le\; \frac{1}{(1-\gamma)^2 T} + \frac{2\eta(W+1)}{(1-\gamma)^3} + \frac{3(1+1/\xi)}{(1-\gamma)^2}\sqrt{\frac{\kappa|A|}{1-\gamma}\epsilon_{\text{stat}}} \\
&\quad + \frac{\log|A|}{\alpha(1-\gamma)T}.
\end{aligned} \tag{20}$$

Differing from Lemma 5, the estimation error enters as,

$$\begin{aligned}
&\mathbb{E}_{s' \sim d_\rho^\star}\big[\big\langle Q_r^t(s',\cdot) - \hat{Q}_r^t(s',\cdot), (\pi^\star - \pi^{t+1})(\cdot \,|\, s')\big\rangle\big] \\
&\le\; \sqrt{\sum_s d_\rho^\star(s)\big\langle Q_r^t(s',\cdot) - \hat{Q}_r^t(s',\cdot), (\pi^\star - \pi^{t+1})(\cdot \,|\, s')\big\rangle^2} \\
&\le\; \sqrt{\frac{\kappa|A|}{1-\gamma}\sum_{s',a'} d_\rho^t(s)\frac{1}{|A|}\big(Q_r^t(s',\cdot) - \hat{Q}_r^t(s',a')\big)^2} \\
&\le\; \sqrt{\frac{\kappa|A|}{1-\gamma}\epsilon_{\text{stat}}}.
\end{aligned}$$

Additionally, $|\mathbb{E}[\lambda^t] - \mathbb{E}[\lambda^{t+1}]| \le \eta(W + \frac{1}{1-\gamma})$ and $\mathbb{E}[\lambda^T] \le \eta T(W + \frac{1}{1-\gamma})$ that result from the dual update (9).

The next proof is similar to proving Theorem 6. Let $\sigma := W^2 + \frac{1}{(1-\gamma)^2}$. The first step is similar to (17),

$$-\,\mathbb{E}\,[\,\text{LHS of (17)}\,] \;\le\; \frac{2\sqrt{\epsilon_{\text{stat}}}}{(1-\gamma)\xi} + \eta\sigma \tag{21}$$

where we use $\mathbb{E}\big(\hat{V}_g^t(\rho) - b\big)^2 \le 2\sigma$, and Assumption 3,

$$\Big|\hat{V}_g^t(\rho) - V_g^t(\rho)\Big| \;\le\; \sqrt{\sum_s \rho(s)(\hat{V}_g^t(s) - V_g^t(s))^2} \;\le\; \sqrt{\epsilon_{\text{stat}}}.$$

If we add (21) to the inequality (20) from both sides, and take $\alpha = \log|A|$ and $\eta = \frac{1-\gamma}{2\sqrt{T}}$, we obtain the first bound. The second step is to apply the same reasoning of proving (19),

$$\mathbb{E}\,[\,\text{LHS of (18)}\,] \;\le\; \frac{\mathbb{E}[(\lambda^0 - \lambda)^2]}{2\eta T} + \frac{\sqrt{\epsilon_{\text{stat}}}}{(1-\gamma)\xi} + \eta\sigma. \tag{22}$$

Adding (22) to the inequality (20) from both sides yields,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \big(V_r^\star(\rho) - V_r^t(\rho)\big) + \frac{\lambda}{T}\big(b - V_g^t(\rho)\big)\right] \\
&\le\; \frac{1}{(1-\gamma)^2 T} + \frac{2\eta(W+1)}{(1-\gamma)^3} + \frac{3(1+1/\xi)}{(1-\gamma)^2}\sqrt{\frac{\kappa|A|}{1-\gamma}\epsilon_{\text{stat}}} \\
&\quad + \frac{\log|A|}{\alpha(1-\gamma)T} + \frac{\lambda^2}{2\eta T} + \frac{\sqrt{\epsilon_{\text{stat}}}}{(1-\gamma)\xi} + \eta\sigma.
\end{aligned} \tag{23}$$

Finally, simplifying RHS of (23) with $\alpha = \log|A|$ and $\eta = \frac{1-\gamma}{2\sqrt{T}}$ and application of Lemma 2 lead to the second bound.