Beyond Exact Gradients: Convergence of Stochastic Soft-Max Policy Gradient Methods with Entropy Regularization

Yuhao Ding
University of California, Berkeley
yuhao_ding@berkeley.edu

Junzi Zhang [1]
Amazon Advertising
junziz@amazon.com

Javad Lavaei
University of California, Berkeley
lavaei@berkeley.edu

Abstract

Entropy regularization is an efficient technique for encouraging exploration and preventing a premature convergence of (vanilla) policy gradient methods in reinforcement learning (RL). However, the theoretical understanding of entropy regularized RL algorithms has been limited. In this paper, we revisit the classical entropy regularized policy gradient methods with the soft-max policy parametrization, whose convergence has so far only been established assuming access to exact gradient oracles. To go beyond this scenario, we propose the first set of (nearly) unbiased stochastic policy gradient estimators with trajectory-level entropy regularization, with one being an unbiased visitation measure-based estimator and the other one being a nearly unbiased yet more practical trajectory-based estimator. We prove that although the estimators themselves are unbounded in general due to the additional logarithmic policy rewards introduced by the entropy term, the variances are uniformly bounded. This enables the development of the first set of convergence results for stochastic entropy regularized policy gradient methods to both stationary points and globally optimal policies. We also develop some improved sample complexity results under a good initialization.

1 Introduction

Entropy regularization is a popular technique to encourage exploration and prevent premature convergence for reinforcement learning (RL) algorithms. It was originally proposed in Williams and Peng (1991) to improve the performance of REINFORCE, a classical family of vanilla policy gradient methods widely used in practice. Since then, the entropy regularization technique has been applied to a large set of other RL algorithms including actor-critic (Mnih et al., 2016; Haarnoja et al., 2018), Q-learning (O'Donoghue et al., 2016; Haarnoja et al., 2017) and trust-region policy optimization methods (Zang et al., 2020). It has also been demonstrated to work well with deep learning approximations to achieve an impressive empirical performance boost. Nevertheless, the theoretical understanding of the convergence of these algorithms has been rather limited and mostly restricted to the exact gradient setting.

In this paper, we revisit the classical entropy regularized (vanilla) policy gradient (PG) methods proposed in the seminal work Williams and Peng (1991) under the soft-max policy parametrization. We focus on the modern trajectory-level entropy regularization proposed in Haarnoja et al. (2017), which is shown to improve over the original one-step entropy regularization adopted in Williams and Peng (1991); Mnih et al. (2016) and O'Donoghue et al. (2016). The literature on the convergence analysis of such algorithms is extremely limited. The work Mei et al. (2020) has recently developed the first set of global convergence results, which is focused on the soft-max policy parametrization assuming access to exact policy gradient evaluations. It remains open whether convergence results can still be obtained in the practical stochastic gradient setting with an arbitrary initial point, for which there is a potentially unbounded logarithmic policy reward component introduced by the entropy term.

The remainder of the paper provides an affirmative answer to the above question. We begin by proposing two new entropy regularized stochastic policy gradient estimators. The first one is an unbiased visitation measure-based estimator, whereas the second one is a nearly unbiased yet more practical trajectory-based estimator. These (nearly) unbiased stochastic policy gradient estimators are the first estimators in the literature with a *trajectory-level* entropy regularization. We show that although the estimators themselves are

[1] Work done prior to joining or outside of Amazon.

unbounded in general due to the entropy-induced logarithmic policy rewards, the variances indeed remain uniformly bounded. This enables the development of the first set of convergence results[2] for stochastic entropy-regularized policy gradient methods, both to stationary points and to globally optimal policies. A discussion about improved sample complexity results with a good initialization is also provided.

Due to the space restriction, we leave the more detailed literature review and notation to Sections 6 and 7 in appendix.

2 Preliminaries

Markov decision processes. Reinforcement learning is generally modeled as a discounted Markov decision process (MDP) defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma)$. Here, $\mathcal{S}$ and $\mathcal{A}$ denote the finite state and action spaces; $\mathbb{P}(s'|s,a)$ is the probability that the agent transits from the state s to the state s' under the action $a \in \mathcal{A}$; $r(s,a)$ is the reward function, i.e., the agent obtains the reward $r(s_h, a_h)$ after it takes the action a_h at the state s_h at time h; $\gamma \in (0,1)$ is the discount factor. Without loss of generality, we assume that $r(s,a) \in [0, \bar{r}]$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. The policy $\pi(a|s)$ at the state s is usually represented by a conditional probability distribution $\pi_\theta(a|s)$ associated to the parameter $\theta \in \mathbb{R}^d$. Let $\tau = \{s_0, a_0, s_1, a_1, \ldots\}$ denote the data of a sampled trajectory under policy π_θ with the probability distribution over the trajectory as

$$p(\tau|\theta, \rho) := \rho(s_0) \prod_{h=1}^{\infty} \mathbb{P}(s_{h+1}|s_h, a_h)\pi_\theta(a_h|s_h), \qquad (1)$$

where $\rho \in \Delta(\mathcal{S})$ is the probability distribution of the initial state s_0. Here, $\Delta(\mathcal{X})$ denotes the probability simplex over a finite set $\mathcal{X}$.

Value functions and Q-functions. Given a policy π, one can define the state-action value function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as

$$Q^\pi(s,a) := \mathbb{E}_{\substack{a_h \sim \pi(\cdot|s_h) \\ s_{h+1} \sim \mathbb{P}(\cdot|s_h, a_h)}} \left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \middle| s_0 = s, a_0 = a \right].$$

The state-value function $V^\pi : \mathcal{S} \to \mathbb{R}$ and the advantage function $A^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ can be defined as

$$V^\pi(s) := \mathbb{E}_{a \sim \pi(\cdot|s)}[Q^\pi(s,a)],$$
$$A^\pi(s,a) := Q^\pi(s,a) - V^\pi(s).$$

The goal is to find an optimal policy in the underlying policy class that maximizes the expected discounted return, namely,

$$\max_{\theta \in \mathbb{R}^d} \quad V^{\pi_\theta}(\rho) := \mathbb{E}_{s_0 \sim \rho}[V^{\pi_\theta}(s_0)]. \qquad (2)$$

For notional convenience, we will denote $V^{\pi_\theta}(\rho)$ by the shorthand notation $V^\theta(\rho)$.

Exploratory initial distribution. The discounted state visitation distribution $d^\pi_{s_0}$ is defined as

$$d^\pi_{s_0}(s) := (1-\gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}(s_h = s|s_0, \pi), \qquad (3)$$

where $\mathbb{P}(s_h = s|s_0, \pi)$ is the state visitation probability that s_h is equal to s under the policy π starting from the state s_0. The discounted state visitation distribution under the initial distribution ρ is defined as $d^\pi_\rho(s) := \mathbb{E}_{s_0 \sim \rho}[d^\pi_{s_0}(s)]$. Furthermore, the state-action visitation distribution induced by π and the initial state distribution ρ is defined as $v^\pi_\rho(s,a) := d^\pi_\rho(s)\pi(a|s)$, which can also be written as

$$v^\pi_\rho(s,a) := (1-\gamma)\mathbb{E}_{s_0 \sim \rho} \sum_{h=0}^{\infty} \gamma^h \mathbb{P}(s_h = s, a_h = a|s_0, \pi),$$

where $\mathbb{P}(s_h = s, a_h = a|s_0, \pi)$ is the state-action visitation probability that $s_h = s$ and $a_h = a$ under π starting from the state s_0. To facilitate the presentation of the main results of the paper, we assume that the state distribution ρ for the performance measure is exploratory (Mei et al., 2020; Bhandari and Russo, 2019), i.e., $\rho(\cdot)$ adequately covers the entire state distribution:

Assumption 2.1 *The state distribution ρ satisfies $\rho(s) > 0$ for all $s \in \mathcal{S}$.*

In practice, when the above assumption is not satisfied, we can optimize under another initial distribution μ, i.e., the gradient is taken with respect to the optimization measure μ, where μ is usually chosen as an exploratory initial distribution that adequately covers the state distribution of some optimal policy. It is shown in Agarwal et al. (2019) that the difficulty of the exploration problem faced by policy gradient algorithms can be captured through the distribution mismatch coefficient defined as $\left\| \frac{d^\pi_\rho}{\mu} \right\|_\infty$, where $\frac{d^\pi_\rho}{\mu}$ denotes component-wise division.

Soft-max policy parameterization. In this work, we consider the soft-max parameterization – a widely adopted scheme that naturally ensures that the policy lies in the probability simplex. Specifically, for an unconstrained parameter $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\pi_\theta(a|s)$ is chosen to be

$$\frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}.$$

The soft-max parameterization is generally used for MDPs with finite state and action spaces. It is complete in the sense that every stochastic policy can be

[2] Note that our main focus here is on studying global convergence, instead of achieving tight sample complexity results, which we leave as future work.

represented by this class. For the soft-max parameterization, it can be shown that the gradient and Hessian of the function $\log \pi_\theta(a|s)$ are bounded, i.e., for all $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we have:

$$\|\nabla \log \pi_\theta(a|s)\|_2 \le 2, \ \left\|\nabla^2 \log \pi_\theta(a|s)\right\|_2 \le 1.$$

Reinforcement learning with entropy regularization. Entropy is a commonly used regularization in RL to promote exploration and discourage premature convergence to suboptimal policies (Haarnoja et al., 2017; Schulman et al., 2017; Eysenbach and Levine, 2019). It is far less aggressive in penalizing small probabilities, in comparison to other common regularizations such as log barrier functions (Agarwal et al., 2019). In the entropy-regularized RL (also known as maximum entropy RL), near-deterministic policies are penalized, which is achieved by modifying the value function to

$$V_\lambda^\pi(\rho) = V^\pi(\rho) + \lambda \mathbb{H}(\rho, \pi), \tag{4}$$

where $\lambda \ge 0$ determines the strength of the penalty and $\mathbb{H}(\rho, \pi)$ stands for the discounted entropy defined as

$$\mathbb{H}(\rho, \pi) \coloneqq \mathbb{E}_{\substack{s_0 \sim \rho, a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t|s_t) \right].$$

Equivalently, $V_\lambda^\pi(\rho)$ can be viewed as the weighted value function of π by adjusting the instantaneous reward to be policy-dependent regularized version as

$$r^\lambda(s,a) \coloneqq r(s,a) - \lambda \log \pi(a|s), \ \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

We also define $V_\lambda^\pi(s)$ analogously when the initial state is fixed at a given state $s \in \mathcal{S}$. The regularized Q-function Q_λ^π of a policy π, also known as the soft Q-function, is related to V_λ^π as (for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$)

$$\begin{aligned} Q_\lambda^\pi(s,a) &= r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[V_\lambda^\pi(s') \right], \\ V_\lambda^\pi(s) &= \mathbb{E}_{a \sim \pi(\cdot|s)} \left[-\lambda \log \pi(a \mid s) + Q_\lambda^\pi(s,a) \right]. \end{aligned}$$

Bias due to entropy regularization. Due to the presence of regularization, the optimal solution will be biased with the bias disappearing as $\lambda \to 0$. More precisely, the optimal policy π_λ^* of the entropy-regularized problem could also be nearly optimal in terms of the unregularized objective function, as long as the regularization parameter λ is chosen to be small. Denote by π^* and π_λ^* the policies that maximize the objective function and the entropy-regularized objective function with the regularization parameter λ, respectively. Let V^* and V_λ^* represent the resulting optimal objective value function and the optimal regularized objective value function. Cen et al. (2020) shows a simple but crucial connection between π^* and π_λ^* via the following sandwich bound:

$$V^{\pi_\lambda^*}(\rho) \le V^{\pi^*}(\rho) \le V^{\pi_\lambda^*}(\rho) + \frac{\lambda \log |\mathcal{A}|}{1-\gamma},$$

which holds for all initial distribution ρ.

3 Stochastic policy gradient methods for entropy regularized RL

3.1 Review: Exact policy gradient methods

The policy gradient method (Algorithm 1) is one of the most popular approaches for a direct policy search in reinforcement learning (Sutton and Barto, 2018).

Algorithm 1 Exact policy gradient method

1: **Inputs**: $\{\eta_t\}_{t=1}^T$, θ_1.
2: **for** $t = 1, 2, \ldots, T-1$ **do**
3: $\quad \theta_{t+1} = \theta_t + \eta_t \nabla V_\lambda^{\theta_t}(\rho)$.
4: **end for**
5: **Outputs**: θ_T.

The uniform boundedness of the reward function r implies that the absolute value of the entropy-regularized state-value function and Q-value function are bounded.

Lemma 3.1 *$V_\lambda^\theta(s) \le \frac{\bar{r} + \lambda \log |\mathcal{A}|}{1-\gamma}$ and $Q_\lambda^\pi(s,a) \le \frac{\bar{r} + \lambda \log |\mathcal{A}|}{1-\gamma}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.*

Under the soft-max policy parameterization, one can obtain the following expression for the gradient of $V_\lambda^\pi(s)$ with respect to the policy parameter θ:

Lemma 3.2 *The entropy regularized policy gradient with respect to θ is*

$$\nabla V_\lambda^\theta(\rho) = \frac{1}{1-\gamma} \mathbb{E}_{s,a \sim v_\rho^{\pi_\theta}} \left[\boldsymbol{e}_{s,a} A_\lambda^\theta(s,a) \right], \tag{5}$$

where $\boldsymbol{e}_{s,a} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ has the value 1 at the entry corresponding to the state s and action a and has 0 at all other entries, and where $A_\lambda^\theta(s,a)$ is the soft advantage function defined as

$$\begin{aligned} A_\lambda^\theta(s,a) &= Q_\lambda^\theta(s,a) - \lambda \log \pi_\theta(a \mid s) - V_\lambda^\theta(s), \\ Q_\lambda^\theta(s,a) &= r(s,a) + \gamma \sum_{s'} \mathbb{P}(s' \mid s,a) V_\lambda^\theta(s'). \end{aligned}$$

Furthermore, the entropy regularized policy gradient is bounded, i.e., $\left\|\nabla V_\lambda^\theta(\rho)\right\| \le G$ for all $\rho \in \Delta(\mathcal{S})$ and $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, where $G \coloneqq \frac{2(\bar{r} + \lambda \log |\mathcal{A}|)}{(1-\gamma)^2}$.

Furthermore, it is shown in Lemmas 7 and 14 of Mei et al. (2020) that the policy gradient $\nabla V_\lambda^\theta(\rho)$ is Lipschitz continuous.

Lemma 3.3 *(Lipschitz-Continuity of Policy Gradient). The policy gradient $\nabla V_\lambda^\theta(\rho)$ is Lipschitz continuous with some constant $L > 0$, i.e.,*

$$\left\|\nabla V_\lambda^{\theta_1}(\rho) - \nabla V_\lambda^{\theta_2}(\rho)\right\| \le L \cdot \left\|\theta^1 - \theta^2\right\|,$$

for all $\theta^1, \theta^2 \in \mathbb{R}^d$, where the value of the Lipschitz constant L is defined as $L \coloneqq \frac{8\bar{r} + \lambda(4 + 8 \log |\mathcal{A}|)}{(1-\gamma)^3}$.

The results in Lemmas 3.1-3.3 are adopted from Mei et al. (2020).

Challenges for designing entropy regularized policy gradient estimators. Existing works either consider one-step entropy regularization (Williams, 1992; Mnih et al., 2016), KL divergence (Schulman et al., 2017), or the re-parametrization technique (Haarnoja et al., 2017, 2018) (which introduces approximation errors that are difficult to quantify exactly). In general, the regularized reward $r - \lambda \log \pi_\theta$ is policy-dependent and unbounded even though the original reward r is uniformly bounded. Hence, the existing estimators for the un-regularized setting must be modified to account for the policy-dependency and unboundedness while maintaining the essential properties of (nearly) unbiasedness and bounded variances. In the subsequent sections, we propose two (nearly) unbiased estimators and show that although the estimators may be unbounded due to unbounded regularized rewards, the variances are indeed bounded. The proofs of the results in this section can be found in Section 8 of the appendix.

3.2 Sampling the unbiased policy gradient

It results from (5) that in order to obtain an unbiased sample of $\nabla V_\lambda^\theta(\rho)$, we need to first draw a state-action pair (s,a) from the distribution $\nu_\rho^{\pi_\theta}(\cdot,\cdot)$ and then obtain an unbiased advantage function $A_\lambda^\theta(s,a)$

For the standard discounted infinite-horizon RL setting with bounded reward functions, Zhang et al. (2020) proposes an unbiased estimate of the advantage-function using the random horizon with a geometric distribution and the Monte-Carlo rollouts of finite horizons. However, their result cannot be immediately applied to the entropy-regularized RL setting since the entropy-regularized instantaneous reward $r(s,a) - \lambda \log \pi(a|s)$ could be unbounded when $\pi(a|s) \to 0$. Fortunately, we can still show that an unbiased policy gradient estimator with the bounded variance for the entropy regularized RL can be obtained in a similar fashion as in Zhang et al. (2020). In particular, we will use a random horizon that follows a certain geometric distribution in the sampling process. To ensure that the condition (i) is satisfied, we will use the last sample (s_H, a_H) of a finite sample trajectory $(s_0, a_0, s_1, a_1, \ldots, s_H, a_H)$ to be the sample at which $A_\lambda^\theta(\cdot,\cdot)$ is evaluated, where the horizon $H \sim \text{Geom}(1-\gamma)$. It can be shown that $(s_H, a_H) \sim \nu_\rho^{\pi_\theta}(s,a)$. Moreover, given (s_H, a_H), we will perform Monte-Carlo rollouts for two other trajectories with horizons $H', H'' \sim \text{Geom}\left(1-\gamma^{1/2}\right)$ independent of H, and estimate the advantage function value $A_\lambda^\theta(s,a)$ along the trajectories $(s_0', a_0', \ldots, s_{H'}')$ and $(s_0'', a_0'', \ldots, s_{H''}'')$ as follows:

$$\hat{A}_\lambda^\theta(s,a) = \hat{Q}_\lambda^\theta(s,a) - \lambda \log \pi_\theta(a|s) - \hat{V}_\lambda^\theta(s), \quad (6)$$

where

$$\begin{aligned}\hat{Q}_\lambda^\theta(s,a) =& r(s_0', a_0') + \sum_{t=1}^{H'} \gamma^{t/2} \cdot \left(r(s_t', a_t') - \lambda \log \pi_\theta(a'|s')\right) \\ & \mid s_0' = s, a_0' = a, \\ \hat{V}_\lambda^\theta(s) =& \sum_{t=0}^{H''} \gamma^{t/2} \cdot \left(r(s_t'', a_t'') - \lambda \log \pi_\theta(a_t''|s_t'')\right) \mid s_0'' = s.\end{aligned}$$

The subroutines of sampling the pair (s,a) from $\nu_\rho^{\pi_\theta}(\cdot,\cdot)$, estimating $\hat{Q}_\lambda^\theta(s,a)$ and estimating $\hat{V}_\lambda^\theta(s)$ are summarized as **Sam-SA**, **Est-EntQ** and **Est-EntV** in Algorithms 4, 5 and 6 in Section 8.3, respectively.

Motivated by the form of policy gradient in (5), we propose the following stochastic estimator:

$$\hat{\nabla} V_\lambda^\theta(\rho) = \frac{1}{1-\gamma} \cdot \boldsymbol{e}_{s_H, a_H} \cdot \hat{A}_\lambda^\theta(s_H, a_H), \quad (7)$$

where $s_H, a_H \leftarrow \textbf{Sam-SA}(\rho, \theta, \gamma)$ and $\hat{A}_\lambda^\theta$ is defined in (6). The following lemma shows that the stochastic policy gradient (7) is an unbiased estimator of $\nabla V_\lambda^\theta(\rho)$.

Lemma 3.4 *For $\hat{\nabla} V_\lambda^\theta(\rho)$ defined in (7), we have $\mathbb{E}[\hat{\nabla} V_\lambda^\theta(\rho)] = \nabla V_\lambda^\theta(\rho)$.*

The next lemma shows that the proposed PG estimator $\hat{\nabla} V_\lambda^\theta(\rho)$ has a bounded variance even if it is unbounded when π_θ approaches a deterministic policy.

Lemma 3.5 *For $\hat{\nabla} V_\lambda^\theta(\rho)$ defined in (7), we have $Var[\hat{\nabla} V_\lambda^\theta(\rho)] \le \sigma^2$, where $\sigma^2 = \frac{8}{(1-\gamma)^2}\left(\frac{\bar{r}^2 + (\lambda \log |\mathcal{A}|)^2}{(1-\gamma^{1/2})^2}\right)$.*

3.3 Sampling the trajectory-based policy gradient

Compared to the unbiased policy gradient with a random horizon in (7), a more practical policy gradient estimator is the trajectory-based policy gradient. To derive the trajectory-based policy gradient for the entropy-regularized RL, we first notice that the gradient $\nabla V_\lambda^\theta(\rho)$ can also be written as

$$\nabla V_\lambda^\theta(\rho) = \lambda \mathbb{E}\left[\left(\sum_{t=0}^{\infty} -\gamma^t \nabla \log \pi_\theta(a_t|s_t)\right)\right] + \quad (8)$$

$$\lambda \mathbb{E}\left[\left(\sum_{t=0}^{\infty} \nabla \log \pi_\theta(a_t|s_t)\right)\left(\sum_{t=0}^{\infty} -\gamma^t \log \pi_\theta(a_t|s_t)\right)\right] \quad (9)$$

$$+ \mathbb{E}\left[\left(\sum_{t=0}^{\infty} \nabla \log \pi_\theta(a_t|s_t)\right)\left(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right)\right], \quad (10)$$

where all expectations are taken over the trajectory distribution, i.e., $\tau \sim p(\tau|\theta)$. The term (10) in the

gradient $\nabla V_\lambda^\theta(\rho)$ is the gradient of the unregularized RL objective and the terms in (8)-(9) are introduced due to the entropy regularization.

Since the distribution $p(\tau|\theta)$ is unknown, $\nabla V_\lambda^\theta(\rho)$ needs to be estimated from samples. The trajectory-based estimators include REINFORCE (Williams, 1992), PGT (Sutton et al., 1999) and GPOMDP (Baxter and Bartlett, 2001). In practice, the truncated versions of these trajectory-based PG estimators are used to approximate the infinite sum in the PG estimator. Let $\tau^H = \{s_0, a_0, s_1, \ldots, s_{H-1}, a_{H-1}, s_H\}$ denote the truncation of the full trajectory τ of length H. For example, the commonly used truncated GPOMDP given by

$$g_1(\tau^H|\theta,\rho) \tag{11}$$
$$= \sum_{h=0}^{H-1}\left(\sum_{j=0}^{h}\nabla\log\pi_\theta(a_j|s_j)\right)\gamma^h r_h(s_h,a_h),$$
$$g_2(\tau^H|\theta,\rho) \tag{12}$$
$$=\lambda\sum_{h=0}^{H-1}\left(\sum_{j=0}^{h}\nabla\log\pi_\theta(a_j|s_j)\right)\left(-\gamma^h\log\pi_\theta(a_h|s_h)\right).$$

can be used to estimate the terms in (10) and (9) separately. In addition, the term in (8) can be approximated by the following estimator:

$$g_3(\tau^H|\theta,\rho) = \lambda\sum_{h=0}^{H-1}-\gamma^t\nabla\log\pi_\theta(a_h,s_h). \tag{13}$$

Then, the truncated PG estimator for ∇V_λ^θ can be written as:

$$\hat\nabla V_\lambda^{\theta,H}(\rho) = g_1(\tau^H|\theta,\rho) + g_2(\tau^H|\theta,\rho) + g_3(\tau^H|\theta,\rho). \tag{14}$$

Due to the horizon truncation, the policy gradient estimator (14) may no longer be unbiased, but its bias can be very small with a large horizon H.

Lemma 3.6 *For $\hat\nabla V_\lambda^{\theta,H}(\rho)$ defined in (14), we have*

$$\left\|\mathbb{E}[\hat\nabla V_\lambda^{\theta,H}(\rho)] - \nabla V_\lambda^\theta(\rho)\right\|_2 \le 2\lambda\gamma^H\left(\frac{\lambda+(\bar r+\lambda\log|\mathcal{A}|)H}{1-\gamma}+\frac{(\bar r+\lambda\log|\mathcal{A}|)}{(1-\gamma)^2}\right).$$

From Lemma 3.6, we can observe that the bias is proportional to γ^H and thus can be controlled to be arbitrarily small with a constant horizon up to some logarithmic term. We then show that the truncated PG estimator $\hat\nabla V^{\theta,H}$ has a bounded variance even if it may be unbounded when π_θ approaches a deterministic policy.

Lemma 3.7 *For $\hat\nabla V_\lambda^{\theta,H}(\rho)$ defined in (14), we have*

$$Var(\hat\nabla V_\lambda^{\theta,H}(\rho)) \le \frac{12\lambda}{(1-\gamma)^2}+\frac{12\bar r^2+24\lambda^2(\log|\mathcal{A}|)^2}{(1-\gamma)^4}.$$

3.4 Batched policy gradient algorithms

In practice, we can sample and compute a batch of independently and identically distributed policy gradient estimators $\{\hat\nabla V_\lambda^{\theta,i}(\rho)\}_{i=1}^B$ or $\{\hat\nabla V_\lambda^{\theta,H,i}(\rho)\}_{i=1}^B$, where B is the batch size, in order to reduce the estimation variance. To maximize the entropy-regularized objective function (4), we can then update the policy parameter θ by iteratively running gradient-ascent-based algorithms, i.e., $\theta_{t+1} = \theta_t + \frac{\eta_t}{B}\sum_{i=1}^B\hat\nabla V_\lambda^{\theta,i}(\rho)$, where $\eta_t > 0$ is the step size. The details of the unbiased policy gradient algorithm with a random horizon (and the trajectory-based policy gradient algorithm with the horizon truncation) for the entropy-regularized RL are summarized in Algorithm 2 (and Algorithm 3).

Algorithm 2 Ent-RPG: Random-horizon policy gradient algorithm for Entropy-regularized RL

Inputs: $\rho, \lambda, \theta_1, B, T, \{\eta_t\}_{t=1}^T$.
for $t = 1, 2, \ldots, T$ **do**
  for $i = 1, 2, \ldots, B$ **do**
    $s_{H_t}^i, a_{H_t}^i \leftarrow$ **SamSA**(ρ, θ_t, γ).
    $\hat Q_\lambda^{\theta_t,i} \leftarrow$ **Est-EntQ**$(s_{H_t}^i, a_{H_t}^i, \theta_t, \gamma, \lambda)$.
    $\hat V_\lambda^{\theta_t,i} \leftarrow$ **Est-EntV**$(s_{H_t}^i, \theta_t, \gamma, \lambda)$.
    $\hat A_\lambda^{\theta_t,i} \leftarrow \hat Q_\lambda^{\theta,i} - \lambda\log\pi_{\theta_t}(s_{H_t}^i \mid a_{H_t}^i) - \hat V_\lambda^{\theta,i}$.
  end for
  Perform policy gradient update: $\theta_{t+1} \leftarrow \theta_t + \frac{\eta_t}{(1-\gamma)B}\sum_{i=1}^B\left[\boldsymbol{e}_{s_{H_t}^i, a_{H_t}^i}\hat A_\lambda^{\theta_t,i}(s_{H_t}^i, a_{H_t}^i)\right]$
end for
Outputs: θ_T.

Algorithm 3 Stochastic PG for entropy regularized RL

Inputs: $\rho, \lambda, \theta_1, B, T, \{\eta_t\}_{t=1}^T$.
for $t = 1, 2, \ldots, T-1$ **do**
  Sample B trajectories $\{\tau_i^H\}_{i=1}^B$ from $p(\cdot|\theta_t, \rho)$;
  Compute $u_t = \frac{1}{B}\sum_{i=1}^B\hat\nabla V_\lambda^{\theta,H,i}(\rho)$ where $\hat\nabla V_\lambda^{\theta,H,i}(\rho)$ is given by (14);
  Update $\theta_{t+1} = \theta_t + \eta_t u_t$;
end for
Outputs: θ_T;

4 Global convergence and sample complexity

In this section, we first review some key results for entropy-regularized RL with the exact policy gradient and highlight the difficulty of generalizing these results to the setting with stochastic policy gradients. Then, we develop the counterparts of these key results when the stochastic policy gradient estimator, time-varying step-sizes and a large batch size are used. Due to space

restrictions and in order to facilitate the presentation of the main ideas, we will mainly focus on the analysis of the unbiased policy gradient estimator in (7). Similar results hold for the trajectory-based policy gradient estimator in (14) since its bias is exponentially small with respect to the horizon (see Lemma 3.6). We leave the formal discussion of these results as future work.

4.1 Review: Linear convergence with exact policy gradient

A key result from Lemma 15 of Mei et al. (2020) shows that, under the soft-max parameterization, the entropy-regularized value function $V_\lambda^\theta(\rho)$ in (4) satisfies a non-uniform Łojasiewicz inequality as follows:

Lemma 4.1 (Mei et al. (2020)) *It holds that* $\left\|\nabla V_\lambda^\theta(\rho)\right\|_2^2 \geq C(\theta)(V_\lambda^{\theta^*}(\rho) - V_\lambda^\theta(\rho))$, *where* $C(\theta) = \frac{2\lambda}{|\mathcal{S}|}\min_s \rho(s) \min_{s,a} \pi_\theta(a|s)^2 \left\|\frac{d_\rho^{\pi_\lambda^*}}{\rho}\right\|_\infty^{-1}$.

It is also shown in Theorem 1 of Cen et al. (2020) that the optimal policy for the entropy-regularized RL problem 4 is unique. Furthermore, it is shown in Lemma 16 of Mei et al. (2020) that the action probabilities under the soft-max parameterization are uniformly bounded away from zero if the exact policy gradient is available.

Lemma 4.2 (Mei et al. (2020)) *Using the exact PG (Algorithm 1) with* $\eta_t \leq \frac{2}{L}$ *for the entropy regularized objective, it holds that* $\inf_{t\geq 1} \min_{s,a} \pi_{\theta_t}(a|s) > 0$.

With Lemmas 3.3, 4.1 and 4.2, it is shown in Theorem 6 of Mei et al. (2020) that the convergence rate for the entropy regularized policy gradient in general MDPs is $O\left(e^{-t}\right)$:

Lemma 4.3 (Mei et al. (2020)) *Consider Algorithm 1 with the entropy regularized objective and soft-max parametrization and* $\eta_t = \frac{1}{L}$. *There exists a problem-dependent constant* $C > 0$ *such that the following inequality holds for all* $t \geq 1$*:*

$$\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) \leq \left\|\frac{1}{\rho}\right\|_\infty \cdot \frac{1 + \lambda \log|\mathcal{A}|}{(1-\gamma)^2} \cdot e^{-C(t-1)}.$$

It is shown in Mei et al. (2020) that the value of C depends on $\inf_{t\geq 1} \min_{s,a} \pi_{\theta_t}(a|s) > 0$, where $\{\theta_t\}_{t=1}^\infty$ is generated by Algorithm 1. With a bad initialization θ_1, $\min_{s,a} \pi_{\theta_1}(a|s)$ could be very small and result in a slow convergence rate. When studying the stochastic policy gradient, this issue of bad initialization will create more severe challenges on the convergence, which we will discuss in the following sections.

One main challenge is the boundedness of iterations under the stochastic policy gradient. If the iterations of Algorithm 2 remain in a bounded region, then the results of Lemma 4.3 can be easily generalized to the stochastic policy gradient setting. However, unfortunately, the iterates of stochastic gradient methods may indeed escape to infinity in general, rendering the entire scheme useless (Benaïm, 1999; Borkar, 2009). In particular, when using the stochastic truncated PG for the entropy regularized RL, the key result of Lemma 4.2 may no longer hold true. This in turn results in the loss of gradient domination condition in guaranteeing the global convergence.

4.2 Landscape of a simple bandit example

To have a better understanding of the landscape of the entropy-regularized value function, we visualize its landscape in this section. For the simplicity of the visualization, we use a simple bandit example (corresponding to $\gamma = 0$) with 2 actions, 2 parameters (θ_1, θ_2), the reward vector $r = [2, 1]$ and the regularization parameter $\lambda = 1$. Then, the entropy-regularized value function can be written as $\pi_\theta^\top (r - \log \pi_\theta)$.

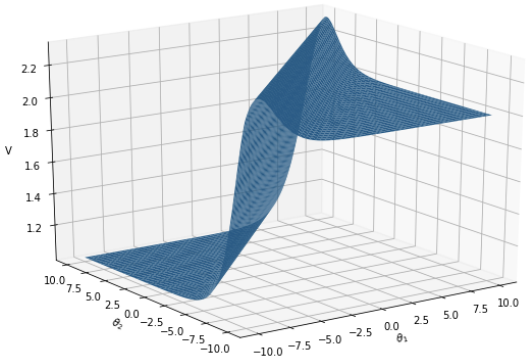

Figure 1: Landscape of $\pi_\theta^\top (r - \log \pi_\theta)$.

As shown in Figure 1, the entropy-regularized value function is not coercive. When θ_1 goes to positive (negative) infinity and θ_2 goes to negative (positive) infinity, the landscape will become highly flat. It can also be seen that there is a line space for (θ_1, θ_2) at which the entropy-regularized value function is maximum.

When the stochastic policy gradient is used, the search direction may be dominated by the gradient estimation noise at the region where the landscape is highly flat. This may further lead to the failure of the convergence for the stochastic policy gradient algorithm if the initial point is at the flat region. However, in the next section, we will show that the stochastic policy gradient can still converge to the optimal policy from an arbitrary initial point, given a sufficiently large number of samples.

4.3 Global convergence with arbitrary initialization

We will first show that the stochastic policy gradient method proposed in Algorithm 2 asymptotically converges to a region where the policy gradient vanishes almost surely if a specific adaptive step-size sequence is used.

Lemma 4.4 *Suppose that the sequence $\{\theta_t\}_{t=1}^{\infty}$ is generated by Algorithm 2 for the entropy regularized objective with the step-sizes satisfying $\sum_{t=1}^{\infty}\eta_t=\infty, \sum_{t=1}^{\infty}\eta_t^2<\infty$ and $\eta_t\le\frac{2}{L}$ for all $t=1,2,\ldots$. It holds that $\lim_{t\to\infty}\left\|\nabla V_\lambda^{\theta_t}(\rho)\right\|_2=0$ with probability 1.*

This result follows from classic results for the Robbins-Monro algorithm (Bertsekas and Tsitsiklis, 2000; Benaïm and Hirsch, 1996; Kushner and Clark, 2012) when an unbiased policy gradient estimator with the bounded variance, as in Algorithm 2, is used in the update rule. No requirement on the batch size B is needed in Lemma 4.4. We refer the reader to the supplement in Section 9 for the details of the proof.

However, since the entropy-regularized value function $V_\lambda^\theta(\rho)$ is not coercive in θ and it may be the case that the gradient $\nabla V_\lambda^{\theta_t}(\rho)$ diminishing to 0 corresponds to θ_t going to infinity instead of converging to a stationary point. In addition, the existing results (Benaïm, 1999; Benaïm and Hirsch, 1996; Kushner and Clark, 2012) on the almost surely stationary point convergence rely on the assumption that the trajectories of the process are bounded, i.e.,

$$\sup_{t\ge0}\|\theta_t\|<\infty,\ \text{almost surely.}$$

This assumption is proven to hold when the function is coercive (Mertikopoulos et al., 2020). However, when the function is not coercive, as in our problem, it is very challenging to characterize the trade-off between the gradient information and the estimation error without additional assumptions.

To overcome this challenge, we will use a large batch size to control the estimation error. Then, with a small estimation error, we can show that if the iterations with the exact policy gradient are bounded, then the iterations with the unbiased stochastic policy gradient will remain bounded with high probability. This will further imply that the unbiased stochastic policy gradient will converge to the globally optimal policy with high probability. Before presenting the theorem, we first denote $D(\theta_t)=V_\lambda^{\theta^*}(\rho)-V_\lambda^{\theta_t}(\rho)$ as the sub-optimality gap between $V_\lambda^{\theta^*}(\rho)$ and $V_\lambda^{\theta_t}(\rho)$.

Theorem 4.5 *Consider arbitrary tolerance levels $\delta>0$ and $\epsilon>0$. For every initial point θ_1, there exists a constant $C_\delta^0>0$ such that if θ_T is generated by Algorithm 2 with $\eta_t=\eta\le\min\left\{\frac{\log T}{TL},\frac{8}{C_\delta^0},\frac{1}{2L}\right\}$ and*

$$T=\mathcal{O}((\delta\epsilon)^{-q}),\ B=\tilde{\mathcal{O}}\left(\max\left\{(\delta\epsilon)^{-1},T\right\}\right),$$

where $q=\frac{8L}{C_\delta^0\ln 2}$, then we have $\mathbb{P}(D(\theta_T)\le\epsilon)\ge1-\delta$. In total, it requires $\tilde{\mathcal{O}}\left(\max\left\{(\delta\epsilon)^{-1-q},(\delta\epsilon)^{-2q}\right\}\right)$ samples to obtain an ϵ-optimal policy with high probability.

The value of C_δ^0 depends on the problem parameters, namely, $|\mathcal{A}|,|\mathcal{S}|,\gamma,\lambda,\rho$, the initial point θ_1 and the constant δ. To facilitate the presentation, we hide the dependency of T and B on the parameters such as the initial sub-optimality gap $D(\theta_1)$, the variance of $\hat{\nabla}V_\lambda^\theta(\rho)$ defined in Lemma 3.5 and the boundedness region of the iterations with the exact policy gradient. The definition of C_δ^0, the exact bound on the number of iterations T and the batch size B can be found in the proof of Theorem 4.5 in Section 10. We refer the reader to the supplement in Section 10 for more details and provide a short proof sketch below:

1. When $\{\theta_t\}_{t=1}^T$ are bounded, we can use Lemma 4.1 to show that $D(\theta_t)$ is linearly convergent up to some aggregated estimation error.

2. Since the iterations with the exact policy gradient are bounded, we can show that the iterations with the unbiased stochastic policy gradient remain bounded with high probability. This is due to using a large batch size to control the aggregated policy gradient estimation error.

4.4 Faster convergence with good initialization

Theorem 4.5 shows the asymptotic global convergence of the unbiased stochastic policy gradient method for the entropy-regularized RL problem, but its sample complexity may be high if the initialization is not good. In this section, we utilize the curvature information around the optimal policy to obtain much better sample complexities under a good initialization. To this end, we first show that, with the stochastic policy gradient, the action probabilities will still remain uniformly bounded with high probability if the initial policy is not too far away from the optimal one.

Lemma 4.6 *Given a tolerance level $\delta>0$, let π_{θ^*} be the optimal policy of $V_\lambda^\theta(\rho)$. Assume further that Algorithm 2 is run for T iterates with a step-size sequence of the form $\eta_t=1/(t+t_0)$ and a batch-size sequence $B\ge\frac{1}{\eta_t}$ for all $t=1,2,\ldots,T$. If $t_0>0$ is large enough, then there exist a constant $\alpha\in(0,1)$ and a neighborhood*

$\mathcal{U}_0$ of π_{θ^*} such that, if $\pi_{\theta_1} \in \mathcal{U}_0$, the event

$$\Omega_{\alpha,T} = \Big\{ \min_{s,a} \pi_{\theta_t}(a|s) \geq (1-\alpha) \min_{s,a} \pi_{\theta^*}(a|s), \quad \text{for all } t = 1, 2, \ldots, T \Big\} \tag{15}$$

occurs with probability at least $1-\delta$.

The definition of the region $\mathcal{U}_0$ appears in the proof of this lemma in the supplementary material. The proof of Lemma 4.6 heavily relies on the fact that the entropy-regularized value function $V_\lambda^\theta(\rho)$ is lower-bounded by a quadratic function around the optimal policy π_{θ^*} with respect to π_θ. It also involves a delicate combination of non-standard techniques, some of which are built on a range of ideas and techniques due to Hsieh et al. (2019, 2020); Mei et al. (2021); Mertikopoulos and Zhou (2019); Mertikopoulos et al. (2020). We refer the reader to the supplement in Section 11 for more details and provide a short sketch below:

1. We first characterize the maximum amount by which $D(\theta_t)$ can grow at each step. This quantity can be large for any given t but we show that, with high probability, the aggregation of these errors remains controllably small under the stated conditions on the step-sizes and batch size.

2. As a result of the above result, if $D(\theta_1)$ is not too big, $D(\theta_t)$ cannot be too large at all times. $D(\theta_1)$ not being too big can be guaranteed by the Lipschitz continuity of the entropy-regularized value function $V_\lambda^\theta(\rho)$ and the initial condition that π_{θ_1} is close to π_{θ^*}.

3. Finally, we show that the entropy-regularized value function $V_\lambda^\theta(\rho)$ is lower-bounded by a quadratic function around the optimal policy π_{θ^*} with respect to π_θ. Thus, $D(\theta_t)$ not being too big also implies that π_{θ_t} remains close to π_{θ^*} at all times.

From Lemma 4.6, we know that, with a good initialization, the policies $\{\pi_{\theta_t}\}_{t=1}^T$ will remain in the interior of the probability simplex with high probability. We conclude our analysis of the stochastic PG for entropy-regularized RL by establishing the algorithm's improved sample complexity when the initial policy is not far away from the optimal policy, as stated below.

Theorem 4.7 *Given some tolerance levels $\delta > 0$ and $\epsilon > 0$, let π_{θ^*} be the optimal policy of $V_\lambda^\theta(\rho)$. Assume that Algorithm 2 is run for T iterations with a step-size sequence of the form $\eta_t = 1/(t+t_0)$ and a batch-size B. If $t_0 > 0$ is large enough, then there exist a constant $\alpha \in (0,1)$ and a neighborhood $\mathcal{U}_0$ of π_{θ^*} such that, if $\pi_{\theta_1} \in \mathcal{U}_0$, it holds that*

- *Conditioned on $\Omega_{\alpha,T}$ defined in (15), we have*

$$\mathbb{E}[D(\theta_t)|\Omega_{\alpha,T}] \leq \frac{t_0 D(\theta_1)}{(T+t_0)(1-\delta)} + \frac{\sigma^2 \ln(T+t_0)}{B(1-\delta)C_\alpha},$$

- *For every $\epsilon > 0$, if*

$$T \geq \frac{t_0 D(\theta_1)}{\delta\epsilon} - t_0, \quad B \geq \frac{\sigma^2 \ln(T+t_0)}{C_\alpha \delta \epsilon}.$$

then we have $\mathbb{P}(D(\theta_T) \leq \epsilon) \geq 1 - \mathcal{O}(\delta)$. In total, it takes $\tilde{\mathcal{O}}(\frac{1}{\epsilon^2})$ samples to have $D(\theta_T) \leq \epsilon$ with high probability.

The constant C_α depends on the problem parameters, namely, $|\mathcal{A}|, |\mathcal{S}|, \gamma, \lambda, \rho$, which can be found in the proof of the theorem in the supplement. For the first claim of Theorem 4.7, by conditioning on the event $\Omega_{\alpha,T}$, we can obtain the result in Lemma 4.1, with $C(\theta)$ being uniformly lower-bounded by a positive constant. Combining with the smoothness, this leads to a recursion to control the optimally gap $D(\theta_T)$. The main challenge here is that, after conditioning, the gradient estimation is no longer unbiased, and therefore we need to adapt our analysis to the new estimation error. Then, the second claim follows from the first claim by applying the law of total expectation and Markov inequality. We refer the reader to the supplement in Section 12 for the details.

5 Conclusion

In this work, we studied the global convergence and the sample complexity of stochastic policy gradient methods for the entropy-regularized RL with the softmax parameterization. We proposed two new (nearly) unbiased policy gradient estimators for the entropy-regularized RL and proved that they have a bounded variance even though they could be unbounded. In addition, this work provided the first global convergence result for stochastic policy gradient methods for the entropy-regularized RL, although the sample complexity could be high due to the loss of curvature for some parameter space. Furthermore, with a good initialization, we showed that the policies $\{\pi_{\theta_t}\}_{t=1}^T$ remain in the interior of the probability simplex with high probability and, therefore, an improved sample complexity can be guaranteed.

References

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

Ahmed, Z., Le Roux, N., Norouzi, M., and Schuurmans, D. (2019). Understanding the impact of entropy on policy optimization. In *International Conference on Machine Learning*, pages 151–160. PMLR.

Baxter, J. and Bartlett, P. L. (2001). Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350.

Benaïm, M. (1999). Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pages 1–68. Springer.

Benaïm, M. and Hirsch, M. W. (1996). Asymptotic pseudotrajectories and chain recurrent flows, with applications. *Journal of Dynamics and Differential Equations*, 8(1):141–176.

Bertsekas, D. P. and Tsitsiklis, J. N. (2000). Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642.

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

Borkar, V. S. (2009). *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer.

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

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Eysenbach, B. and Levine, S. (2019). If MaxEnt RL is the answer, what is the question? *arXiv preprint arXiv:1910.01913*.

Eysenbach, B. and Levine, S. (2021). Maximum entropy rl (provably) solves some robust rl problems. *arXiv preprint arXiv:2103.06257*.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. (2017). Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pages 1352–1361. PMLR.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR.

Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. (2019). On the convergence of single-call stochastic extra-gradient methods. *arXiv preprint arXiv:1908.08465*.

Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. (2020). Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. *arXiv preprint arXiv:2003.10162*.

Kushner, H. J. and Clark, D. S. (2012). *Stochastic approximation methods for constrained and unconstrained systems*, volume 26. Springer Science & Business Media.

Mei, J., Gao, Y., Dai, B., Szepesvari, C., and Schuurmans, D. (2021). Leveraging non-uniformity in first-order non-convex optimization. *International Conference on Machine Learning*.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. (2020). On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR.

Mertikopoulos, P., Hallak, N., Kavis, A., and Cevher, V. (2020). On the almost sure convergence of stochastic gradient descent in non-convex problems. *arXiv preprint arXiv:2006.11144*.

Mertikopoulos, P. and Zhou, Z. (2019). Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1):465–507.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR.

Neu, G., Jonsson, A., and Gómez, V. (2017). A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*.

O'Donoghue, B., Munos, R., Kavukcuoglu, K., and Mnih, V. (2016). Combining policy gradient and q-learning. *arXiv preprint arXiv:1611.01626*.

Schulman, J., Abbeel, P., and Chen, X. (2017). Equivalence between policy gradients and soft q-learning. *CoRR*, abs/1704.06440.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, R. S., McAllester, D. A., Singh, S. P., Mansour, Y., et al. (1999). Policy gradient methods for reinforcement learning with function approximation. In *NIPs*, volume 99, pages 1057–1063. Citeseer.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Williams, R. J. and Peng, J. (1991). Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268.

Yuan, R., Gower, R. M., and Lazaric, A. (2021). A general sample complexity analysis of vanilla policy gradient. *arXiv preprint arXiv:2107.11433*.

Zang, H., Li, X., Zhang, L., Zhao, P., and Wang, M. (2020). Teac: Intergrating trust region and max entropy actor critic for continuous control.

Zhang, K., Koppel, A., Zhu, H., and Basar, T. (2020). Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612.

Supplementary Materials

6 Related work

One line of theoretical research on entropy regularization focuses on its connection to other methodologies in RL. O'Donoghue et al. (2016) shows that policy gradient and Q-learning are (approximately) equivalent when a one-step entropy regularization is adopted. The exact equivalence between these algorithms is then established in Schulman et al. (2017) under a trajectory-level KL regularization[3] More recently, Eysenbach and Levine (2019, 2021) showed that entropy-regularized RL is equivalent to POMDPs, two-player games and robust RL.

Stochastic policy gradient estimators with the original one-step entropy regularization has been proposed and adopted in Williams and Peng (1991); Mnih et al. (2016); O'Donoghue et al. (2016). For trajectory-level entropy regularization, exact (visitation measure-based) policy gradient formula has been derived in Ahmed et al. (2019) and later re-derived in the soft-max policy parametrization setting in Mei et al. (2020), while stochastic policy gradient estimators have not been formally proposed or studied in the literature. The only exception is Schulman et al. (2017), which provides stochastic policy gradient estimators with a related but different trajectory-level KL regularization term.

The convergence analyses of entropy regularized RL algorithms have so far been focused on natural policy gradient and trust-region policy optimization methods (Neu et al., 2017; Cen et al., 2020), with the only exception being Mei et al. (2020) that studied the exact gradient setting. The prior literature lacks convergence results for stochastic (vanilla) policy gradient methods with (trajectory-level) entropy regularization.

7 Notation

The set of real numbers is shown as $\mathbb{R}$. $u \sim \mathcal{U}$ means that u is a random vector sampled from the distribution $\mathcal{U}$. We use $|\mathcal{X}|$ to denote the number of elements in a finite set X. The notions $\mathbb{E}_\xi[\cdot]$ and $\mathbb{E}[\cdot]$ refer to the expectation over the random variable ξ and over all of the randomness. The notion $\mathrm{Var}[\cdot]$ refers to the variance. For vectors $x, y \in \mathbb{R}^d$, let $\|x\|_1$, $\|x\|_2$ and $\|x\|_\infty$ denote the ℓ_1-norm, ℓ_2-norm and ℓ_∞-norm. We use $\langle x, y\rangle$ denote the inner product. For a matrix A, $A \succcurlyeq 0$ means that A is positive semi-definite. Given a variable x, the notation $a = \mathcal{O}(b(x))$ means that $a \leq C \cdot b(x)$ for some constant $C > 0$ that is independent of x. Similarly, $a = \widetilde{\mathcal{O}}(b(x))$ indicates that the previous inequality may also depend on the function $\log(x)$, where $C > 0$ is again independent of x. We use $\mathrm{Geom}(x)$ to denote a geometric distribution with the parameter x.

[3]Note that this is related to but different from the widely-used trajectory-level entropy regularization later introduced in Haarnoja et al. (2017).

8 Properties of stochastic policy gradient

8.1 Proof of Lemma 3.1

Proof. We first show that the entropy-regularized state-value function is upper bounded:

$$\begin{aligned}
V_\lambda^\theta(s) &= \mathbb{E}_{s_0=s, a_t\sim\pi_{\theta_t}(\cdot|s_t), s_{t+1}\sim\mathbb{P}(\cdot|s_t,a_t)}\left[\sum_{t=0}^{\infty}\gamma^t\left(r\left(s_t,a_t\right)-\lambda\log\pi_{\theta_t}\left(a_t\mid s_t\right)\right)\right] \\
&= \frac{1}{1-\gamma}\sum_s d_s^{\pi_{\theta_t}}(s)\cdot\left[\sum_a \pi_{\theta_t}(a\mid s)\cdot\left(r(s,a)-\lambda\log\pi_{\theta_t}(a\mid s)\right)\right] \\
&\le \frac{1}{1-\gamma}\sum_s d_s^{\pi_{\theta_t}}(s)\cdot(\bar{r}+\lambda\log A) \\
&\le \frac{\bar{r}+\lambda\log|\mathcal{A}|}{1-\gamma},
\end{aligned}$$

where the first inequality is due to $-\sum_a \pi(a|s)\cdot\log\pi(a|s)\le\log|\mathcal{A}|$. Then, by the definition of $Q_\lambda^\pi(s,a)$, we have

$$\begin{aligned}
Q_\lambda^\pi(s,a) &= r(s,a)+\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[V_\lambda^\pi\left(s'\right)\right] \\
&\le \frac{\bar{r}}{1-\gamma}+\frac{\gamma(\lambda\log|\mathcal{A}|)}{1-\gamma} \\
&\le \frac{\bar{r}+\lambda\log|\mathcal{A}|}{1-\gamma}.
\end{aligned}$$

This completes the proof. □

8.2 Proof of Lemma 3.2

This result is similar to Lemma 10 in Mei et al. (2020) and Lemma S.2 in Ahmed et al. (2019). We provide a proof here for completeness.

Proof. Since $V_\lambda^\theta(\rho)=\mathop{\mathbb{E}}_{s\sim\rho}\sum_a\pi_\theta(a\mid s)\cdot\left[Q_\lambda^\theta(s,a)-\lambda\log\pi_\theta(a\mid s)\right]$, taking derivative with respect to θ rise to,

$$\begin{aligned}
&\frac{\partial V_\lambda^\theta(\rho)}{\partial\theta} \\
&= \mathop{\mathbb{E}}_{s\sim\rho}\sum_a\frac{\partial\pi_\theta(a\mid s)}{\partial\theta}\cdot\left[Q_\lambda^\theta(s,a)-\lambda\log\pi_\theta(a\mid s)\right]+\mathop{\mathbb{E}}_{s\sim\rho}\sum_a\pi_\theta(a\mid s)\cdot\left[\frac{\partial Q_\lambda^\theta(s,a)}{\partial\theta}-\lambda\cdot\frac{1}{\pi_\theta(a\mid s)}\cdot\frac{\partial\pi_\theta(a\mid s)}{\partial\theta}\right] \qquad (16)\\
&= \mathop{\mathbb{E}}_{s\sim\rho}\sum_a\frac{\partial\pi_\theta(a\mid s)}{\partial\theta}\cdot\left[Q_\lambda^\theta(s,a)-\lambda\log\pi_\theta(a\mid s)\right]+\mathop{\mathbb{E}}_{s\sim\rho}\sum_a\pi_\theta(a\mid s)\cdot\frac{\partial Q_\lambda^\theta(s,a)}{\partial\theta} \\
&= \mathop{\mathbb{E}}_{s\sim\rho}\sum_a\frac{\partial\pi_\theta(a\mid s)}{\partial\theta}\cdot\left[Q_\lambda^\theta(s,a)-\lambda\log\pi_\theta(a\mid s)\right]+\gamma\cdot\mathop{\mathbb{E}}_{s\sim\rho}\sum_a\pi_\theta(a\mid s)\sum_{s'}\mathbb{P}\left(s'\mid s,a\right)\cdot\frac{\partial\tilde{V}^{\pi_\theta}\left(s'\right)}{\partial\theta} \\
&= \frac{1}{1-\gamma}\sum_s d_\rho^{\pi_\theta}(s)\sum_a\frac{\partial\pi_\theta(a\mid s)}{\partial\theta}\cdot\left[Q_\lambda^\theta(s,a)-\lambda\log\pi_\theta(a\mid s)\right] \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi_\theta},a\sim\pi_\theta(a|s)}\left[\nabla\log\pi_\theta(a\mid s)\cdot\left[Q_\lambda^\theta(s,a)-\lambda\log\pi_\theta(a\mid s)\right]\right]. \qquad (17)
\end{aligned}$$

where the second equation is because of

$$\sum_a\pi_\theta(a\mid s)\cdot\left[\frac{1}{\pi_\theta(a\mid s)}\cdot\frac{\partial\pi_\theta(a\mid s)}{\partial\theta}\right]=\sum_a\frac{\partial\pi_\theta(a\mid s)}{\partial\theta}=\frac{\partial}{\partial\theta}\sum_a\pi_\theta(a\mid s)=\frac{\partial 1}{\partial\theta}=0.$$

Due to $s' \neq s$, $\frac{\partial \pi_\theta(a|s)}{\partial \theta(s',\cdot)} = 0$, we obtain

$$\begin{aligned}\frac{\partial V_\lambda^\theta(\rho)}{\partial \theta(s,\cdot)} &= \frac{1}{1-\gamma} \cdot d_\rho^{\pi_\theta}(s) \cdot \left[\sum_a \frac{\partial \pi_\theta(a \mid s)}{\partial \theta(s,\cdot)} \cdot \left[Q_\lambda^\theta(s,a) - \lambda \log \pi_\theta(a \mid s)\right]\right] \\ &= \frac{1}{1-\gamma} \cdot d_\rho^{\pi_\theta}(s) \cdot \left(\frac{d\pi(\cdot \mid s)}{d\theta(s,\cdot)}\right)^\top \left[Q_\lambda^\theta(s,\cdot) - \lambda \log \pi_\theta(\cdot \mid s)\right].\end{aligned}$$

Since $\frac{d\pi(\cdot|s)}{d\theta(s,\cdot)} = \mathrm{diag}(\pi(\cdot \mid s)) - \pi(\cdot \mid s)\pi(\cdot \mid s)^\top$, where $\mathrm{diag}(x)$ denotes the diagonal matrix that has x on the diagonal, we can write

$$\begin{aligned}\frac{\partial V_\lambda^\theta(\rho)}{\partial \theta(s,a)} &= \frac{1}{1-\gamma} \cdot d_\rho^{\pi_\theta}(s) \cdot \pi_\theta(a \mid s) \cdot \left[Q_\lambda^\theta(s,a) - \lambda \log \pi_\theta(a \mid s) - \sum_a \pi_\theta(a \mid s) \cdot \left[Q_\lambda^\theta(s,a) - \lambda \log \pi_\theta(a \mid s)\right]\right] \\ &= \frac{1}{1-\gamma} \cdot d_\rho^{\pi_\theta}(s) \cdot \pi_\theta(a \mid s) \cdot \left[Q_\lambda^\theta(s,a) - \lambda \log \pi_\theta(a \mid s) - V_\lambda^\theta(s)\right] \\ &= \frac{1}{1-\gamma} \cdot d_\rho^{\pi_\theta}(s) \cdot \pi_\theta(a \mid s) \cdot A_\lambda^\theta(s,a), \quad \forall a \in \mathcal{A}.\end{aligned}$$

By stacking all components s, a into a vector, we obtain

$$\begin{aligned}\frac{\partial V_\lambda^\theta(\rho)}{\partial \theta} &= \frac{1}{1-\gamma} \cdot \sum_s d_\rho^{\pi_\theta}(s) \sum_a \pi_\theta(a \mid s) \cdot \boldsymbol{e}_{s,a} \cdot A_\lambda^\theta(s,a) \\ &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\rho^{\pi_\theta}, a \sim \pi_\theta(a|s)} \left[\boldsymbol{e}_{s,a} A_\lambda^\theta(s,a)\right].\end{aligned}$$

Finally, from (17) and Jensen's inequality, we have

$$\begin{aligned}\left\|\frac{\partial V_\lambda^\theta(\rho)}{\partial \theta}\right\| &\leq \frac{1}{1-\gamma} \max_{a,s} \|\nabla \log \pi_\theta(a \mid s)\| \cdot \max_s \left\|\sum_a \pi_\theta(a \mid s) \left[Q_\lambda^\theta(s,a) - \lambda \log \pi_\theta(a \mid s)\right]\right\| \\ &= \frac{1}{1-\gamma} \max_{a,s} \|\nabla \log \pi_\theta(a \mid s)\| \cdot \max_s \left\|V_\lambda^\theta(s)\right\| \\ &\leq \frac{2(\bar{r} + \lambda \log |\mathcal{A}|)}{(1-\gamma)^2}.\end{aligned}$$

This completes the proof. □

8.3 Subroutines for Algorithm 2

The subroutines of sampling one pair (s,a) from $\nu_\rho^{\pi_\theta}(\cdot,\cdot)$, estimating $\hat{Q}_\lambda^\theta(s,a)$, and estimating $\hat{V}_\lambda^\theta(s)$ are summarized as **Sam-SA**, **Est-EntQ** and **Est-EntV** in Algorithms 4, 5 and 6, respectively.

Algorithm 4 Sam-SA: Sample for $s, a \sim \nu_\rho^{\pi_\theta}(\cdot,\cdot)$

1: **Inputs**: ρ, θ, γ.
2: Draw $H \sim \mathrm{Geom}(1-\gamma)$.
3: Draw $s_0 \sim \rho$ and $a_0 \sim \pi_\theta(\cdot|s_0)$
4: **for** $h = 1, 2, \ldots, H-1$ **do**
5: Simulate the next state $s_{h+1} \sim \mathbb{P}(\cdot|s_h, a_h)$ and action $a_{h+1} \sim \pi_{\theta_t}(\cdot|s_{h+1})$.
6: **end for**
7: **Outputs**: s_H, a_H.

Algorithm 5 Est-EntQ: Unbiasedly estimating entropy-regularized Q function

Inputs: s, a, γ, λ and θ.
Initialize $s_0 \leftarrow s, a_0 \leftarrow a, \hat{Q} \leftarrow r(s_0, a_0)$.
Draw $H \sim \text{Geom}(1-\gamma^{1/2})$.
for $h = 0, 1, \ldots, H-1$ **do**
  Simulate the next state $s_{h+1} \sim \mathbb{P}(\cdot|s_h, a_h)$ and action $a_{h+1} \sim \pi_\theta(\cdot|s_{h+1})$.
  Collect the instantaneous reward $r(s_{h+1}, a_{h+1}) - \lambda \log \pi_\theta(a_{h+1}|s_{h+1})$ and add to the value $\hat{Q}$: $\hat{Q} \leftarrow \hat{Q} + \gamma^{(h+1)/2}\left(r(s_{h+1}, a_{h+1}) - \lambda \log \pi_\theta(a_{h+1}|s_{h+1})\right)$,
end for
Outputs: $\hat{Q}$.

Algorithm 6 Est-EntV: Unbiasedly estimating entropy-regularized state-value function

Inputs: s, γ, λ and θ.
Initialize $s_0 \leftarrow s$, draw $a_0 \sim \pi_\theta(\cdot|s_0)$ and let $\hat{V} \leftarrow r(s_0, a_0) - \lambda \log \pi_\theta(a_0|s_0)$.
Draw $H \sim \text{Geom}(1-\gamma^{1/2})$.
for $h = 0, 1, \ldots, H-1$ **do**
  Simulate the next state $s_{h+1} \sim \mathbb{P}(\cdot|s_h, a_h)$ and action $a_{h+1} \sim \pi_\theta(\cdot|s_{h+1})$.
  Collect the instantaneous reward $r(s_{h+1}, a_{h+1}) - \lambda \log \pi_\theta(a_{h+1}|s_{h+1})$ and add to the value $\hat{V}$: $\hat{V} \leftarrow \hat{V} + \gamma^{(h+1)/2}\left(r(s_{h+1}, a_{h+1}) - \lambda \log \pi_\theta(a_{h+1}|s_{h+1})\right)$,
end for
Outputs: $\hat{V}$.

8.4 Proof of Lemma 3.4

Proof. We first show the unbiasedness of the Q-estimate, i.e., $\mathbb{E}\left[\hat{Q}_\lambda^\theta(s,a) \mid \theta, s, a\right] = Q_\lambda^\theta(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $\theta \in \mathbb{R}^d$. In particular, from the definition of $Q_\lambda^\theta(s,a)$, we have

$$\begin{aligned}
&\mathbb{E}\left[\hat{Q}_\lambda^\theta(s,a) \mid \theta, s, a\right] \\
&\quad = \mathbb{E}\left[r(s_0,a_0) + \sum_{h=1}^{H'} \gamma^{h/2} \cdot \left(r(s_h,a_h) - \lambda \log \pi_\theta(a_h|s_h)\right) \mid \theta, s_0 = s, a_0 = a\right] \\
&\quad = \mathbb{E}\left[r(s_0,a_0) + \sum_{h=1}^{\infty} \mathbb{1}_{H' \geq h \geq 0} \gamma^{h/2} \cdot \left(r(s_h,a_h) - \lambda \log \pi_\theta(a_h|s_h)\right) \mid \theta, s_0 = s, a_0 = a\right],
\end{aligned}$$

where we have replaced H' by ∞ since we use the indicator function $\mathbb{1}$ such that the summand for $h \geq H'$ is null. In addition, by the law of total expectation, we have

$$\begin{aligned}
&\mathbb{E}\left[\hat{Q}_\lambda^\theta(s,a) \mid \theta, s, a\right] \\
&\quad = \mathbb{E}_{H'}\left[\mathbb{E}_\tau\left[r(s_0,a_0) + \sum_{h=1}^{\infty} \mathbb{1}_{H' \geq h \geq 0} \gamma^{h/2} \cdot \left(r(s_h,a_h) - \lambda \log \pi_\theta(a_h|s_h)\right) \mid \theta, s_0 = s, a_0 = a, H'\right]\right],
\end{aligned} \tag{18}$$

where the trajectory τ equal to $\{s_0, a_0, s_1, a_1, \ldots\}$. The inner expectation over τ can be written as

$$\begin{aligned}
&\mathbb{E}_\tau\left[r(s_0,a_0) + \sum_{h=1}^{\infty} \mathbb{1}_{H' \geq h \geq 0} \gamma^{h/2} \cdot \left(r(s_h,a_h) - \lambda \log \pi_\theta(a_h|s_h)\right) \mid \theta, s_0 = s, a_0 = a, H'\right] \\
&\quad = \sum_\tau \left[r(s_0,a_0) + \sum_{h=1}^{\infty} \mathbb{1}_{H' \geq h \geq 0} \gamma^{h/2} \cdot \left(r(s_h,a_h) - \lambda \log \pi_\theta(a_h|s_h)\right)\right] \cdot \mathbb{P}(\tau) \mid \theta, s_0 = s, a_0 = a, H' \\
&\quad = r(s_0,a_0) + \sum_\tau \sum_{h=1}^{\infty} \left[\mathbb{1}_{H' \geq h \geq 0} \gamma^{h/2} \cdot \left(r(s_h,a_h) - \lambda \log \pi_\theta(a_h|s_h)\right)\right] \cdot \mathbb{P}(\tau) \mid \theta, s_0 = s, a_0 = a, H'.
\end{aligned} \tag{19}$$

By the definition of the probability over the sample trajectory $\mathbb{P}(\tau)$, for every $h \in \{0, 1, 2, \ldots\}$, it holds that

$$\begin{aligned}
&|(r(s_h, a_h) - \lambda \log \pi_\theta(a_h|s_h)) \cdot \mathbb{P}(\tau)| \\
&\quad = |(r(s_h, a_h) - \lambda \log \pi_\theta(a_h|s_h)) \cdot \pi_\theta(a_h|s_h) \cdot \mathbb{P}(s_1|s_0, a_0) \cdot \pi_\theta(a_1|s_1) \ldots \\
&\quad \cdot \mathbb{P}(s_h|s_{h-1}, a_{h-1}) \cdot \mathbb{P}(s_{h+1}|s_h, a_h) \ldots \mathbb{P}(s_{H'}|s_{H'-1}, a_{H'-1}) \cdot \cdot \pi_\theta(a_{H'}|s_{H'})| \\
&\quad \leq \bar{r} + \frac{\lambda}{e}.
\end{aligned}$$

where the last inequality follows from $\mathbb{P}(s'|s, a) \leq 1$, $\pi_\theta(a|s) \leq 1$ for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$ together with $|x \log x| \leq \frac{1}{e}$ for $x \in [0, 1]$. Thus, for each trajectory τ and $N > 0$, we have

$$\sum_{h=1}^{N} \left[\mathbb{1}_{H' \geq h \geq 0} \gamma^{h/2} \cdot (r(s_h, a_h) - \lambda \log \pi_\theta(a_h|s_h))\right] \cdot \mathbb{P}(\tau) \leq \frac{1}{1 - \gamma^{1/2}} \left(\bar{r} + \frac{\lambda}{e}\right). \tag{20}$$

Since left-hand side of (20) is non-decreasing and the limit as $N \to \infty$ exists, by the Monotone Convergence Theorem, we can interchange the limit with the summation over the trajectory τ in (19) as follows:

$$\begin{aligned}
&\mathbb{E}_\tau \left[r(s_0, a_0) + \sum_{h=1}^{\infty} \mathbb{1}_{H' \geq h \geq 0} \gamma^{h/2} \cdot (r(s_h, a_h) - \lambda \log \pi_\theta(a_h|s_h)) \mid \theta, s_0 = s, a_0 = a, H' \right] \\
&\quad = r(s_0, a_0) + \sum_{h=1}^{\infty} \sum_{\tau} \left[\mathbb{1}_{H' \geq h \geq 0} \gamma^{h/2} \cdot (r(s_h, a_h) - \lambda \log \pi_\theta(a_h|s_h))\right] \cdot \mathbb{P}(\tau) \mid \theta, s_0 = s, a_0 = a, H' \\
&\quad = r(s_0, a_0) + \sum_{h=1}^{\infty} \mathbb{E}_\tau \left[\mathbb{1}_{H' \geq h \geq 0} \gamma^{h/2} (r(s_h, a_h) - \lambda \log \pi_\theta(a_h|s_h)) \mid \theta, s_0 = s, a_0 = a, H'\right].
\end{aligned}$$

In addition, for every $N > 0$, we have

$$\begin{aligned}
&r(s_0, a_0) + \sum_{h=1}^{N} \mathbb{E}_\tau \left[\mathbb{1}_{H' \geq h \geq 0} \gamma^{h/2} (r(s_h, a_h) - \lambda \log \pi_\theta(a_h|s_h)) \mid \theta, s_0 = s, a_0 = a, H'\right] \qquad (21) \\
&\quad \leq r(s_0, a_0) + \gamma^{1/2} \sum_{h=0}^{\infty} \mathbb{E}_\tau \left[\gamma^{(h+1)/2} (r(s_h, a_h) - \lambda \log \pi_\theta(a_h|s_h)) \mid \theta, s_0 = s, a_0 = a, H'\right] \\
&\quad \leq r(s_0, a_0) + \gamma^{1/2} \mathbb{E}_{s_1} \left[V^\theta_{\lambda, \gamma/2}(s_1) \mid s_0, a_0\right] \\
&\quad \leq \bar{r} + \frac{\gamma/2(\bar{r} + \lambda \log |\mathcal{A}|)}{1 - \gamma/2} \\
&\quad \leq \frac{\bar{r} + \lambda \log |\mathcal{A}|}{1 - \gamma/2},
\end{aligned}$$

where the third inequality is due to the boundedness of the enrtopy-regularized value function in Lemma 3.1. Furthermore, since (21) is non-decreasing and the limit as $N \to \infty$ exists, by the Monotone Convergence Theorem, we can interchange the limit with the outer-expectation over H' in (18) as follows:

$$\begin{aligned}
&\mathbb{E}\left[\hat{Q}^\theta_\lambda(s, a) \mid \theta, s, a\right] \qquad (22) \\
&\quad = r(s_0, a_0) + \lim_{N \to \infty} \mathbb{E}_{H'} \left[\mathbb{E}_\tau \left[\sum_{h=1}^{N} \mathbb{1}_{H' \geq h \geq 0} \gamma^{h/2} \cdot (r(s_h, a_h) - \lambda \log \pi_\theta(a_h|s_h)) \mid \theta, s_0 = s, a_0 = a, H'\right]\right] \qquad (23) \\
&\quad = r(s_0, a_0) + \lim_{N \to \infty} \sum_{h=1}^{N} \left[\mathbb{E}_\tau \left[\mathbb{E}_{H'} \left[\mathbb{1}_{H' \geq h \geq 0}\right] \gamma^{h/2} \cdot (r(s_h, a_h) - \lambda \log \pi_\theta(a_h|s_h)) \mid \theta, s_0 = s, a_0 = a\right]\right] \qquad (24) \\
&\quad = r(s_0, a_0) + \lim_{N \to \infty} \sum_{h=1}^{N} \left[\mathbb{E}_\tau \left[\gamma^h \cdot (r(s_h, a_h) - \lambda \log \pi_\theta(a_h|s_h)) \mid \theta, s_0 = s, a_0 = a\right]\right] \qquad (25) \\
&\quad = r(s_0, a_0) + \mathbb{E}_\tau \left[\sum_{h=1}^{\infty} \left[\gamma^h \cdot (r(s_h, a_h) - \lambda \log \pi_\theta(a_h|s_h)) \mid \theta, s_0 = s, a_0 = a\right]\right] \qquad (26) \\
&\quad = Q^\theta_\lambda(s, a) \qquad (27)
\end{aligned}$$

where we have also used the fact that H' is drawn independently from the trajectory τ in the first equality, the fact that $H' \sim \text{Geom}(1 - \gamma^{1/2})$ and thus $\mathbb{E}_{H'}\left[\mathbb{1}_{H' \geq h \geq 0}\right] = \gamma^{t/2}$ in the second equality, and the interchangeability

between the limit and the summation over the trajectory τ in the third equality. This completes the proof of the unbiasedness of $\hat{Q}^\theta_\lambda(s,a)$.

Similar logic allows us to establish that $\hat{V}^\theta_\lambda(s)$ is an unbiased estimate of $V^\theta_\lambda(s)$, i.e.,

$$\mathbb{E}\left[\hat{V}^\theta_\lambda(s) \mid \theta, s\right] = V^\theta_\lambda(s), \quad \forall s \in \mathcal{S},\ \theta \in \mathbb{R}^d,$$

where the expectation is taken along the trajectory as well as with respect to the random horizon $H' \sim \text{Geom}\left(1-\gamma^{1/2}\right)$. Therefore, if $s' \sim \mathbb{P}(\cdot \mid s,a)$ and $a' \sim \pi_\theta(\cdot \mid s')$, we have

$$\mathbb{E}\left[\hat{A}^\theta_\lambda(s,a)\right] = \mathbb{E}\left[\hat{Q}^\theta_\lambda(s,a) - \lambda \log \pi_\theta(s \mid a) - \hat{V}^\theta_\lambda(s)\right] = A^\theta_\lambda(s,a) \tag{28}$$

That is, $\hat{A}^\theta_\lambda(s,a)$ is an unbiased estimate of the advantage function $A^\theta_\lambda(s,a)$.

Now, we are ready to show unbiasedness of the stochastic gradients $\hat{\nabla} V^\theta_\lambda(\rho)$. It follows from Lemma 3.2 that

$$\begin{aligned}
\mathbb{E}[\hat{\nabla} V^\theta_\lambda(\rho) \mid \theta] &= \mathbb{E}_{H,(s_H,a_H)}\left\{\mathbb{E}_{H',\left(s'_{1:H'},a'_{1:H'}\right)}\left[\hat{\nabla} V^\theta_\lambda(\rho) \mid \theta, s'_0 = s_H, a'_0 = a_H\right] \mid \theta\right\} \\
&= \mathbb{E}_{H,(s_H,a_H)}\left(\mathbb{E}_{H',\left(s'_{1:H'},a'_{1:H'}\right)}\left\{\frac{1}{1-\gamma}\boldsymbol{e}_{s'_0,a'_0} \cdot \hat{A}^\theta_\lambda(s'_0,a'_0)\cdot \mid \theta, s'_0 = s_H, a'_0 = a_H, \right\} \mid \theta\right) \\
&= \mathbb{E}_{H,(s_H,a_H)}\left\{\frac{1}{1-\gamma}\cdot \boldsymbol{e}_{s_H,a_H} \cdot A^\theta_\lambda(s_H,a_H) \mid \theta\right\}.
\end{aligned}$$

where we have used (28) in the last equality. By using the identity function $\mathbb{1}_{h=H}$, the above expression can be further written as

$$\mathbb{E}[\hat{\nabla} V^\theta_\lambda(\rho) \mid \theta] = \frac{1}{1-\gamma}\cdot \mathbb{E}_{H,(s_H,a_H)}\left\{\sum_{h=0}^{\infty} \mathbb{1}_{h=H}\cdot \boldsymbol{e}_{s_H,a_H}\cdot A^\theta_\lambda(s_H,a_H) \mid \theta\right\} \tag{29}$$

Since $\sum_{h=0}^{N} \mathbb{1}_{h=H}\cdot \boldsymbol{e}_{s_H,a_H}\cdot A^\theta_\lambda(s_H,a_H)$ is uniformly bounded by the boundedness of A^θ_λ for every $N>0$ and non-decreasing with respect to N, we can interchange the limit and the expectation in (29) by the Monotone Convergence Theorem to obtain

$$\begin{aligned}
\mathbb{E}[\hat{\nabla} V^\theta_\lambda(\rho) \mid \theta] &= \sum_{h=0}^{\infty} \frac{\mathbb{P}(H=h)}{1-\gamma}\cdot \mathbb{E}_{H,(s_H,a_H)}\left\{\cdot \boldsymbol{e}_{s_H,a_H}\cdot A^\theta_\lambda(s_H,a_H) \mid \theta\right\} \\
&= \sum_{h=0}^{\infty} \gamma^h \cdot \mathbb{E}_{(s_h,a_h)}\left\{\boldsymbol{e}_{s_h,a_h}\cdot A^\theta_\lambda(s_h,a_h) \mid \theta\right\} \\
&= \sum_{h=0}^{\infty}\gamma^h \sum_{s\in\mathcal{S},a\in\mathcal{A}} \mathbb{P}(s_h=s,a_h=a|s_0\sim\rho,\theta)\cdot \boldsymbol{e}_{s,a}\cdot A^\theta_\lambda(s,a) \\
&= \sum_{s\in\mathcal{S},a\in\mathcal{A}} \boldsymbol{e}_{s,a}\cdot A^\theta_\lambda(s,a)\cdot \sum_{h=0}^{\infty}\gamma^h \mathbb{P}(s_h=s,a_h=a|s_0\sim\rho,\theta) \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_\theta}_\rho, a\sim\pi_\theta(\cdot|s)}\left[\boldsymbol{e}_{s,a}A^\theta_\lambda(s,a)\right].
\end{aligned}$$

where the second equality is due to the fact that $H \sim \text{Geom}(1-\gamma)$ and thus $\mathbb{P}(h=H) = (1-\gamma)\gamma^h$, and the forth equality is due to the linearity of the integral and the finiteness of the state and action spaces. This completes the proof of unbiasedness of $\hat{\nabla} V^\theta_\lambda(\rho)$. □

8.5 Proof of Lemma 3.5

Proof. We first note that the policy gradient estimator $\hat{\nabla} V^\theta_\lambda(\rho)$ can be decomposed as:

$$\hat{A}^\theta_\lambda(s_H,a_H) = \sum_{i=0}^{H'}\gamma^{i/2}\left(r(s'_i,a'_i) - \lambda\log\pi_\theta(a'_i|s'_i)\right) - \sum_{j=0}^{H''}\gamma^{j/2}\left(r(s''_j,a''_j) - \lambda\log\pi_\theta(a''_j|s''_j)\right) \tag{30}$$

$$= \sum_{i=0}^{H'}\gamma^{i/2} r(s'_i,a'_i) - \sum_{i=0}^{H'}\gamma^{i/2}\lambda\log\pi_\theta(a'_i|s'_i) - \sum_{j=0}^{H''}\gamma^{j/2} r(s''_j,a''_j) + \sum_{j=0}^{H''}\gamma^{j/2}\lambda\log\pi_\theta(a''_j|s''_j) \tag{31}$$

where $H \sim \text{Geom}(1-\gamma), H' \sim \text{Geom}(1-\gamma^1/2), H'' \sim \text{Geom}(1-\gamma^1/2)$, $(s^H, a^H) \sim \nu_\rho^{\pi_\theta}(s,a), s_0' = s_0'' = s_H, a_0' = a_H$. To streamline the presentation, we introduce the following notations:

$$g_1(s_H, a_H) = \sum_{i=0}^{H'} \gamma^{i/2} r(s_i', a_i'), \quad g_2(s_H) = \sum_{j=0}^{H''} \gamma^{j/2} r(s_j'', a_j''), \tag{32}$$

$$g_3(s_H, a_H) = \sum_{i=0}^{H'} \gamma^{i/2} \lambda \log \pi_\theta(a_i'|s_i'), \quad g_4(s_H) = \sum_{j=0}^{H''} \gamma^{j/2} \lambda \log \pi_\theta(a_j''|s_j''). \tag{33}$$

Then, the policy gradient estimator $\hat{\nabla} V_\lambda^\theta(\rho)$ can be decomposed as:

$$\hat{\nabla} V_\lambda^\theta(\rho) = \frac{1}{1-\gamma} \boldsymbol{e}_{s_H, a_H} \cdot \left(g_1(s_H, a_H) - g_2(s_H) - g_3(s_H, a_H) + g_4(s_H) \right). \tag{34}$$

By the definition of the variance and the Cauchy-Schwarz inequality, we have

$$\begin{aligned} \text{Var}(V_\lambda^\theta(\rho)) =& \frac{1}{(1-\gamma)^2} \text{Var}\left(g_1(s_H, a_H) - g_2(s_H) - g_3(s_H, a_H) + g_4(s_H) \right) && (35) \\ \le& \frac{4}{(1-\gamma)^2} \left(\text{Var}\left(g_1(s_H, a_H) \right) + \text{Var}\left(g_2(s_H) \right) \right. && (36) \\ &+ \left. \text{Var}\left(g_3(s_H, a_H) \right) + \text{Var}\left(g_4(s_H) \right) \right). && (37) \end{aligned}$$

Since $g_1(s,a)$ and $g_2(s)$ are uniformly bounded, i.e., $\|g_1(s,a)\| \le \frac{\bar{r}}{1-\gamma^{1/2}}$ and $\|g_2(s)\| \le \frac{\bar{r}}{1-\gamma^{1/2}}$ for all $s \in \mathcal{S}, a \in \mathcal{A}$, we must have

$$\text{Var}\left(g_1(s_H, a_H) \right) \le \frac{\bar{r}^2}{(1-\gamma^{1/2})^2}, \quad \text{Var}\left(g_2(s_H) \right) \le \frac{\bar{r}^2}{(1-\gamma^{1/2})^2}.$$

Then, it remains to prove the bounded variance of g_3 and g_4. Firstly, it can be seen that

$$\begin{aligned} \|g_3\|^2 \le& \lambda^2 \left(\sum_{i=0}^{H'} \gamma^{i/2} \log \pi_\theta(a_i'|s_i') \right)^2 \\ =& \lambda^2 \left(\sum_{i=0}^{H'} \gamma^{i/4} \gamma^{i/4} \log \pi_\theta(a_i'|s_i') \right)^2 \\ \le& \lambda^2 \left(\sum_{i=0}^{H'} \gamma^{i/2} \right) \left(\sum_{i=0}^{H'} \gamma^{i/2} \left(\log \pi_\theta(a_i'|s_i') \right)^2 \right) \\ \le& \frac{\lambda^2}{1-\gamma^{1/2}} \left(\sum_{i=0}^{H'} \gamma^{i/2} \left(\log \pi_\theta(a_i'|s_i') \right)^2 \right), \end{aligned}$$

where the second inequality is due to the Cauchy-Schwarz inequality. By fixing the state action pair (s_H, a_H) and the horizon H' for now and taking expectation of g_3 only over the sample trajectory $\tau' = \{s_0', a_0', \ldots, s_H', a_H'\}$, it holds that

$$\mathbb{E}_{\tau' \sim p(\tau'|\theta)} \left[\|g_3\|^2 \right] \le \frac{\lambda^2}{1-\gamma^{1/2}} \sum_{i=0}^{H'} \gamma^{i/2} \mathbb{E}_{\tau' \sim p(\tau'|\theta)} \left[\left(\log \pi_\theta(a_i'|s_i') \right)^2 \right]. \tag{38}$$

Since the realizations of a_i' and s_i' do not depend on the randomness in $s_{i+1}', a_{i+1}', \ldots, s_H'$, we have

$$\begin{aligned} & \mathbb{E}_{\tau' \sim p(\tau'|\theta)} \left[\left(\log \pi_\theta(a_i'|s_i') \right)^2 \right] \\ =& \mathbb{E}_{s_1' \sim p(\cdot|a_0', s_0') \ldots a_{H-1}' \sim \pi_\theta(\cdot|s_{H-1}'), s_H' \sim p(\cdot|s_{H-1}', a_{H-1}')} \left[\left(\log \pi_\theta(a_i'|s_i') \right)^2 \right] \\ =& \mathbb{E}_{s_1' \sim p(\cdot|a_0', s_0') \ldots a_i' \sim \pi_\theta(\cdot|s_i')} \left[\left(\log \pi_\theta(a_i'|s_i') \right)^2 \right] \\ =& \mathbb{E}_{s_1' \sim p(\cdot|a_0', s_0') \ldots s_i' \sim p(\cdot|a_{i-1}', s_{i-1}')} \left[\sum_{a_i' \in \mathcal{A}} \pi_\theta(a_i'|s_i') \left(\log \pi_\theta(a_i'|s_i') \right)^2 \right]. \end{aligned}$$

By checking the optimality conditions for the optimization problem

$$\max \sum_{i=1}^{n} x_i (\log x_i)^2 \quad \text{such that } \sum_{i=1}^{n} x_i = 1, \tag{39}$$

it can be concluded that the maximizer for the constrained problem (39) is $x_1 = x_2 = \ldots = x_n = \frac{1}{n}$ and the maximum solution is $(\log n)^2$. Thus, we have $\sum_{a_h \in \mathcal{A}} \pi_\theta(a_h|s_h) \left(\log \pi_\theta(a_h^i|s_h^i)\right)^2 \leq (\log|\mathcal{A}|)^2$ and

$$\begin{aligned}
&\mathbb{E}_{\tau' \sim p(\tau'|\theta)} \left[\left(\log \pi_\theta(a_h^i|s_h^i)\right)^2 \right] \\
&\quad \leq \mathbb{E}_{s_0 \sim \rho, a_0 \sim \pi_\theta(\cdot|s_0), s_1 \sim p(\cdot|a_0,s_0) \ldots s_h \sim p(\cdot|a_{h-1},s_{h-1})} \left[(\log|\mathcal{A}|)^2 \right] \\
&\quad = (\log|\mathcal{A}|)^2.
\end{aligned}$$

By substituting the above inequality into (38), we obtain that

$$\begin{aligned}
\mathbb{E}_{\tau' \sim p(\tau'|\theta)} \left[\|g_3\|^2 \right] &\leq \frac{\lambda^2}{1-\gamma^{1/2}} \sum_{i=0}^{H'} \gamma^{i/2} \mathbb{E}_{\tau' \sim p(\tau'|\theta)} \left[(\log \pi_\theta(a_i'|s_i'))^2 \right] \\
&\leq \frac{(\lambda \log|\mathcal{A}|)^2}{1-\gamma^{1/2}} \sum_{i=0}^{H'} \gamma^{i/2} \\
&\leq \frac{(\lambda \log|\mathcal{A}|)^2}{(1-\gamma^{1/2})^2},
\end{aligned}$$

for every $H' > 0$. By taking expectation of g_3 over the state action pair (s_H, a_H) and the horizon H', it yields that

$$\mathbb{E}\left[\|g_3\|^2 \right] \leq \frac{(\lambda \log|\mathcal{A}|)^2}{(1-\gamma^{1/2})^2},$$

which further implies that $\mathrm{Var}\left[\|g_3\|^2 \right] \leq \frac{(\lambda \log|\mathcal{A}|)^2}{(1-\gamma^{1/2})^2}$. Similarly, we can bound the variance of g_4 as $\mathrm{Var}\left[\|g_4\|^2 \right] \leq \frac{(\lambda \log|\mathcal{A}|)^2}{(1-\gamma^{1/2})^2}$. Finally, through (35), we obtain

$$\mathrm{Var}(V_\lambda^\theta(\rho)) \leq \frac{8}{(1-\gamma)^2} \left(\frac{\bar{r}^2 + (\lambda \log|\mathcal{A}|)^2}{(1-\gamma^{1/2})^2} \right).$$

This completes the proof. □

8.6 Proof of Lemma 3.6

Proof. By definition, we have

$$\begin{aligned}
\mathbb{E}[\hat{\nabla} V_\lambda^{\theta,H}(\rho)] - \nabla V_\lambda^\theta(\rho) =& \mathbb{E}\left[\sum_{h=0}^{\infty} \nabla \log \pi_\theta(a_h|s_h) \left(\sum_{j=h}^{\infty} \gamma^j \left(r_j(s_j,a_j) - \lambda \log \pi_\theta(a_j|s_j) \right) \right) \right. \\
&- \sum_{h=0}^{H-1} \nabla \log \pi_\theta(a_h|s_h) \left(\sum_{j=h}^{H-1} \gamma^j \left(r_h(s_j,a_j) - \lambda \log \pi_\theta(a_j|s_j) \right) \right) \\
&\left. + \lambda \sum_{h=H}^{\infty} -\gamma^t \nabla \log \pi_\theta(a_h,s_h) \right] \\
=& \mathbb{E}\left[\sum_{h=0}^{H-1} \nabla \log \pi_\theta(a_h|s_h) \left(\sum_{j=H}^{\infty} \gamma^j \left(r_j(s_j,a_j) - \lambda \log \pi_\theta(a_j|s_j) \right) \right) \right. \\
&+ \sum_{h=H}^{\infty} \nabla \log \pi_\theta(a_h|s_h) \left(\sum_{j=h}^{\infty} \gamma^j \left(r_j(s_j,a_j) - \lambda \log \pi_\theta(a_j|s_j) \right) \right) \\
&\left. + \lambda \sum_{h=H}^{\infty} -\gamma^t \nabla \log \pi_\theta(a_h,s_h) \right]
\end{aligned}$$

Then, by the Cauchy-Schwarz inequality and the triangle inequality, we obtain

$$\begin{aligned}\left\|\mathbb{E}[\hat{\nabla} V_\lambda^{\theta,H}(\rho)] - \nabla V_\lambda^\theta(\rho)\right\|_2 \leq & \left\|\mathbb{E}\left[\sum_{h=0}^{H-1} \nabla \log \pi_\theta(a_h|s_h)\left(\sum_{j=H}^{\infty} \gamma^j \left(r_j(s_j,a_j) - \lambda \log \pi_\theta(a_j|s_j)\right)\right)\right]\right\| \\ & + \left\|\mathbb{E}\left[\sum_{h=H}^{\infty} \nabla \log \pi_\theta(a_h|s_h)\left(\sum_{j=h}^{\infty} \gamma^j \left(r_j(s_j,a_j) - \lambda \log \pi_\theta(a_j|s_j)\right)\right)\right]\right\| \\ & + \left\|\mathbb{E}\left[\lambda \sum_{h=H}^{\infty} -\gamma^t \nabla \log \pi_\theta(a_h,s_h)\right]\right\| \\ \leq & \mathbb{E}\left[\sum_{h=0}^{H-1} \|\nabla \log \pi_\theta(a_h|s_h)\|\left(\sum_{j=H}^{\infty} \gamma^j \left(r_j(s_j,a_j) - \lambda \log \pi_\theta(a_j|s_j)\right)\right)\right] \\ & + \mathbb{E}\left[\sum_{h=H}^{\infty} \|\nabla \log \pi_\theta(a_h|s_h)\|\left(\sum_{j=h}^{\infty} \gamma^j \left(r_j(s_j,a_j) - \lambda \log \pi_\theta(a_j|s_j)\right)\right)\right] \\ & + \mathbb{E}\left[\lambda \sum_{h=H}^{\infty} \gamma^t \|\nabla \log \pi_\theta(a_h,s_h)\|\right]\end{aligned}$$

Since $\|\nabla \log \pi_\theta(a|s)\|_2 \leq 2$ for all $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, it holds that

$$\begin{aligned}\left\|\mathbb{E}[\hat{\nabla} V_\lambda^{\theta,H}(\rho)] - \nabla V_\lambda^\theta(\rho)\right\|_2 \leq & 2\mathbb{E}\left[\sum_{h=0}^{H-1}\left(\sum_{j=H}^{\infty} \gamma^j \left(r_j(s_j,a_j) - \lambda \log \pi_\theta(a_j|s_j)\right)\right)\right] && (40) \\ & + 2\mathbb{E}\left[\sum_{h=H}^{\infty}\left(\sum_{j=h}^{\infty} \gamma^j \left(r_j(s_j,a_j) - \lambda \log \pi_\theta(a_j|s_j)\right)\right)\right] && (41) \\ & + 2\mathbb{E}\left[\lambda \sum_{h=H}^{\infty} \gamma^t\right]. && (42)\end{aligned}$$

For the term in (40), we can rewrite it as

$$2\mathbb{E}_\tau\left[\sum_{h=0}^{H-1}\left(\sum_{j=H}^{\infty} \gamma^j \left(r_j(s_j,a_j) - \lambda \log \pi_\theta(a_j|s_j)\right)\right)\right] = \sum_\tau \left[\sum_{h=0}^{H-1}\left(\sum_{j=H}^{\infty} \gamma^j \left(r_j(s_j,a_j) - \lambda \log \pi_\theta(a_j|s_j)\right)\right)\right] \cdot \mathbb{P}(\tau)$$

Then, by following the arguments in (20) and the Monotone Convergence Theorem, we can interchange the limit with the summation over the trajectory τ in (40) as follows:

$$2\mathbb{E}_\tau\left[\sum_{h=0}^{H-1}\left(\sum_{j=H}^{\infty} \gamma^j \left(r_j(s_j,a_j) - \lambda \log \pi_\theta(a_j|s_j)\right)\right)\right] = 2\sum_{h=0}^{H-1}\left(\sum_{j=H}^{\infty} \gamma^j \mathbb{E}_\tau\left[r_j(s_j,a_j) - \lambda \log \pi_\theta(a_j|s_j)\right]\right).$$

Due to $-\sum_a \pi(a|s) \cdot \log \pi(a|s) \leq \log|\mathcal{A}|$, the term in (40) can be upper bounded as

$$\begin{aligned}2\mathbb{E}_\tau\left[\sum_{h=0}^{H-1}\left(\sum_{j=H}^{\infty} \gamma^j \left(r_j(s_j,a_j) - \lambda \log \pi_\theta(a_j|s_j)\right)\right)\right] &\leq 2(\bar{r} + \lambda \log|\mathcal{A}|)\sum_{h=0}^{H-1}\left(\sum_{j=H}^{\infty} \gamma^j\right) \\ &\leq \frac{2(\bar{r} + \lambda \log|\mathcal{A}|)H\gamma^H}{1-\gamma}.\end{aligned}$$

Similarly, we can interchange the limit with the summation over the trajectory τ in (41) and upper bound it as

$$\begin{aligned}2\mathbb{E}\left[\sum_{h=H}^{\infty}\left(\sum_{j=h}^{\infty} \gamma^j \left(r_j(s_j,a_j) - \lambda \log \pi_\theta(a_j|s_j)\right)\right)\right] &\leq 2(\bar{r} + \lambda \log|\mathcal{A}|)\sum_{h=H}^{\infty}\sum_{j=h}^{\infty} \gamma^j \\ &\leq \frac{2(\bar{r} + \lambda \log|\mathcal{A}|)\gamma^H}{(1-\gamma)^2}.\end{aligned}$$

For the term in (42), it can be easily bounded as

$$2\mathbb{E}\left[\lambda \sum_{h=H}^{\infty} \gamma^t\right] \leq \frac{2\lambda\gamma^H}{1-\gamma}.$$

This completes the proof. □

8.7 Proof of Lemma 3.7

Proof. By the definition of the variance and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathrm{Var}(\hat{\nabla} V_\lambda^{\theta,H}(\rho)) =& \mathbb{E}\left[\left(g_1(\tau^H|\theta,\rho) + g_2(\tau^H|\theta,\rho)\right) + g_3(\tau^H|\theta,\rho))\right. \\
& \left. -\mathbb{E}[g_1(\tau^H|\theta,\rho)] - \mathbb{E}[g_2(\tau^H|\theta,\rho)] - \mathbb{E}[g_3(\tau^H|\theta,\rho)]\right)^2\Big] \\
\leq& 3\mathbb{E}[(g_1(\tau^H|\theta,\rho) - \mathbb{E}[g_1(\tau^H|\theta,\rho)])^2] + 3\mathbb{E}[(g_2(\tau^H|\theta,\rho)) - \mathbb{E}[g_2(\tau^H|\theta,\rho))])^2] \\
& + 3\mathbb{E}[(g_3(\tau^H|\theta,\rho)) - \mathbb{E}[g_3(\tau^H|\theta,\rho))])^2] \\
=& 3\left(\mathrm{Var}(g_1(\tau^H|\theta,\rho)) + \mathrm{Var}(g_2(\tau^H|\theta,\rho)) + \mathrm{Var}(g_3(\tau^H|\theta,\rho))\right). \qquad (43)
\end{aligned}$$

As shown in Lemma 4.2 of Yuan et al. (2021), the fact that $\|\nabla \log \pi_\theta(a|s)\|_2 \leq 2$ for all $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ directly implies that $\mathrm{Var}(g_1(\tau^H|\theta,\rho)) \leq \frac{4\bar{r}^2}{(1-\gamma)^4}$ for all $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

Similar, due to $\|\nabla \log \pi_\theta(a|s)\|_2 \leq 2$ for all $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, it can be verified that $\mathrm{Var}(g_3(\tau^H|\theta,\rho)) \leq \frac{4\lambda}{(1-\gamma)^2}$ for all $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

Then, it remains to prove the bounded variance of g_2. Firstly, it can be observed that

$$\begin{aligned}
\|g_2\| =& \lambda \left\| \sum_{h=0}^{H-1} \left(\sum_{j=0}^{h} \nabla \log \pi_\theta(a_j^i|s_j^i) \right) \left(-\gamma^h \log \pi_\theta(a_h^i|s_h^i)\right) \right\| \\
\leq& -\lambda \sum_{h=0}^{H-1} \left(\sum_{j=0}^{h} \left\| \nabla \log \pi_\theta(a_j^i|s_j^i) \right\| \right) \gamma^h \log \pi_\theta(a_h^i|s_h^i) \\
\leq& -2\lambda \sum_{h=0}^{H-1} (h+1)\gamma^h \log \pi_\theta(a_h^i|s_h^i).
\end{aligned}$$

where the first inequality is due to the triangle inequality and the second inequality is due to $\left\|\nabla \log \pi_\theta(a_j^i|s_j^i)\right\| \leq 2$. Then, by taking the suqre of $\|g_2\|$, we obtain

$$\begin{aligned}
\|g_2\|^2 \leq& 4\lambda^2 \left(\sum_{h=0}^{H-1} (h+1)\gamma^h \log \pi_\theta(a_h^i|s_h^i) \right)^2 \\
=& 4\lambda^2 \left(\sum_{h=0}^{H-1} (h+1)\sqrt{\gamma^h}\sqrt{\gamma^h} \log \pi_\theta(a_h^i|s_h^i) \right)^2 \\
\leq& 4\lambda^2 \left(\sum_{h=0}^{H-1} (h+1)^2\gamma^h \right) \left(\sum_{h=0}^{H-1} \gamma^h \left(\log \pi_\theta(a_h^i|s_h^i)\right)^2 \right) \\
=& 4\lambda^2 \left(\sum_{h=0}^{H-1} (h^2+2h+1)\gamma^h \right) \left(\sum_{h=0}^{H-1} \gamma^h \left(\log \pi_\theta(a_h^i|s_h^i)\right)^2 \right) \\
\leq& 4\lambda^2 \left(\frac{\gamma^2+\gamma}{(1-\gamma)^3} + \frac{2\gamma}{(1-\gamma)^2} + \frac{1}{1-\gamma} \right) \left(\sum_{h=0}^{H-1} \gamma^h \left(\log \pi_\theta(a_h^i|s_h^i)\right)^2 \right) \\
=& 4\lambda^2 \left(\frac{\gamma+1}{(1-\gamma)^3} \right) \left(\sum_{h=0}^{H-1} \gamma^h \left(\log \pi_\theta(a_h^i|s_h^i)\right)^2 \right)
\end{aligned}$$

where the second inequality is due to the Cauchy-Schwarz inequality and the last inequality is due to $\sum_{h=0}^{\infty} h^2\gamma^h = \frac{\gamma^2+\gamma}{(1-\gamma)^3}$, $\sum_{h=0}^{\infty} h\gamma^h = \frac{\gamma}{(1-\gamma)^2}$ and $\sum_{h=0}^{\infty} \gamma^h = \frac{1}{1-\gamma}$.

By taking expectation of g_2 over the sample trajectory τ, it holds that

$$\mathbb{E}_{\tau\sim p(\tau|\theta)}\left[\|g_2\|^2\right]\leq 4\lambda^2\left(\frac{\gamma+1}{(1-\gamma)^3}\right)\sum_{h=0}^{H-1}\gamma^h\mathbb{E}_{\tau\sim p(\tau|\theta)}\left[\left(\log\pi_\theta(a_h^i|s_h^i)\right)^2\right]. \tag{44}$$

Since the realizations of a_h^i and s_h^i do not depend on the randomness in $s_{h+1}, a_{h+1}, \ldots, s_H$, we have

$$\begin{aligned}
&\mathbb{E}_{\tau\sim p(\tau|\theta)}\left[\left(\log\pi_\theta(a_h^i|s_h^i)\right)^2\right]\\
=&\mathbb{E}_{s_0\sim\rho,a_0\sim\pi_\theta(\cdot|s_0),s_1\sim p(\cdot|a_0,s_0)\ldots a_{H-1}\sim\pi_\theta(\cdot|s_{H-1}),s_H\sim p(\cdot|s_{H-1},a_{H-1})}\left[\left(\log\pi_\theta(a_h^i|s_h^i)\right)^2\right]\\
=&\mathbb{E}_{s_0\sim\rho,a_0\sim\pi_\theta(\cdot|s_0),s_1\sim p(\cdot|a_0,s_0)\ldots a_h\sim\pi_\theta(\cdot|s_h)}\left[\left(\log\pi_\theta(a_h^i|s_h^i)\right)^2\right]\\
=&\mathbb{E}_{s_0\sim\rho,a_0\sim\pi_\theta(\cdot|s_0),s_1\sim p(\cdot|a_0,s_0)\ldots s_h\sim p(\cdot|a_{h-1},s_{h-1})}\left[\sum_{a_h\in\mathcal{A}}\pi_\theta(a_h|s_h)\left(\log\pi_\theta(a_h^i|s_h^i)\right)^2\right].
\end{aligned}$$

As proved earlier in Lemma 3.5, we know that the maximizer for the constrained problem (39) is $x_1 = x_2 = \cdots = x_n = \frac{1}{n}$ and the maximum solution is $(\log n)^2$. Thus, we have $\sum_{a_h\in\mathcal{A}}\pi_\theta(a_h|s_h)\left(\log\pi_\theta(a_h^i|s_h^i)\right)^2 \leq (\log|\mathcal{A}|)^2$ and

$$\begin{aligned}
\mathbb{E}_{\tau\sim p(\tau|\theta)}\left[\left(\log\pi_\theta(a_h^i|s_h^i)\right)^2\right]\leq&\mathbb{E}_{s_0\sim\rho,a_0\sim\pi_\theta(\cdot|s_0),s_1\sim p(\cdot|a_0,s_0)\ldots s_h\sim p(\cdot|a_{h-1},s_{h-1})}\left[(\log|\mathcal{A}|)^2\right]\\
=&(\log|\mathcal{A}|)^2.
\end{aligned} \tag{45}$$

By combining (44) and (45), we have

$$\begin{aligned}
\mathrm{Var}(g_2)\leq&\mathbb{E}_{\tau\sim p(\tau|\theta)}\left[\|g_2\|^2\right]\\
\leq&4\lambda^2\left(\frac{\gamma+1}{(1-\gamma)^3}\right)\sum_{h=0}^{H-1}\gamma^h\mathbb{E}_{\tau\sim p(\tau|\theta)}\left[\left(\log\pi_\theta(a_h^i|s_h^i)\right)^2\right]\\
\leq&4\lambda^2\left(\frac{\gamma+1}{(1-\gamma)^3}\right)\sum_{h=0}^{H-1}\gamma^h(\log|\mathcal{A}|)^2\\
\leq&\frac{8\lambda^2(\log|\mathcal{A}|)^2}{(1-\gamma)^4}.
\end{aligned}$$

Finally, by substituting $\mathrm{Var}(g_1), \mathrm{Var}(g_2)$ and $\mathrm{Var}(g_3)$ into (43), it holds that

$$\mathrm{Var}(\hat{\nabla}V_\lambda^{\theta,H}(\rho)) \leq \frac{12\lambda}{(1-\gamma)^2}+\frac{12\bar{r}^2+24\lambda^2(\log|\mathcal{A}|)^2}{(1-\gamma)^4}.$$

This completes the proof. □

9 Proof of Lemma 4.4

We first introduce some useful results before proceeding with the proof. The following result describes the asymptotic behavior of the true gradient when an unbiased gradient estimator with a bounded variance is used in the update rule.

Proposition 9.1 (Proposition 3 in Bertsekas and Tsitsiklis (2000)) *Consider the problem* $\max_{x\in\mathbb{R}^d} f(x)$*, where* $\mathbb{R}^d$ *denotes the d-dimensional Euclidean space. Let* $\{x_t\}_{t=0}^{\infty}$ *be a sequence generated by the iterative method* $\mathbf{x}_{t+1} = x_t + \eta_t(u_t + w_t)$*, where* η_t *is a deterministic positive step-size,* u_t *is an update direction, and* w_t *is a random noise term. Let* $\mathcal{F}_t$ *be an increasing sequence of* $\sigma-$*fields. Assume that*

1. *f is a continuously differentiable function and there exists a constant L such that*

$$\|\nabla f(x)-\nabla f(\bar{x})\| \leq L\|x-\bar{x}\|, \quad \forall x,\bar{x}\in\mathbb{R}^d.$$

2. x_t *and* u_t *are* $\mathcal{F}_t-$*measurable.*

3. *There exist positive scalars* c_1 *and* c_2 *such that*

$$c_1 \|\nabla f(x_t)\|^2 \leq \nabla f(x_t)^\top u_t, \quad \|u_t\| \leq c_2(1 + \|\nabla f(x_t)\|), \quad \forall t \in \{1, 2, \ldots\}.$$

4. *We have*

$$\mathbb{E}[w_t|\mathcal{F}_t] = 0, \quad \mathbb{E}[\|w_t\|^2 |\mathcal{F}_t] \leq A(1 + \|\nabla f(x_t)\|^2),$$

for all $t \in \{1, 2, \ldots\}$ *with probability 1, where* A *is a positive deterministic constant.*

5. *We have*

$$\sum_{t=1}^{\infty} \eta_t = \infty, \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty.$$

Then, either $f(x_t) \to \infty$ *or else* $f(x_t)$ *converges to a finite value and* $\lim_{t\to\infty} \nabla f(x_t) = 0$ *with probability* 1.

9.1 Proof of Lemma 4.4

Proof.

To prove Lemma 4.4, it suffices to check the conditions in Proposition 9.1 for the objective function $V_\lambda^\theta(\rho)$ and the update rule $\theta_{t+1} = \theta_t + \eta_t(u_t + w_t)$, where $u_t = \nabla V_\lambda^{\theta_t}(\rho)$ and $w_t = \hat{\nabla} V_\lambda^{\theta_t}(\rho) - \nabla V_\lambda^{\theta_t}(\rho)$.

1. From Lemma 3.3, we know that Condition 1 in 9.1 is satisfied with $L = \frac{8\bar{r}+\lambda(4+8\log|\mathcal{A}|)}{(1-\gamma)^3}$.
2. Condition 1 in 9.1 is satisfied by the definition of θ_t and $\nabla V_\lambda^\theta(\rho)$.
3. Condition 1 in 9.1 is satisfied with $c_1 = 1$ and $c_2 = 1$.
4. From Lemma 3.4 and 3.5, we know that Condition 4 in 9.1 is satisfied with $A = \frac{8}{(1-\gamma)^2}\left(\frac{\bar{r}^2+(\lambda\log|\mathcal{A}|)^2}{(1-\gamma^{1/2})^2}\right)$.
5. Condition 1 in 9.1 is satisfied by the definition of η_t.

In addition, it results from Lemma 3.1 wee know that the entropy-regularized value function $V_\lambda^\theta(\rho)$ is bounded. Thus, by Proposition 9.1, we must have $\lim_{t\to\infty} \nabla V_\lambda^{\theta_t}(\rho) = 0$ with probability 1. This completes the proof.

□

10 Proof of Theorem 4.5

We begin by introducing some helpful definitions. Let $\{\bar{\theta}_t\}_{t=1}^T$ denote the iterates of the algorithm with the exact PG (Algorithm 1) with $\eta_t \leq \frac{1}{2L}$ starting from the initial point θ_1. Let θ^*, which depends on the initial point θ_1, be the optimal solution that the Algorithm 1 will converge to. Then, by Lemma 4.2, there must exist a bounded constant $\bar{\Delta}$ such that $\left\|\bar{\theta}_t - \theta^*\right\|_2 \leq \bar{\Delta}$ for all $t = \{1, 2, \ldots\}$. Then, with a fair degree of hindsight and for some $\delta > 0$, we define the stopping time for the iterates $\{\theta_t\}_{t=1}^T$ as

$$\tau \coloneqq \min\left\{t \middle| \|\theta_t - \theta^*\|_2 > \left(1 + \frac{1}{\delta}\right)\bar{\Delta}\right\},$$

which is the index of the first iterate that exits the bounded region

$$\mathcal{G}_\delta^0 \coloneqq \left\{\theta : \|\theta - \theta^*\|_2 \leq \left(1 + \frac{1}{\delta}\right)\bar{\Delta}\right\}.$$

Furthermore, we define the constant

$$C_\delta^0 = \min_{\theta\in\mathcal{G}_\delta^0} C(\theta),$$

where $C(\theta)$ is defined in Lemma 4.1. Finally, we define $D(\theta_t) = \|\theta_t - \theta^*\|_2$.

Lemma 10.1 *Suppose that $f(x)$ is L-smooth. Given $0 < \eta_t \le \frac{1}{2L}$ for all $t \ge 1$, let $\{x_t\}_{t=1}^T$ be generated by a general update of the form $x_{t+1} = x_t + \eta_t u_t$ and let $e_t = u_t - \nabla f(x_t)$. We have*

$$f(x_{t+1}) \ge f(x_t) + \frac{\eta_t}{4}\|u_t\|_2^2 - \frac{\eta_t}{2}\|e_t\|_2^2.$$

Proof. Since $f(f)$ is L-smooth, one can write

$$\begin{aligned}
&f(x_{t+1}) - f(x_t) - \langle u_t, x_{t+1} - x_t\rangle \\
&\quad = f(x_{t+1}) - f(x_t) - \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \langle \sqrt{\eta_t}(\nabla f(x_t) - u_t), \frac{1}{\sqrt{\eta_t}}(x_{t+1} - x_t)\rangle \\
&\quad \ge -\frac{L}{2}\|x_{t+1} - x_t\|^2 - \frac{b\eta_t}{2}\|\nabla f(x_t) - u_t\|_2^2 - \frac{1}{2b\eta_t}\|x_{t+1} - x_t\|_2^2 \\
&\quad = (-\frac{1}{2b\eta_t} - \frac{L}{2})\|x_{t+1} - x_t\|_2^2 - \frac{b\eta_t}{2}\|e_t\|_2^2,
\end{aligned}$$

where the constant $b > 0$ is to be determined later. By the above inequality and the definition of x_{t+1}, we have

$$\begin{aligned}
f(x_{t+1}) &\ge f(x_t) + \langle u_t, x_{t+1} - x_t\rangle - (\frac{1}{2b\eta_t} + \frac{L}{2})\|x_{t+1} - x_t\|_2^2 - \frac{b\eta_t}{2}\|e_t\|_2^2 \\
&= f(x_t) + \eta_t\|u_t\|^2 - (\frac{\eta_t}{2b} + \frac{L\eta_t^2}{2})\|u_t\|_2^2 - \frac{b\eta_t}{2}\|e_t\|_2^2
\end{aligned}$$

By choosing $b = 1$ and using the fact that $0 < \eta_t \le \frac{1}{2L}$, we have

$$\begin{aligned}
f(x_{t+1}) &\ge f(x_t) + (\frac{\eta_t}{2} - \frac{L\eta_t^2}{2})\|u_t\|_2^2 - \frac{\eta_t}{2}\|e_t\|_2^2 \\
&\ge f(x_t) + \frac{\eta_t}{4}\|u_t\|_2^2 - \frac{\eta_t}{2}\|e_t\|_2^2.
\end{aligned}$$

This completes the proof. □

Lemma 10.2 *Let $e_t = \nabla V_\lambda^{\theta_t}(\rho) - u_t$, where $u_t = \frac{1}{B}\sum_{i=1}^B \hat{\nabla} V_\lambda^{\theta_t, i}(\rho)$ and $\hat{\nabla} V_\lambda^{\theta_t, i}(\rho)$ is an unbiased estimator of $\nabla V_\lambda^{\theta_t}(\rho)$. If $\eta_t = \eta \le \frac{1}{2L}$, then*

$$\mathbb{E}[D(\theta_T)\mathbf{1}_{\tau > T}] \le \left(1 - \frac{\eta C_\delta^0}{8}\right)^{T-1} D(\theta_1) + \frac{5\sigma^2}{8C_\delta^0 B}.$$

Proof. Let $\mathcal{F}_t$ denote the sigma field generated by the randomness up to iteration t. We define $\mathbb{E}^t \coloneqq \mathbb{E}[\cdot|\mathcal{F}_t]$ as the expectation operator conditioned on the sigma field $\mathcal{F}_t$. Since $\nabla V_\lambda^\theta(\rho)$ is L-smooth due to Lemma 3.3, it follows from Lemma 10.1 that

$$\begin{aligned}
\mathbb{E}^t[D(\theta_{t+1}) - D(\theta_t)]\mathbf{1}_{\tau > t} &= \mathbb{E}^t\left[V_\lambda^{\theta_t}(\rho) - V_\lambda^{\theta_{t+1}}(\rho)\right]\mathbf{1}_{\tau > t} \\
&\le \mathbb{E}^t\left[-\frac{\eta}{8}\|u_t\|_2^2 + \frac{3\eta}{4}\|e_t\|_2^2\right]\mathbf{1}_{\tau > t} \\
&\le \mathbb{E}^t\left[-\frac{\eta}{8}\left\|u_t - \nabla V_\lambda^{\theta_t}(\rho) + \nabla V_\lambda^{\theta_t}(\rho)\right\|_2^2 + \frac{3\eta}{4}\|e_t\|_2^2\right]\mathbf{1}_{\tau > t} \\
&= \mathbb{E}^t\left[-\frac{\eta}{8}\left\|u_t - \nabla V_\lambda^{\theta_t}(\rho)\right\|_2^2 - \frac{\eta}{8}\left\|\nabla V_\lambda^{\theta_t}(\rho)\right\|_2^2 + \frac{3\eta}{4}\|e_t\|_2^2\right]\mathbf{1}_{\tau > t} \\
&= \mathbb{E}^t\left[-\frac{\eta}{8}\left\|\nabla V_\lambda^{\theta_t}(\rho)\right\|_2^2 + \frac{5\eta}{8}\|e_t\|_2^2\right]\mathbf{1}_{\tau > t} \\
&\le \mathbb{E}^t\left[-\frac{\eta C(\theta_t)}{8}D(\theta_t) + \frac{5\eta}{8}\|e_t\|_2^2\right]\mathbf{1}_{\tau > t},
\end{aligned}$$

for every $\eta \le \frac{1}{2L}$, where the second inequality uses the fact that u_t is an unbiased estimator of $\nabla V_\lambda^{\theta_t}(\rho)$ and the last inequality is due to Lemma 4.1. We now consider two cases:

- Case 1: Assume that $\tau > t$, which implies that $\theta_t \in \mathcal{G}_\delta^0$ and $C(\theta_t) \geq C_\delta^0$. Then,

$$\mathbb{E}[D(\theta_{t+1})|\mathcal{F}_t] \leq \left(1 - \frac{\eta C_\delta^0}{8}\right) D(\theta_t) + \frac{5\eta}{8}\mathbb{E}\left[\|e_t\|_2^2|\mathcal{F}_t\right].$$

- Case 2: Assume that $\tau \leq t$ which leads to

$$\mathbb{E}[D(\theta_{t+1})|\mathcal{F}_t]\mathbf{1}_{\tau>t} = 0.$$

Now combining the above two cases yields the inequality

$$\begin{aligned}\mathbb{E}[D(\theta_{t+1})|\mathcal{F}_t]\mathbf{1}_{\tau>t} &\leq \left\{(1 - \frac{\eta C_\delta^0}{8})D(\theta_t) + \frac{5\eta}{8}\mathbb{E}\left[\|e_t\|_2^2|\mathcal{F}_t\right]\right\}\mathbf{1}_{\tau>t} \\ &\leq \left(1 - \frac{\eta C_\delta^0}{8}\right) D(\theta_t)\mathbf{1}_{\tau>t} + \frac{5\eta}{8}\mathbb{E}\left[\|e_t\|_2^2|\mathcal{F}_t\right].\end{aligned}$$

In addition, conditioning on $\mathcal{F}_t$ yields that

$$\mathbb{E}[D(\theta_{t+1})\mathbf{1}_{\tau>t+1}|\mathcal{F}_t] \leq \mathbb{E}[D(\theta_{t+1})\mathbf{1}_{\tau>t}|\mathcal{F}_t] = \mathbb{E}[D(\theta_{t+1})|\mathcal{F}_t]\mathbf{1}_{\tau>t},$$

where the last equality uses the fact that τ is a stopping time and the random variable $\mathbf{1}_{\tau>t}$ is determined completely by the sigma-field $\mathcal{F}_t$. Taking the expectations over the sigma-field $\mathcal{F}_t$ and then arguing inductively gives rise to

$$\begin{aligned}\mathbb{E}[D(\theta_{t+1})\mathbf{1}_{\tau>t+1}] &\leq \prod_{i=0}^{t}\left(1 - \frac{\eta C_\delta^0}{8}\right) D(\theta_1) + \sum_{i=0}^{t}\left(1 - \frac{\eta C_\delta^0}{8}\right)^i \frac{5\eta}{8}\mathbb{E}\left[\|e_i\|_2^2\right] \\ &\leq \left(1 - \frac{\eta C_\delta^0}{8}\right)^t D(\theta_1) + \frac{5\sigma^2}{C_\delta^0 B}.\end{aligned}$$

By setting $t + 1 = T$, we obtain that

$$\mathbb{E}[D(\theta_T)\mathbf{1}_{\tau>T}] \leq \left(1 - \frac{\eta C_\delta^0}{8}\right)^{T-1} D(\theta_1) + \frac{5\sigma^2}{C_\delta^0 B}.$$

This completes the proof. □

Lemma 10.3 *Let $e_t = \nabla V_\lambda^{\theta_t}(\rho) - u_t$, where $u_t = \frac{1}{B}\sum_{i=1}^{B}\hat{\nabla} V_\lambda^{\theta_t,i}(\rho)$ and $\hat{\nabla} V_\lambda^{\theta_t,i}(\rho)$ is an unbiased estimator of $\nabla V_\lambda^{\theta_t}(\rho)$. Then, it holds that*

$$\mathbb{P}(\tau \leq T) \leq \frac{\delta \cdot \eta \cdot T \cdot (1 + \eta L)^{T-1} \cdot \sigma}{\bar{\Delta} B}.$$

Proof. By the triangle inequality and the fact that the iterations of the algorithm with the exact policy gradient are bounded by $\bar{\Delta}$, we have

$$D(\theta_t) \leq \left\|\theta_t - \bar{\theta}_t\right\|_2 + \left\|\theta^* - \bar{\theta}_t\right\|_2 = \left\|\theta_t - \bar{\theta}_t\right\|_2 + \bar{\Delta}.$$

Using the update rule of the algorithm with the exact policy gradient $\nabla V_\lambda^{\bar{\theta}_i}(\rho)$ and the stochastic policy gradient $u_i = \frac{1}{B}\sum_{j=1}^{B}\hat{\nabla} V_\lambda^{\theta_i,j}(\rho)$, one can write

$$\begin{aligned}D(\theta_i) &= \left\|(\theta_1 + \sum_{i=1}^{t-1}\eta_i u_i) - (\theta_1 + \sum_{i=1}^{t-1}\eta_i \nabla V_\lambda^{\bar{\theta}_i}(\rho))\right\|_2 + \bar{\Delta} \\ &\leq \sum_{i=1}^{t-1}\eta_i \left\|u_i - \nabla V_\lambda^{\bar{\theta}_i}(\rho)\right\|_2 + \bar{\Delta} \\ &= \sum_{i=1}^{t-1}\eta_i \left\|u_i - \nabla V_\lambda^{\theta_i}(\rho) + \nabla V_\lambda^{\theta_i}(\rho) - \nabla V_\lambda^{\bar{\theta}_i}(\rho)\right\|_2 + \bar{\Delta} \\ &\leq \sum_{i=1}^{t-1}\eta_i \|e_i\|_2 + \sum_{i=1}^{t-1}\eta_i L \left\|\theta_i - \bar{\theta}_i\right\|_2 + \bar{\Delta}.\end{aligned}$$

By expanding $\left\|\theta_i - \bar{\theta}_i\right\|_2$ recursively, it can be concluded that

$$
\begin{aligned}
D(\theta_t) \leq & \sum_{i=1}^{t-1} \eta_i \left\|e_i\right\|_2 + \eta_{t-1} L \left\|\theta_{t-1} - \bar{\theta}_{t-1}\right\|_2 + \sum_{i=1}^{t-2} \eta_i L \left\|\theta_i - \bar{\theta}_i\right\|_2 + \bar{\Delta} \\
\leq & \sum_{i=1}^{t-1} \eta_i \left\|e_i\right\|_2 + \eta_{t-1} L \sum_{i=1}^{t-2} \eta_i \left\|e_i\right\|_2 + \eta_{t-1} L^2 \sum_{i=1}^{t-2} \eta_i \left\|\theta_i - \bar{\theta}_i\right\|_2 + \sum_{i=1}^{t-2} \eta_i L \left\|\theta_i - \bar{\theta}_i\right\|_2 + \bar{\Delta} \\
= & \sum_{i=1}^{t-1} \eta_i \left\|e_i\right\|_2 + \eta_{t-1} L \sum_{i=1}^{t-2} \eta_i \left\|e_i\right\|_2 + \sum_{i=1}^{t-2} \left(\eta_i L + \eta_{t-1}\eta_i L^2\right) \left\|\theta_i - \bar{\theta}_i\right\|_2 + \bar{\Delta} \\
\leq & \sum_{i=1}^{t-1} \eta_i \left\|e_i\right\|_2 + \eta_{t-1} L \sum_{i=1}^{t-2} \eta_i \left\|e_i\right\|_2 + \left(\eta_{t-2} L + \eta_{t-1}\eta_{t-2} L^2\right) \sum_{i=1}^{t-3} \eta_i \left\|e_i\right\|_2 \\
& + \sum_{i=1}^{t-3} \left(\left(\eta_{t-2} L + \eta_{t-1}\eta_{t-2} L^2\right) \eta_i L + \left(\eta_i L + \eta_{t-1}\eta_i L^2\right)\right) \left\|\theta_i - \bar{\theta}_i\right\|_2 + \bar{\Delta} \\
\leq & \sum_{i=1}^{t-1} \eta_i \left\|e_i\right\|_2 + \eta_{t-1} L \sum_{i=1}^{t-2} \eta_i \left\|e_i\right\|_2 + \left(\eta_{t-2} L + \eta_{t-1}\eta_{t-2} L^2\right) \sum_{i=1}^{t-3} \eta_i \left\|e_i\right\|_2 \\
& + \sum_{i=1}^{t-4} \left(\left(\eta_{t-2} L + \eta_{t-1}\eta_{t-2} L^2\right) \eta_{t-3} L + \left(\eta_{t-3} L + \eta_{t-1}\eta_{t-3} L^2\right)\right) \left\|e_i\right\|_2 \\
& + \sum_{i=1}^{t-4} \left(\left(\eta_{t-2}\eta_{t-3} L^2 + \eta_{t-1}\eta_{t-2}\eta_{t-3} L^3\right) \eta_i L + \left(\eta_{t-3} L + \eta_{t-1}\eta_{t-3} L^2\right) \eta_i\right) \left\|\theta_i - \bar{\theta}_i\right\|_2 + \bar{\Delta} \\
= & \sum_{i=1}^{t-1} \eta_i \prod_{j=i+1}^{t-1} (1 + \eta_j L) \left\|e_i\right\|_2 + \bar{\Delta}.
\end{aligned}
$$

Then, by the definition of τ and Markov inequality, we obtain

$$
\begin{aligned}
\mathbb{P}(\tau \leq T) = & \mathbb{P}\Big(\max_{t \in \{1,\ldots,T\}} D(\theta_t) \geq (1 + \frac{1}{\delta})\bar{\Delta}\Big) \\
\leq & \mathbb{P}\Big(\sum_{i=1}^{T-1} \eta_i \prod_{j=i+1}^{T-1} (1 + \eta_j L) \left\|e_i\right\|_2 + \bar{\Delta} \geq (1 + \frac{1}{\delta})\bar{\Delta}\Big) \\
\leq & \frac{\sum_{i=1}^{T-1} \eta_i \prod_{j=i+1}^{T-1} (1 + \eta_j L) \mathbb{E}[\left\|e_i\right\|_2]}{\frac{1}{\delta}\bar{\Delta}} \\
\leq & \frac{\delta \sum_{i=1}^{T-1} \eta_i \prod_{j=1}^{T-1} (1 + \eta_j L) \mathbb{E}[\left\|e_i\right\|_2]}{\bar{\Delta}} \\
\leq & \frac{\delta\eta(1 + \eta L)^{T-1} \sum_{i=1}^{T-1} \mathbb{E}[\left\|e_i\right\|_2]}{\bar{\Delta}},
\end{aligned}
$$

where we use the fact that $\eta_t = \eta$ for all $t \in \{1, 2, \ldots\}$. Furthermore, since $\mathbb{E}[\left\|e_i\right\|_2] \leq \sqrt{\mathbb{E}[\left\|e_i\right\|_2^2]} \leq \frac{\sigma}{B}$, we have

$$
\mathbb{P}(\tau \leq T) \leq \frac{\delta \cdot \eta \cdot T \cdot (1 + \eta L)^{T-1} \cdot \sigma}{\bar{\Delta} B}.
$$

This completes the proof. □

10.1 Proof of Theorem 4.5

Proof. By combining Lemmas 10.2 and 10.3, we obtain that

$$\begin{aligned}
\mathbb{P}(D(\theta_t)\geq\epsilon)\leq&\mathbb{P}(\tau>T, D(\theta_t)\geq\epsilon)+\mathbb{P}(\tau\leq T, D(\theta_t)\geq\epsilon)\\
\leq&\frac{\mathbb{E}[\mathbf{1}_{\tau>T}D(\theta_t)]}{\epsilon}+\mathbb{P}(\tau\leq T)\\
\leq&\left(1-\frac{\eta C_\delta^0}{8}\right)^{T-1}\frac{D(\theta_1)}{\epsilon}+\frac{5\sigma^2}{C_\delta^0 B\epsilon}+\frac{\delta\cdot\eta\cdot T\cdot(1+\eta L)^{T-1}\cdot\sigma}{\bar{\Delta}B}\\
\leq&\left(1-\frac{\eta C_\delta^0}{8}\right)^{\frac{8}{\eta C_\delta^0}\frac{\eta C_\delta^0 T}{8}}\frac{D(\theta_1)}{\epsilon}+\frac{5\sigma^2}{C_\delta^0 B\epsilon}+\frac{\delta\cdot\eta\cdot T\cdot(1+\eta L)^{T-1}\cdot\sigma}{\bar{\Delta}B}\\
\leq&\frac{1}{2}^{\frac{\eta C_\delta^0 T}{8}}\frac{D(\theta_1)}{\epsilon}+\frac{5\sigma^2}{C_\delta^0 B\epsilon}+\frac{\delta\cdot\eta\cdot T\cdot(1+\eta L)^{T-1}\cdot\sigma}{\bar{\Delta}B},
\end{aligned}$$

where the second inequality holds due to the Markov inequality, and the last inequality holds because of $(1-\frac{1}{m})^m\leq\frac{1}{2}$ for all $m\geq 1$ and $\frac{8}{\eta C_\delta^0}\geq 1$. By taking $\eta\leq\min\left\{\frac{\log T}{TL},\frac{8}{C_\delta^0},\frac{1}{2L}\right\}$, we obtain

$$\begin{aligned}
\mathbb{P}(D(\theta_t)\geq\epsilon)\leq&\frac{1}{2}^{\frac{C_\delta^0\log T}{8L}}\frac{D(\theta_1)}{\epsilon}+\frac{5\sigma^2}{C_\delta^0 B\epsilon}+\frac{\delta\cdot\log T\cdot(1+\frac{\log T}{T})^{T-1}\cdot\sigma}{\bar{\Delta}BL}\\
\leq&\frac{1}{2}^{\frac{C_\delta^0\log T}{8L}}\frac{D(\theta_1)}{\epsilon}+\frac{5\sigma^2}{C_\delta^0 B\epsilon}+\frac{\delta\cdot\log T\cdot(1+\frac{\log T}{T})^{\frac{T}{\log T}\cdot\log T}\cdot\sigma}{\bar{\Delta}BL}\\
\leq&\frac{1}{2}^{\frac{C_\delta^0\log T}{8L}}\frac{D(\theta_1)}{\epsilon}+\frac{5\sigma^2}{C_\delta^0 B\epsilon}+\frac{\delta\cdot\log T\cdot T\cdot\sigma}{\bar{\Delta}BL}\\
\leq&\frac{1}{T^{\frac{\ln 2 C_\delta^0}{8L}}}\frac{D(\theta_1)}{\epsilon}+\frac{5\sigma^2}{C_\delta^0 B\epsilon}+\frac{\delta\cdot\log T\cdot T\cdot\sigma}{\bar{\Delta}BL},
\end{aligned}$$

where we have used $(1+x)^{1/x}\leq e$ in the third inequality and $a^{\ln b}=b^{\ln a}$ in the last inequality. To guarantee $\mathbb{P}(D(\theta_t)\geq\epsilon)\leq\delta$, it suffices to have

$$T=\mathcal{O}\left(\left(\frac{3D(\theta_1)}{\delta\epsilon}\right)^{\frac{8L}{C_\delta^0\ln 2}}\right)\quad\text{and}\quad B=\tilde{\mathcal{O}}\left(\max\left\{\frac{15\sigma^2}{C_\delta^0\epsilon\delta},\frac{3\sigma}{\bar{\Delta}L}\cdot T\cdot\log T\right\}\right).$$

It total, it takes

$$\tilde{\mathcal{O}}(\max\left\{\epsilon^{\frac{8L}{C_\delta^0\ln 2}+1},\epsilon^{\frac{16L}{C_\delta^0\ln 2}}\right\})$$

samples to have $\mathbb{P}(D(\theta_t)\geq\epsilon)\leq\delta$. □

11 Proof of Lemma 4.6

We first introduce some useful results before proceeding with the proof.

Lemma 11.1 *The entropy-regularized value function $V_\lambda^\theta(\rho)$ is locally quadratic around the optimal policy π_{θ^*}. In particular, for every policy policy π_θ, we have*

$$D(\theta)\geq\frac{\lambda\min_s\rho(s)}{2\ln 2}\left|\pi_\theta(a\mid s)-\pi_{\theta^*}(a\mid s)\right|^2,\quad\forall s\in\mathcal{S},a\in\mathcal{A}.$$

Proof. It follows from the soft sub-optimality difference lemma (Lemma 26 in Mei et al. (2020)) that

$$\begin{aligned}
V_\lambda^{\theta^*}(\rho) - V_\lambda^{\theta}(\rho) =& \frac{1}{1-\gamma}\sum_s \left[d_\rho^{\pi_\theta}(s)\cdot\lambda\cdot D_{\mathrm{KL}}\left(\pi_\theta(\cdot\mid s)\|\pi_{\theta^*}(\cdot\mid s)\right)\right]\\
\geq& \frac{1}{1-\gamma}\sum_s \left[d_\rho^{\pi_\theta}(s)\cdot\lambda\cdot\frac{1}{2\ln 2}\|\pi_\theta(\cdot\mid s)-\pi_{\theta^*}(\cdot\mid s)\|_1^2\right]\\
\geq& \frac{\lambda}{2\ln 2}\sum_s \left[\rho(s)\cdot\|\pi_\theta(\cdot\mid s)-\pi_{\theta^*}(\cdot\mid s)\|_1^2\right]\\
\geq& \frac{\lambda}{2\ln 2}\sum_s \left[\rho(s)\cdot\|\pi_\theta(\cdot\mid s)-\pi_{\theta^*}(\cdot\mid s)\|_2^2\right]\\
\geq& \frac{\lambda}{2\ln 2}\left[\rho(s)\|\pi_\theta(\cdot\mid s)-\pi_{\theta^*}(\cdot\mid s)\|_2^2\right]\quad \forall s\in\mathcal{S}\\
\geq& \frac{\lambda\min_s\rho(s)}{2\ln 2}\left|\pi_\theta(a\mid s)-\pi_{\theta^*}(a\mid s)\right|^2,\quad \forall s\in\mathcal{S}, a\in\mathcal{A},
\end{aligned}$$

where the first inequality is due to Theorem 11.6 in Cover (1999) stating that

$$D_{\mathrm{KL}}\left[P(\cdot)\mid Q(\cdot)\right]\geq\frac{1}{2\ln 2}\|P(\cdot)-Q(\cdot)\|_1^2$$

for every two discrete distributions $P(\cdot)$ and $Q(\cdot)$. Moreover, the second inequality is due to $d_\rho^{\pi_\theta}(s)\geq(1-\gamma)\rho(s)$ and the third inequality is due to the equivalence between ℓ_1-norm and ℓ_2-norm. This completes the proof. □

Lemma 11.2 *Suppose that $\{\theta_t\}$ is generated by Algorithm 2 with $0<\eta_t\leq\frac{(1-\gamma)^3}{16\bar{r}+\lambda(8+16\log|\mathcal{A}|)}$ for all $t\geq 1$. We have*

$$D(\theta_{t+1})\leq\left(1-\frac{\eta_t C(\theta_t)}{4}\right)D(\theta_t)-\frac{\eta_t}{2}\xi_t+\frac{\eta_t}{4}\|e_t\|_2^2, \tag{46}$$

where $\xi_t=\langle e_t,\nabla V_\lambda^{\theta_t}(\rho)\rangle$ and $e_t=\hat{\nabla}V_\lambda^{\theta_t}(\rho)-\nabla V_\lambda^{\theta_t}(\rho)$.

Proof. Since $\nabla V_\lambda^{\theta}(\rho)$ is L-smooth in light of Lemma 3.3, it follows from Lemma 10.1 that

$$\begin{aligned}
D(\theta_{t+1})-D(\theta_t)\leq& -\frac{\eta_t}{4}\left\|\hat{\nabla}V_\lambda^{\theta}(\rho)\right\|_2^2+\frac{\eta_t}{2}\|e_t\|_2^2\\
\leq& -\frac{\eta_t}{4}\left\|\hat{\nabla}V_\lambda^{\theta_t}(\rho)-\nabla V_\lambda^{\theta_t}(\rho)+\nabla V_\lambda^{\theta_t}(\rho)\right\|_2^2+\frac{\eta_t}{2}\|e_t\|_2^2\\
=& -\frac{\eta_t}{4}\left\|\hat{\nabla}V_\lambda^{\theta_t}(\rho)-\nabla V_\lambda^{\theta_t}(\rho)\right\|_2^2-\frac{\eta_t}{4}\left\|\hat{\nabla}V_\lambda^{\theta_t}(\rho)\right\|_2^2-\frac{\eta_t}{2}\langle e_t,\nabla V_\lambda^{\theta_t}(\rho)\rangle+\frac{\eta_t}{2}\|e_t\|_2^2\\
=& -\frac{\eta_t}{4}\left\|\nabla V_\lambda^{\theta_t}(\rho)\right\|_2^2-\frac{\eta_t}{2}\langle e_t,\nabla V_\lambda^{\theta_t}(\rho)\rangle+\frac{\eta_t}{4}\|e_t\|_2^2\\
\leq& -\frac{\eta_t C(\theta_t)}{4}D(\theta_t)-\frac{\eta_t}{2}\langle e_t,\nabla V_\lambda^{\theta_t}(\rho)\rangle+\frac{\eta_t}{4}\|e_t\|_2^2,
\end{aligned}$$

for every $\eta_t\leq\frac{1}{2L}$, where the last inequality is due to Lemma 4.1. This completes the proof. □

We now encode the error terms in (46) as

$$M_n=\sum_{t=1}^n\eta_t\xi_t, \tag{47}$$

and

$$S_n=\sum_{t=1}^n\frac{\eta_t}{4}\|e_t\|_2^2. \tag{48}$$

Since $\mathbb{E}[\xi_n]=0$, we have $\mathbb{E}[M_n]=M_{n-1}$. Therefore, M_n is a zero-mean martingale; likewise, $\mathbb{E}[S_n]\geq S_{n-1}$, and therefore S_n is a submartingale. The difficulty of controlling the errors in (47) and (48) lies in the fact that the

estimation error e_n may be unbounded. Because of this, we need to take a less direct, step-by-step approach to bound the total error increments conditioned on the event that $D(\theta_n)$ remains close to $D(\theta^*)$.

We begin by introducing the "cumulative mean square error"

$$R_n = M_n^2 + S_n.$$

By construction, we have

$$\begin{aligned} R_n &= (M_{n-1} + \eta_n \xi_n)^2 + S_{n-1} + \frac{1}{4}\eta_n \|e_n\|^2 \\ &= R_{n-1} + 2M_{n-1}\eta_n \xi_n + \eta_n^2 \xi_n^2 + \frac{1}{4}\eta_n \|e_n\|^2 \end{aligned}$$

Hence,

$$\mathbb{E}[R_n] = R_{n-1} + 2M_{n-1}\eta_n \mathbb{E}[\xi_n] + \eta_n^2 \mathbb{E}\left[\xi_n^2\right] + \frac{1}{4}\eta_n \mathbb{E}\left[\|e_n\|^2\right] \geq R_{n-1},$$

i.e., R_n is a submartingale. With a fair degree of hindsight, we will choose $\varepsilon > 0$, and define $\mathcal{U}$ and $\mathcal{U}_1$ as:

$$\mathcal{U} = \left\{\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|} : D(\pi) \leq 2\epsilon + \sqrt{\varepsilon}\right\}. \tag{49}$$

We also assume that π_{θ_1} is initialized in a neighborhood $\mathcal{U}_1 \subseteq \mathcal{U}$ such that

$$\mathcal{U}_1 \subseteq \left\{\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|} : D(\pi) \leq \epsilon\right\}. \tag{50}$$

To condition it further, we also define the events

$$\Omega_n \equiv \Omega_n(\epsilon) = \{\pi_{\theta_t} \in \mathcal{U} \text{ for all } t = 1, 2, \ldots, n\}$$

and

$$E_n \equiv E_n(\varepsilon) = \{R_t \leq \varepsilon \text{ for all } t = 1, 2, \ldots, n\}$$

By definition, we also have $\Omega_0 = E_0 = \Omega$ (because the set-building index set for k is empty in this case, and every statement is true for the elements of the empty set). These events will play a crucial role in the sequel as indicators of whether π_{θ_t} has escaped the vicinity of π_{θ^*}.

Let the notation $\mathbb{1}_A$ indicate the logical indicator of an event $A \subseteq \Omega$, i.e., $\mathbb{1}_A(\omega) = 1$ if $\omega \in A$ and $\mathbb{1}_A(\omega) = 0$ otherwise. For brevity, we write $\mathcal{F}_n = \sigma(\theta_1, \ldots, \theta_n)$ for the natural filtration of θ_n. Now, we are ready to state the next lemma.

Lemma 11.3 *Let π_{θ^*} be the optimal policy. Then, for all $n \in \{1, 2, \ldots\}$, the following statements hold:*

1. *$\Omega_{n+1} \subseteq \Omega_n$ and $E_{n+1} \subseteq E_n$.*
2. *$E_{n-1} \subseteq \Omega_n$.*
3. *Consider the "large noise" event*

$$\begin{aligned} \tilde{E}_n &\equiv E_{n-1} \backslash E_n = E_{n-1} \cap \{R_n > \varepsilon\} \\ &= \{R_t \leq \varepsilon \text{ for all } t = 1, 2, \ldots, n-1 \text{ and } R_n > \varepsilon\} \end{aligned}$$

and let $\tilde{R}_n = R_n \mathbb{1}_{E_{n-1}}$ denote the cumulative error subject to the noise being "small" until time n. Then,

$$\mathbb{E}\left[\tilde{R}_n\right] \leq \mathbb{E}\left[\tilde{R}_{n-1}\right] + G^2\sigma^2\eta_n^2 + \frac{\eta_n \sigma^2}{4B} - \varepsilon \mathbb{P}\left(\tilde{E}_{n-1}\right). \tag{51}$$

By convention, we write $\tilde{E}_0 = \varnothing$ and $\tilde{R}_0 = 0$.

Proof. Statement 1 is obviously true. For Statement 2, we proceed inductively:

1. For the base case $n = 1$, we have $\Omega_1 = \{\pi_{\theta_1} \in \mathcal{U}\} \supseteq \{\pi_{\theta_1} \in \mathcal{U}_1\} = \Omega$ because π_{θ_1} is initialized in $\mathcal{U}_1 \subseteq \mathcal{U}$. Since $E_0 = \Omega$, our claim follows.

2. For the inductive step, assume that $E_{n-1} \subseteq \Omega_n$ for some $n \geq 1$. To show that $E_n \subseteq \Omega_{n+1}$, we fix a realization in E_n such that $R_t \leq \varepsilon$ for all $t = 1, 2, \ldots, n$. Since $E_n \subseteq E_{n-1}$, the inductive hypothesis posits that Ω_n also occurs, i.e., $\pi_{\theta_t} \in \mathcal{U}$ for all $t = 1, 2, \ldots, n$; hence, it suffices to show that $\pi_{\theta_{n+1}} \in \mathcal{U}$. To that end, given that $\pi_{\theta_t} \in \mathcal{U}$ for all $t = 1, 2, \ldots n$, the distance estimate (46) readily gives

$$D(\theta_{t+1}) \leq D(\theta_t) + \eta_t \xi_t + \frac{\eta_t}{4} \|e_t\|_2^2, \quad \forall t = 1, 2, \ldots, n.$$

Therefore, after telescoping, we obtain

$$D(\theta_{n+1}) \leq D(\theta_1) + M_n + S_n \leq D(\theta_1) + \sqrt{R_n} + R_n \leq \varepsilon + \sqrt{\varepsilon} + \varepsilon = 2\varepsilon + \sqrt{\varepsilon}$$

by the inductive hypothesis. This completes the induction.

For Statement 3, we decompose $\tilde{R}_n$ as

$$\begin{aligned}
\tilde{R}_n = R_n \mathbb{1}_{E_{n-1}} &= R_{n-1}\mathbb{1}_{E_{n-1}} + (R_n - R_{n-1})\,\mathbb{1}_{E_{n-1}} \\
&= R_{n-1}\mathbb{1}_{E_{n-2}} - R_{n-1}\mathbb{1}_{\tilde{E}_{n-1}} + (R_n - R_{n-1})\,\mathbb{1}_{E_{n-1}} \\
&= \tilde{R}_{n-1} + (R_n - R_{n-1})\,\mathbb{1}_{E_{n-1}} - R_{n-1}\mathbb{1}_{\tilde{E}_{n-1}}
\end{aligned}$$

where we have used the fact that $E_{n-1} = E_{n-2} \backslash \tilde{E}_{n-1}$ so $\mathbb{1}_{E_{n-1}} = \mathbb{1}_{E_{n-2}} - \mathbb{1}_{\tilde{E}_{n-1}}$ (recall that $E_{n-1} \subseteq E_{n-2}$). Then, by the definition of R_n, we have

$$R_n - R_{n-1} = 2M_{n-1}\eta_n\xi_n + \eta_n^2\xi_n^2 + \frac{1}{4}\eta_n \|e_n\|^2$$

and therefore

$$\mathbb{E}\left[(R_n - R_{n-1})\,\mathbb{1}_{E_{n-1}}\right] = 2\eta_n \mathbb{E}\left[M_{n-1}\xi_n \mathbb{1}_{E_{n-1}}\right] + \eta_n^2 \mathbb{E}\left[\xi_n^2 \mathbb{1}_{E_{n-1}}\right] + \frac{1}{4}\eta_n \mathbb{E}\left[\|e_n\|^2 \mathbb{1}_{E_{n-1}}\right]. \tag{52}$$

However, since E_{n-1} and M_{n-1} are both $\mathcal{F}_n$-measurable, we have the following estimates:

- For the term in (52), by the unbiasedness of the gradient estimator shown in Lemma 3.4, we have:

$$\mathbb{E}\left[M_{n-1}\xi_n \mathbb{1}_{E_{n-1}}\right] = \mathbb{E}\left[M_{n-1}\mathbb{1}_{E_{n-1}}\mathbb{E}\left[\xi_n \mid \mathcal{F}_n\right]\right] = 0.$$

- The second term in (52) is where the conditioning on E_{n-1} plays the most important role. It holds that:

$$\begin{aligned}
\mathbb{E}\left[\xi_n^2 \mathbb{1}_{E_{n-1}}\right] &= \mathbb{E}\left[\mathbb{1}_{E_{n-1}} \mathbb{E}\left[\left\langle e_n, \nabla V_\lambda^{\theta_n}(\rho)\right\rangle^2 \mid \mathcal{F}_n\right]\right] \\
&\leq \mathbb{E}\left[\mathbb{1}_{E_{n-1}} \left\|\nabla V_\lambda^{\theta_n}(\rho)\right\|^2 \mathbb{E}\left[\|e_n\|^2 \mid \mathcal{F}_n\right]\right] \\
&\leq \mathbb{E}\left[\mathbb{1}_{\Omega_n} \left\|\nabla V_\lambda^{\theta_n}(\rho)\right\|^2 \mathbb{E}\left[\|e_n\|^2 \mid \mathcal{F}_n\right]\right] \\
&\leq G^2\sigma^2
\end{aligned}$$

where the first inequality is due to the Cauchy-Schwarz inequality, the second inequality follows from $E_{n-1} \subseteq \Omega_n$ and the last inequality results from Lemmas 3.2 and 3.5.

- Finally, for the third term in (52), we have:

$$\frac{\eta_n}{4}\mathbb{E}\left[\|e_n\|_2^2 \mathbb{1}_{E_{n-1}}\right] \leq \frac{\eta_n \sigma^2}{4B}. \tag{53}$$

Thus, putting together all of the above, we obtain:

$$\mathbb{E}\left[(R_n - R_{n-1})\,\mathbb{1}_{E_{n-1}}\right] \leq G^2\sigma^2\eta_n^2 + \frac{\eta_n\sigma^2}{4B}.$$

Since $R_{n-1} > \varepsilon$ if $\tilde{E}_{n-1}$ occurs, we obtain

$$\mathbb{E}\left[R_{n-1}\mathbb{1}_{\tilde{E}_{n-1}}\right] \geq \varepsilon \mathbb{E}\left[\mathbb{1}_{\tilde{E}_{n-1}}\right] = \varepsilon \mathbb{P}\left(\tilde{E}_{n-1}\right).$$

This completes the proof of Statement 3. □

Lemma 11.4 *Consider an arbitrary tolerance level $\delta > 0$. If Algorithm 2 is run with a step-size schedule of the form $\eta_t = 1/(t + t_0)$ for some sufficiently large $m > 0$ and a batch size schedule $B_t \geq \frac{1}{\eta_t}$, we have*

$$\mathbb{P}(E_n) \geq 1 - \delta \quad \text{for all } n = 1, 2, \ldots$$

Proof. We begin by bounding the probability of the "large noise" event $\tilde{E}_n = E_{n-1} \backslash E_n$ as follows:

$$\begin{aligned}
\mathbb{P}\left(\tilde{E}_n\right) &= \mathbb{P}\left(E_{n-1} \backslash E_n\right) = \mathbb{P}\left(E_{n-1} \cap \{R_n > \varepsilon\}\right) \\
&= \mathbb{E}\left[\mathbb{1}_{E_{n-1}} \times \mathbb{1}_{\{R_n > \varepsilon\}}\right] \\
&\leq \mathbb{E}\left[\mathbb{1}_{E_{n-1}} \times (R_n/\varepsilon)\right] \\
&= \mathbb{E}\left[\tilde{R}_n\right]/\varepsilon
\end{aligned}$$

which is derived by using the fact that $R_n \geq 0$ (so $\mathbb{1}_{\{R_n > \varepsilon\}} \leq R_n/\varepsilon$). Now, by summing up (51), we conclude that

$$\mathbb{E}\left[\tilde{R}_n\right] \leq \mathbb{E}\left[\tilde{R}_0\right] + \frac{\sigma^2}{4B} \sum_{t=1}^{n} \eta_t - \varepsilon \sum_{t=1}^{n} \mathbb{P}\left(\tilde{E}_{t-1}\right).$$

Hence, combining the above results, we obtain the estimate

$$\sum_{t=1}^{n} \mathbb{P}\left(\tilde{E}_k\right) \leq \frac{\sigma^2}{4B\epsilon} \sum_{t=1}^{n} \eta_t \leq \frac{\sigma^2}{4\epsilon} \sum_{t=1}^{n} \eta_t^2 \leq \frac{\sigma^2 \Gamma}{4\epsilon},$$

where $\Gamma = \sum_{t=1}^{\infty} \eta_t^2 = \sum_{t=1}^{\infty} (t + t_0)^{-2}$, and we have used the relations that $\tilde{R}_0 = 0$ and $\tilde{E}_0 = \varnothing$ (by convention).

By choosing t_0 to be sufficiently large, we ensure that $\frac{\sigma^2 \Gamma}{4\epsilon} < \delta$; moreover, since the events $\tilde{E}_t$ are disjoint for all $t = 1, 2, \ldots$, we obtain

$$\mathbb{P}\left(\bigcup_{t=1}^{n} \tilde{E}_t\right) = \sum_{t=1}^{n} \mathbb{P}\left(\tilde{E}_t\right) \leq \delta.$$

Hence,

$$\mathbb{P}(E_n) = \mathbb{P}\left(\bigcap_{t=1}^{n} \tilde{E}_t^{\mathrm{c}}\right) \geq 1 - \delta$$

as claimed.

11.1 Proof of Lemma 4.6

Proof. To begin, define $\mathcal{U}$ and $\mathcal{U}_1$ as in (49) and (50). Moreover, with a fair degree of hindsight, we define $\mathcal{U}_0$ as

$$\mathcal{U}_0 = \left\{\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|} : \|\pi - \pi_{\theta^*}\|_2 \leq \frac{\epsilon}{L_\pi}\right\}$$

where $\epsilon \leq \min\left\{\left(\frac{\lambda \min_s \rho(s)}{6 \ln 2}\right)^2 \left(\alpha \min_{s,a} \pi_{\theta^*}(a|s)\right)^4, 1\right\}$ and L_π is the upper bound on the norm of the exact gradient of $V_\lambda^\pi(\rho)$ with respect to the policy π. The existence of such upper bound L_π is warranted by Section 4 in Agarwal et al. (2019) and the Lipschitz continuity of the discounted entropy $\mathbb{H}(\rho, \pi)$. Then, by construction, we have

$$\mathcal{U}_0 \subseteq \mathcal{U}_1.$$

Since the sequence Ω_n is decreasing and $\Omega_n \supseteq E_{n-1}$ (by the second part of Lemma 11.3), Lemma 11.4 yields that

$$\mathbb{P}(\Omega_T) \geq \inf_n \mathbb{P}(\Omega_n) \geq \inf_n \mathbb{P}(E_{n-1}) \geq 1 - \delta$$

provided that t_0 is chosen large enough.

Now, it remains to show that $\Omega_T \subseteq \Omega_{\alpha,T}$. We fix a realization in Ω_T such that $D(\theta_t) \leq 2\epsilon + \sqrt{\epsilon}$ for all $t = 1, 2, \ldots, T$.

By Lemma 11.1, we have

$$\begin{aligned}|\pi_{\theta_t}(a \mid s) - \pi_{\theta^*}(a \mid s)| \leq & \sqrt{\frac{2D(\theta_t)\ln 2}{\lambda \min_s \rho(s)}} \\ \leq & \sqrt{\frac{2(2\epsilon + \sqrt{\epsilon})\ln 2}{\lambda \min_s \rho(s)}} \\ \leq & \sqrt{\frac{6\sqrt{\epsilon}\ln 2}{\lambda \min_s \rho(s)}} \\ \leq & \alpha \min_{s,a} \pi_{\theta^*}(a \mid s),\end{aligned}$$

where the second inequality is due to the condition that the event Ω_T occurs, the third inequality is due to $\epsilon \leq \sqrt{\epsilon}$ when $\epsilon \leq 1$, and the last inequality is due to the definition of ϵ. Now, it can be easily verified that

$$\pi_{\theta_t}(a \mid s) \geq \pi_{\theta^*}(a \mid s) - \alpha \min_{s,a} \pi_{\theta^*}(a \mid s).$$

For every $t \in \{1, 2, \ldots, T\}$, let $\bar{s}, \bar{a} = \operatorname{argmin}_{s,a} \pi_{\theta_t}(a \mid s)$. One can write

$$\begin{aligned}\min_{s,a} \pi_{\theta_t}(a \mid s) = & \pi_{\theta_t}(\bar{a} \mid \bar{s}) \\ \geq & \pi_{\theta^*}(\bar{a} \mid \bar{s}) - \alpha \min_{s,a} \pi_{\theta^*}(a \mid s) \\ \geq & (1-\alpha) \min_{s,a} \pi_{\theta^*}(\bar{a} \mid \bar{s}),\end{aligned}$$

where the last inequality is due to $\pi(a|s) \geq \min_{s,a} \pi(a|s)$ for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Thus, we obtain

$$\mathbb{P}(\Omega_{\alpha,T}) \geq \mathbb{P}(\Omega_T) \geq 1 - \delta.$$

This completes the proof. □

12 Proof of Theorem 4.7

Proof. It follows from Lemma 11.2 that

$$D(\theta_{t+1})\mathbb{1}_{\Omega_{\alpha,t}} \leq \left(1 - \frac{\eta_t C(\theta_t)}{4}\right) D(\theta_t)\mathbb{1}_{\Omega_{\alpha,t}} - \frac{\eta_t}{2}\xi_t \mathbb{1}_{\Omega_{\alpha,t}} + \frac{\eta_t}{4}\|e_t\|_2^2 \mathbb{1}_{\Omega_{\alpha,t}}, \tag{54}$$

where $\xi_t = \langle e_t, \nabla V_\lambda^{\theta_t}(\rho)\rangle$. When the event $\Omega_{\alpha,t}$ occurs, we have $C(\theta_t) \geq C_\alpha$, where

$$C_\alpha \coloneqq \frac{2\lambda}{|\mathcal{S}|} \min_s \rho(s)(1-\alpha)^2 \min_{s,a} \pi_{\theta^*}(a|s)^2 \left\| \frac{d_\rho^{\pi_\lambda^*}}{\rho} \right\|_\infty^{-1} > 0.$$

By taking the expectation, we can obtain

$$\begin{aligned}\mathbb{E}\left[-\frac{\eta_t}{2}\xi_t \mathbb{1}_{\Omega_{\alpha,t}} + \frac{\eta_t}{4}\|e_t\|_2^2 \mathbb{1}_{\Omega_{\alpha,t}}\right] = & \mathbb{E}\left[\mathbb{1}_{\Omega_{\alpha,t}} \mathbb{E}\left[-\frac{\eta_t}{2}\xi_t + \frac{\eta_t}{4}\|e_t\|_2^2 \middle| \mathcal{F}_t\right]\right] \\ = & \mathbb{E}\left[\mathbb{1}_{\Omega_{\alpha,t}} \mathbb{E}\left[\frac{\eta_t}{4}\|e_t\|_2^2 \middle| \mathcal{F}_t\right]\right] \\ \leq & \frac{\eta_t \sigma^2}{4B},\end{aligned}$$

where the first equality is due to the fact that $\Omega_{\alpha,t}$ is deterministic conditioning on $\mathcal{F}_t$, the second equality is due to the unbiasedness of ξ_t conditioning on $\mathcal{F}_t$, and the first inequality is due to (53). Therefore,

$$\mathbb{E}[D(\theta_{t+1})\mathbb{1}_{\Omega_{\alpha,t}}] \leq \left(1 - \frac{\eta_t C_\alpha}{4}\right) \mathbb{E}\left[D(\theta_t)\mathbb{1}_{\Omega_{\alpha,t}}\right] + \frac{\eta_t \sigma^2}{4B}.$$

Arguing inductively yields that

$$\begin{aligned}\mathbb{E}[D(\theta_{t+1})\mathbb{1}_{\Omega_{\alpha,t}}] \leq& \prod_{i=1}\left(1-\frac{\eta_i C_\alpha}{4}\right)\Delta_1 + \sum_{i=1}\left(1-\frac{\eta_i C_\alpha}{4}\right)^i \frac{\eta_i\sigma^2}{4B}\\ \leq& \prod_{i=1}\left(1-\frac{\eta_i C_\alpha}{4}\right)\Delta_1 + \sum_{i=1}\frac{\eta_i\sigma^2}{4B}.\end{aligned}$$

By setting $t+1=T$ and taking $\eta_i = \frac{4}{C_\alpha(i+t_0)}$, we obtain that

$$\begin{aligned}\mathbb{E}[D(\theta_T)\mathbb{1}_{\Omega_{\alpha,T-1}}] \leq& \prod_{i=1}^{T-1}\left(1-\frac{\eta_i C_\alpha}{4}\right)\Delta_1 + \sum_{i=1}^{T-1}\frac{\eta_i\sigma^2}{4B}\\ \leq& \prod_{i=1}^{T}\left(1-\frac{\eta_i C_\alpha}{4}\right)\Delta_1 + \sum_{i=1}^{T}\frac{\eta_i\sigma^2}{4B}\\ =& \prod_{i=1}^{T}\left(\frac{i+t_0-1}{i+t_0}\right)\Delta_1 + \frac{\sigma^2}{C_\alpha B}\sum_{i=1}^{T}\frac{1}{i+t_0}\\ \leq& \frac{t_0}{T+t_0}\Delta_1 + \frac{\sigma^2\ln(T+t_0)}{BC_\alpha}.\end{aligned}$$

Since $\Omega_{\alpha,T}\subseteq\Omega_{\alpha,t}$ for all $t=1,2,\ldots,T$, it can be concluded that

$$\mathbb{E}[D(\theta_T)\mathbb{1}_{\Omega_{\alpha,T}}] \leq \frac{t_0}{T+t_0}\Delta_1 + \frac{\sigma^2\ln(T+t_0)}{BC_\alpha}.$$

Then, the claim of the theorem follows by noting that

$$\begin{aligned}\mathbb{E}[D(\theta_T)\mid\Omega_{\alpha,T}] \leq& \frac{\mathbb{E}[D(\theta_T)\mathbb{1}_{\Omega_{\alpha,T}}]}{\mathbb{P}(\Omega_{\alpha,T})}\\ \leq& \frac{t_0}{(T+t_0)(1-\delta)}\Delta_1 + \frac{\sigma^2\ln(T+t_0)}{B(1-\delta)C_\alpha}.\end{aligned}$$

By the law of total probability and the Markov inequality, we obtain that

$$\begin{aligned}\mathbb{P}(D(\theta_T)\geq\epsilon) =& \mathbb{P}(D(\theta_T)\geq\epsilon\mathbb{1}_{\Omega_{\alpha,T}}) + \mathbb{P}(D(\theta_T)\geq\epsilon\mathbb{1}^c_{\Omega_{\alpha,T}})\\ =& \mathbb{P}(D(\theta_T)\geq\epsilon\mid\Omega_{\alpha,T})\mathbb{P}(\Omega_{\alpha,T}) + \mathbb{P}(D(\theta_T)\geq\epsilon\mid\Omega^c_{\alpha,T})\mathbb{P}(\Omega^c_{\alpha,T})\\ \leq& \frac{\mathbb{E}[D(\theta_T)\mid\Omega_{\alpha,T}]}{\epsilon}\mathbb{P}(\Omega_{\alpha,T}) + \mathbb{P}(D(\theta_T)\geq\epsilon\mathbb{1}^c_{\Omega_{\alpha,T}})\mathbb{P}(\Omega^c_{\alpha,T})\\ \leq& \frac{\mathbb{E}[D(\theta_T)\mathbb{1}_{\Omega_{\alpha,T}}]}{\epsilon} + \delta\\ \leq& \frac{t_0}{(T+t_0)\epsilon}\Delta_1 + \frac{\sigma^2\ln(T+t_0)}{BC_\alpha\epsilon} + \delta.\end{aligned}$$

To guarantee that $\mathbb{P}(D(\theta_T)\geq\epsilon)\leq\mathcal{O}(\delta)$, it suffices to have

$$T \geq \frac{t_0\Delta_1}{\delta\epsilon} - t_0, \quad B \geq \frac{\sigma^2\ln(T+t_0)}{C_\alpha\delta\epsilon}.$$

Thus, the total sample complexity is $T\cdot B = \tilde{\mathcal{O}}(\frac{1}{\epsilon^2})$. This completes the proof. □