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

Yuhao Ding | Junzi Zhang | Javad Lavaei
---|---|---
University of California, Berkeley | Amazon Advertising | University of California, Berkeley
yuhao_ding@berkeley.edu | junziz@amazon.com | 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

1Work 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\footnote{Note that our main focus here is on studying global convergence, instead of achieving tight sample complexity results, which we leave as future work.} 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\[63 and \[72 in appendix.

2 Preliminaries

Markov decision processes.\footnote{\textbf{Note} that our main focus here is on studying global convergence, instead of achieving tight sample complexity results, which we leave as future work.} 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 \(r(s, a) \in [0, R]\) 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 \(\pi(\theta|s)\) 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 \(\pi(\theta)\) with the probability distribution over the trajectory as

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

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 \(\pi\), one can define the state-action value function \(Q^\pi: \mathcal{S} \times \mathcal{A} \to \mathbb{R}\) as

\[
Q^\pi(s, a) := \mathbb{E}_{s_0 \sim \rho \hookrightarrow \pi(\cdot|s_0)} \left[ \sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \right] | s_0 = s, a_0 = a.
\]

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

\[
V^\pi(s) := \mathbb{E}_{s_0 \sim \rho \hookrightarrow \pi(\cdot)} [Q^\pi(s, a)],
\]

\[
A^\pi(s, a) := Q^\pi(s, a) - V^\pi(s).
\]

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

\[
\max_{\theta \in \mathbb{R}^d} V^\pi(\rho) := \mathbb{E}_{s_0 \sim \rho} [V^\pi(s_0)].
\]

For notational convenience, we will denote \(V^\pi(\rho)\) by the shorthand notation \(V^\theta(\rho)\).

**Exploratory initial distribution.** The discounted state visitation distribution \(d^\rho_{s_0}(s)\) is defined as

\[
d^\rho_{s_0}(s) := (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}(s_h = s|s_0, \pi),
\]

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

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

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 \(\pi\) starting from the state \(s_0\). To facilitate the presentation of the main results of the paper, we assume that the state distribution \(\rho\) for the performance measure is exploratory \((\text{Mei et al. 2020}; \text{Bhandari and Russo 2019})\), i.e., \(\rho(\cdot)\) adequately covers the entire state distribution:

**Assumption 2.1** The state distribution \(\rho\) 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 \(\mu\), i.e., the gradient is taken with respect to the optimization measure \(\mu\), where \(\mu\) is usually chosen as an exploratory initial distribution that adequately covers the state distribution of some optimal policy. It is shown in \(\text{Agarwal et al. 2019}\) that the difficulty of the exploration problem faced by policy gradient algorithms can be captured through the distribution mismatch coefficient defined as \(\|d^\rho - d^\mu\|_\infty\), where \(\|d\|\) 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 \(\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}\), \(\pi(\theta|s)\) is chosen to be

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

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
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 \( \theta \in \mathbb{R}^{[S][A]} \), \( s \in S \) and \( a \in A \), we have:

\[
\| \nabla \log \pi_\theta(a|s) \|_2 \leq 2, \quad \| \nabla^2 \log \pi_\theta(a|s) \|_2 \leq 1.
\]

Reinforcement learning with entropy regularization. Entropy is a commonly used regularizer 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

\[
V_\lambda^\pi\rho(p) = V^\pi\rho(p) + \lambda H(\rho, \pi),
\]

where \( \lambda \geq 0 \) determines the strength of the penalty and \( H(\rho, \pi) \) stands for the discounted entropy defined as

\[
H(\rho, \pi) = \mathbb{E}_{s_0 \sim \rho, a_0 \sim \pi(s_0)} \left[ \sum_{t=0}^\infty -\gamma^t \log \pi(a_t|s_t) \right].
\]

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

\[
r^\lambda(s, a) := r(s, a) - \lambda \log \pi(a|s), \quad \forall (s, a) \in S \times A.
\]

We also define \( V_\lambda^\pi\rho(s) \) analogously when the initial state is fixed at a given state \( s \in S \). The regularized Q-function \( Q_\lambda^\pi\rho(s, a) \), also known as the soft Q-function, is related to \( V_\lambda^\pi\rho(s) \) as (for every \( s \in S \) and \( a \in A \))

\[
Q_\lambda^\pi\rho(s, a) = r(s, a) + \gamma \mathbb{E}_{s', \rho \sim \pi(s, a)} [V_\lambda^\pi\rho(s')] = V_\lambda^\pi\rho(s) - \lambda \log \pi(\pi(a|s) + Q_\lambda^\pi\rho(s, a)).
\]

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 \( \pi_\lambda^* \) of the entropy-regularized problem could also be nearly optimal in terms of the unregularized objective function, as long as the regularization parameter \( \lambda \) is chosen to be small. Denote by \( \pi^* \) and \( \pi_\lambda^* \) the policies that maximize the objective function and the entropy-regularized objective function with the regularization parameter \( \lambda \), respectively. Let \( V^* \) and \( V_\lambda^\pi\rho \) 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 \( \pi^* \) and \( \pi_\lambda^* \) via the following sandwich bound:

\[
V_\lambda^\pi\rho(p) \geq V^\pi(p) \geq V_\lambda^\pi\rho(p) + \frac{\lambda \log |A|}{1 - \gamma},
\]

which holds for all initial distribution \( \rho \).

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: \( \{\eta_t\}_{t=1}^t, \theta_1 \).
2: for \( t = 1, 2, \ldots, T - 1 \) do
3: \( \theta_{t+1} = \theta_{t} + \eta_t \nabla V_\lambda^\theta(p) \).
4: end for
5: Outputs: \( \theta_T \).

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** \( V_\lambda^\theta(s) \leq \frac{r + \lambda \log |A|}{1 - \gamma} \) and \( Q_\lambda^\theta(s, a) \leq \frac{r + \lambda \log |A|}{1 - \gamma} \) for all \( (s, a) \in S \times A \) and \( \theta \in \mathbb{R}^{[S][A]} \).

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

\[
\nabla V_\lambda^\theta(p) = \frac{1}{1 - \gamma} \mathbb{E}_{s, a \sim \pi_\theta(s, a)} [e_{s, a} A_\lambda^\theta(s, a)],
\]

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

\[
A_\lambda^\theta(s, a) = Q_\lambda^\theta(s, a) - \lambda \log \pi_\theta(a|s) - V_\lambda^\theta(s),
\]

\[
Q_\lambda^\theta(s, a) = r(s, a) + \gamma \mathbb{E}_{s', \rho \sim \pi_\theta(s, a)} [V_\lambda^\theta(s')].
\]

Furthermore, the entropy regularized policy gradient is bounded, i.e., \( \| \nabla V_\lambda^\theta(p) \| \leq G \) for all \( \rho \in \Delta(S) \) and \( \theta \in \mathbb{R}^{[S][A]} \), where \( G := \frac{2(r + \lambda \log |A|)}{(1 - \gamma)^2} \).

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

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

\[
\| \nabla V_\lambda^\theta(p) - \nabla V_\lambda^\theta(\theta^2) \| \leq L \| \theta^1 - \theta^2 \|,
\]

for all \( \theta^1, \theta^2 \in \mathbb{R}^d \), where the value of the Lipschitz constant \( L \) is defined as \( L := \frac{8r + 2\lambda(4 + 8 \log |A|)}{(1 - \gamma)^2} \).
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 \( r - \log \pi_\theta \) 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 (6) that in order to obtain an unbiased sample of \( \nabla V^\theta_\lambda (\rho) \), we need to first draw a state-action pair \((s, a)\) from the distribution \( \nu_\rho (\cdot, \cdot) \) and then obtain an unbiased advantage function \( A^\theta_\lambda (s, a) \)

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 \( r(s, a) = \log \pi_\theta (a | s) \) 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 \( A^\theta_\lambda (\cdot, \cdot) \) is evaluated, where the horizon \( H \sim \text{Geom}(1 - \gamma) \). It can be shown that \((s_H, a_H) \sim \nu_\rho (\cdot, \cdot) \). Moreover, given \((s_H, a_H)\), we will perform Monte-Carlo rollouts for two other trajectories with horizons \( H', H'' \sim \text{Geom}(1 - \gamma^{1/2}) \) independent of \( H \), and estimate the advantage function value \( A^\theta_\lambda (s, a) \) along the trajectories \((s_0, a_0, \ldots, s_H')\) and \((s_0', a_0', \ldots, s_{H''}')\) as follows:

\[
\hat{A}^\theta_\lambda (s, a) = \hat{Q}^\theta_\lambda (s, a) - \lambda \log \pi_\theta (a | s) - \hat{V}^\theta_\lambda (s), \tag{6}
\]

where

\[
\hat{Q}^\theta_\lambda (s, a) = r(s_0', a_0') + \sum_{t=1}^{H'} \gamma^{t/2} \cdot (r(s_t', a_t') - \lambda \log \pi_\theta (a_t' | s_t')) \mid s_0' = s, a_0' = a,
\]

\[
\hat{V}^\theta_\lambda (s) = \sum_{t=0}^{H''} \gamma^{t/2} \cdot (r(s_t'', a_t'') - \lambda \log \pi_\theta (a_t'' | s_t'')) \mid s_0'' = s.
\]

The subroutines of sampling the pair \((s, a)\) from \( \nu_\rho (\cdot, \cdot) \), estimating \( \hat{Q}^\theta_\lambda (s, a) \) and estimating \( \hat{V}^\theta_\lambda (s) \) 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:

\[
\hat{\nabla} V^\theta_\lambda (\rho) = \frac{1}{1 - \gamma} \cdot e_{s_H, a_H} \cdot \hat{A}^\theta_\lambda (s_H, a_H), \tag{7}
\]

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

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

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

**Lemma 3.5** For \( \hat{\nabla} V^\theta_\lambda (\rho) \) defined in (7), we have \( \mathbb{V} a r [\hat{\nabla} V^\theta_\lambda (\rho)] \leq \sigma^2 \), where \( \sigma^2 = \frac{8}{(1 - \gamma)^2} \left( \frac{r^2 + (\lambda \log |A|)^2}{(1 - \gamma^{1/2})^2} \right) \).

### 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 \( \nabla V^\theta_\lambda (\rho) \) can also be written as

\[
\nabla V^\theta_\lambda (\rho) = \lambda \mathbb{E} \left[ \left( \sum_{t=0}^{\infty} -\gamma^t \nabla \log \pi_\theta (a_t | s_t) \right) \right] + \frac{\lambda \mathbb{E} \left[ \left( \sum_{t=0}^{\infty} \nabla \log \pi_\theta (a_t | s_t) \right) \left( \sum_{t=0}^{\infty} -\gamma^t \log \pi_\theta (a_t | s_t) \right) \right]}{(1 - \gamma)^2} \tag{9}
\]

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

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 (1)–(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 \( \tau \) of length \( H \). For example, the commonly used truncated GPOMDP given by

\[
g_1(\tau^H|\theta, \rho) = \frac{1}{H} \sum_{h=0}^{H-1} \nabla \log \pi_\theta(a_j|s_j) \gamma^h r_h(s_h, a_h),
\]

\[
g_2(\tau^H|\theta, \rho) = \frac{1}{H} \sum_{h=0}^{H-1} \nabla \log \pi_\theta(a_j|s_j) \gamma^h \lambda \theta(s_h, a_h, s_h).
\]

can be used to estimate the terms in (1) and (9) separately. In addition, the term in (8) can be approximated by the following estimator:

\[
g_3(\tau^H|\theta, \rho) = \frac{1}{H} \sum_{h=0}^{H-1} -\gamma^h \nabla \log \pi_\theta(a_h, s_h).
\]

Then, the truncated PG estimator for \( \nabla V_\lambda^\theta \) can be written as:

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

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 \( \hat{\nabla} V_\lambda^\theta(\rho) \) defined in (14), we have

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

From Lemma 3.6, we can observe that the bias is proportional to \( \gamma^H \) 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 \( \hat{\nabla} V_\lambda^\theta, H \) has a bounded variance even if it may be unbounded when \( \pi_\theta \) approaches a deterministic policy.

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

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

### 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 (7), we can then update the policy parameter \( \theta \) by iteratively running gradient-ascent-based algorithms, i.e., \( \theta_{t+1} = \theta_t + \eta_t \sum_{i=1}^B \hat{\nabla} V_\lambda^{\theta,i}(\rho) \), 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

1: **Inputs:** \( \rho, \lambda, \theta_1, B, T, \{\eta_t\}_{t=1}^T \).
2: for \( t = 1, 2, \ldots, T \) do
3: for \( i = 1, 2, \ldots, B \) do
4:  \( s_{H,i}, a_{H,i} \leftarrow \text{Sample}(\rho, \theta_1, \gamma) \).
5:  \( \tilde{Q}_\lambda^{\theta_1,i} \leftarrow \text{Est-Max}(s_{H,i}, a_{H,i}, \theta_1, \gamma) \).
6:  \( \hat{V}_{\theta_1,i} \leftarrow \text{Est-Max}(s_{H,i}, a_{H,i}, \theta_1, \gamma) \).
7:  \( \hat{A}_{\theta_1,i} \leftarrow \tilde{Q}_{\lambda}^{\theta_1,i} - \lambda \log \pi_\theta(s_{H,i}, a_{H,i}) - \tilde{V}_{\lambda}^{\theta_1,i} \).
8:  end for
9:  Perform policy gradient update: \( \theta_{t+1} \leftarrow \theta_t + \sum_{i=1}^B \left\{ e_{s_{H,i}, a_{H,i}} \hat{A}_{\theta_1,i}(s_{H,i}, a_{H,i}) \right\} \).
10: end for
11: **Outputs:** \( \theta_T \).

**Algorithm 3** Stochastic PG for entropy regularized RL

1: **Inputs:** \( \rho, \lambda, \theta_1, B, T, \{\eta_t\}_{t=1}^T \).
2: for \( t = 1, 2, \ldots, T - 1 \) do
3:  Sample \( B \) trajectories \( \{\tau^H_t\}_{t=1}^B \) from \( p(\theta_1, \rho) \).
4:  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).
5:  Update \( \theta_{t+1} = \theta_t + \eta_t u_t \).
6:  end for
7: **Outputs:** \( \theta_T \).

### 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 [17]. 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 $V^*_\lambda(\rho)$ in [3] satisfies a non-uniform Łojasiewicz inequality as follows:

\[
\|\nabla V^*_{\lambda}(\rho)\|^2 \geq C(\theta)(V^*_{\lambda}(\rho) - V^0_{\lambda}(\rho)),
\]

where

\[
C(\theta) = \frac{2\lambda}{|\mathcal{S}|} \min_{s} \rho(s) \min_{a,\theta} \pi_\theta(a|s)^2 \left\| \frac{\pi_\theta^*}{\pi_\theta} \right\|_\infty.
\]

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 $\eta_t \leq \frac{\lambda}{\gamma}$ for the entropy regularized objective, it holds that $\inf_{t \geq 1} \min_{s, \theta} \pi_\theta(a|s) > 0$.

With Lemmas 3.9, 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 $O(e^{-\gamma})$:

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

\[
\nabla V^*_{\lambda}(\rho) - \nabla V_{\lambda}(\rho) \leq \frac{1}{\rho} \frac{1 + \lambda \log |\mathcal{A}|}{(1 - \gamma)^2} \cdot e^{-C(t-1)}.
\]

It is shown in Mei et al. (2020) that the value of $C$ depends on $\inf_{t \geq 1} \min_{s, \theta} \pi_\theta(a|s) > 0$, where $\{\theta_t\}_{t=1}^\infty$ is generated by Algorithm 1. With a bad initialization $\theta_1$, $\min_{s, \theta} \pi_\theta(a|s)$ 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 (Benaim 1999; Borkar 2009). In particular, when using the stochastic truncated PG for the entropy regularized RL, the key result of Lemma 1.2 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 ($\theta_1, \theta_2$), the reward vector $r = \{2, 1\}$ and the regularization parameter $\lambda = 1$. Then, the entropy-regularized value function can be written as $\pi_\theta^* (r - \log \pi_\theta)$.

As shown in Figure 1, the entropy-regularized value function is not coercive. When $\theta_1$ goes to positive (negative) infinity and $\theta_2$ goes to negative (positive) infinity, the landscape will become highly flat. It can also be seen that there is a line space for $(\theta_1, \theta_2)$ 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 \( \eta_t \leq \frac{\delta}{2} \) for all \( t = 1, 2, \ldots \). It holds that \( \lim_{t\to\infty} \|\nabla V_{\theta_t}(\rho)\|_2 = 0 \) with probability 1.

This result follows from classic results for the Robbins-Monro algorithm (Bertsekas and Tsitsiklis, 2000; Benaim 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 \( V_{\theta_t}(\rho) \) is not coercive in \( \theta \) and it may be the case that the gradient \( \nabla V_{\theta_t}(\rho) \) diminishing to 0 corresponds to \( \theta_t \) going to infinity instead of converging to a stationary point. In addition, the existing results (Benaim, 1999; Benaim 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.,

\[
\sup_{t \geq 0} |\theta_t| < \infty, \text{ almost surely.}
\]

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 \( D(\theta_t) = V_{\theta_t}(\rho) - V_{\theta_t}(\rho) \) as the sub-optimality gap between \( V_{\theta_t}(\rho) \) and \( V_{\theta_t}(\rho) \).

**Theorem 4.5** Consider arbitrary tolerance levels \( \delta > 0 \) and \( \epsilon > 0 \). For every initial point \( \theta_1 \), there exists a constant \( C_3^\delta > 0 \) such that if \( \theta_T \) is generated by Algorithm 2 with \( \eta_t = \eta \leq \min \left\{ \log T, \frac{8L}{C_3^\delta}, \frac{1}{2T} \right\} \) and

\[
T = O((\delta e^{-\delta})^q), \quad B = \tilde{O} \left( \max \{ (\delta e^{-1})^{-q}, (\delta e)^{-2q} \} \right)
\]

where \( q = \frac{4L}{C_3^\delta \gamma^2} \), then we have \( P \left( D(\theta_T) \leq \epsilon \right) \geq 1 - \delta \). In total, it requires \( \tilde{O} \left( \max \{ (\delta e^{-1})^{-q}, (\delta e)^{-2q} \} \right) \) samples to obtain an \( \epsilon \)-optimal policy with high probability.

The value of \( C_3^\delta \) depends on the problem parameters, namely, \( |A|, |S|, \gamma, \lambda, \rho \), the initial point \( \theta_1 \) and the constant \( \delta \). 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 \( \nabla V_{\theta_t}(\rho) \) defined in Lemma 3.5 and the boundedness region of the iterations with the exact policy gradient. The definition of \( C_3^\delta \), 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 \( \pi_{\theta_0} \) be the optimal policy of \( V_{\theta_0}(\rho) \). Assume further that Algorithm 3 is run for \( T \) iterates with a step-size sequence of the form \( \eta_t = 1/(t + t_0) \) and a batch-size sequence \( B \geq \frac{1}{\lambda} \) 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...
Theorem 4.7 Given some tolerance levels $\delta > 0$ and $\epsilon > 0$, let $\pi_\theta^*$ be the optimal policy of $V^\theta_\lambda(\rho)$. Assume that Algorithm 2 is run for $T$ iterations with a step-size sequence of the form $\eta_t = \frac{1}{(t + t_0)}$ 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 $\pi_\theta^*$ such that, if $\pi_{\theta_t} \in \mathcal{U}_0$, it holds that


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. 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 \(|X|\) 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 \( \xi \) and over all of the randomness. The notion \( \text{Var}[\cdot] \) refers to the variance. For vectors \( x, y \in \mathbb{R}^d \), let \(|x|_1\), \(|x|_2\) and \(|x|_\infty\) denote the \( \ell_1 \)-norm, \( \ell_2 \)-norm and \( \ell_\infty \)-norm. We use \( \langle x, y \rangle \) denote the inner product. For a matrix \( A \), \( A \succeq 0 \) means that \( A \) is positive semi-definite. Given a variable \( x \), the notation \( a = O(b(x)) \) means that \( a \leq C \cdot b(x) \) for some constant \( C > 0 \) that is independent of \( x \). Similarly, \( a = \tilde{O}(b(x)) \) indicates that the previous inequality may also depend on the function \( \log(x) \), where \( C > 0 \) is again independent of \( x \). We use Geom(\( x \)) to denote a geometric distribution with the parameter \( x \).

---

3Note 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:

\[ V_\lambda^\theta(s) = \mathbb{E}_{s_0=s, a_0 \sim \pi_\theta(\cdot | s), s_t \sim \mathbb{P} (\cdot | s_t, a_t)} \left[ \sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t) - \lambda \log \pi_\theta (a_t | s_t)) \right] \]

\[ = \frac{1}{1 - \gamma} \sum_s d^\pi_\theta(s) \cdot \left[ \sum_a \pi_\theta(a | s) \cdot (r(s, a) - \lambda \log \pi_\theta(a | s)) \right] \]

\[ \leq \frac{1}{1 - \gamma} \sum_s d^\pi_\theta(s) \cdot (\bar{r} + \lambda \log A) \]

\[ \leq \frac{\bar{r} + \lambda \log |A|}{1 - \gamma}, \]

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

\[ Q_\lambda^\theta(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot | s, a)} [V_\lambda^\theta(s')] \]

\[ \leq \frac{\bar{r}}{1 - \gamma} + \frac{\gamma (\lambda \log |A|)}{1 - \gamma} \]

\[ \leq \frac{\bar{r} + \lambda \log |A|}{1 - \gamma}. \]

This completes the proof. \(\square\)

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 \(V_\lambda^\theta(\rho) = \mathbb{E}_{s \sim \rho} \sum_a \pi_\theta(a | s) \cdot [Q_\lambda^\theta(s, a) - \lambda \log \pi_\theta(a | s)]\), taking derivative with respect to \(\theta\) rise to,

\[
\frac{\partial V_\lambda^\theta(\rho)}{\partial \theta} = \mathbb{E}_{s \sim \rho} \sum_a \frac{\partial \pi_\theta(a | s)}{\partial \theta} \cdot \left[ Q_\lambda^\theta(s, a) - \lambda \log \pi_\theta(a | s) \right]
+ \mathbb{E}_{s \sim \rho} \sum_a \pi_\theta(a | s) \cdot \left[ \frac{\partial Q_\lambda^\theta(s, a)}{\partial \theta} - \lambda \cdot \frac{1}{\pi_\theta(a | s)} \cdot \frac{\partial \pi_\theta(a | s)}{\partial \theta} \right] \quad (16)
\]

\[
= \mathbb{E}_{s \sim \rho} \sum_a \frac{\partial \pi_\theta(a | s)}{\partial \theta} \cdot \left[ Q_\lambda^\theta(s, a) - \lambda \log \pi_\theta(a | s) \right]
+ \mathbb{E}_{s \sim \rho} \sum_a \pi_\theta(a | s) \cdot \left[ \frac{\partial Q_\lambda^\theta(s, a)}{\partial \theta} - \lambda \cdot \frac{1}{\pi_\theta(a | s)} \right]
\]

\[
= \mathbb{E}_{s \sim \rho} \sum_a \frac{\partial \pi_\theta(a | s)}{\partial \theta} \cdot \left[ Q_\lambda^\theta(s, a) - \lambda \log \pi_\theta(a | s) \right]
+ \gamma \mathbb{E}_{s \sim \rho} \sum_a \pi_\theta(a | s) \sum_{s'} \mathbb{P}(s' | s, a) \cdot \frac{\partial V_\lambda^\theta(s')}{\partial \theta}
\]

\[
= \frac{1}{1 - \gamma} \sum_s d^\pi_\theta(s) \sum_a \frac{\partial \pi_\theta(a | s)}{\partial \theta} \cdot \left[ Q_\lambda^\theta(s, a) - \lambda \log \pi_\theta(a | s) \right]
\]

\[
= \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d^\pi_\theta, a \sim \pi_\theta(a | s)} \left[ \nabla \log \pi_\theta(a | s) \cdot \left[ Q_\lambda^\theta(s, a) - \lambda \log \pi_\theta(a | s) \right] \right]. \quad (17)
\]

where the second equation is because of

\[
\sum_a \pi_\theta(a | s) \cdot \left[ \frac{1}{\pi_\theta(a | s)} \cdot \frac{\partial \pi_\theta(a | s)}{\partial \theta} \right] = \sum_a \frac{\partial \pi_\theta(a | s)}{\partial \theta} = \frac{\partial}{\partial \theta} \sum_a \pi_\theta(a | s) = \frac{\partial 1}{\partial \theta} = 0.
\]
Due to $s' \neq s$, \( \frac{\partial \pi(a | s)}{\partial \theta(s', \gamma)} = 0 \), we obtain

\[
\frac{\partial V^\theta_\lambda(\rho)}{\partial \theta(s, a)} = \frac{1}{1 - \gamma} \cdot d_{\rho}^{\pi_\theta}(s) \cdot \left[ Q^\theta_\lambda(s, a) - \lambda \log \pi_\theta(a | s) - \sum_a \pi_\theta(a | s) \cdot [Q^\theta_\lambda(s, a) - \lambda \log \pi_\theta(a | s)] \right]
\]

Since \( \frac{\partial \pi(\cdot | s)}{\partial \theta(s, \gamma)} = \text{diag}(\pi(\cdot | s)) - \pi(\cdot | s) \pi(\cdot | s)^T \), where \( \text{diag}(x) \) denotes the diagonal matrix that has \( x \) on the diagonal, we can write

\[
\frac{\partial V^\theta_\lambda(\rho)}{\partial \theta(s, a)} = \frac{1}{1 - \gamma} \cdot d_{\rho}^{\pi_\theta}(s) \cdot \pi_\theta(a | s) \cdot \left[ Q^\theta_\lambda(s, a) - \lambda \log \pi_\theta(a | s) - \sum_a \pi_\theta(a | s) \cdot [Q^\theta_\lambda(s, a) - \lambda \log \pi_\theta(a | s)] \right]
\]

By stacking all components \( s, a \) into a vector, we obtain

\[
\frac{\partial V^\theta_\lambda(\rho)}{\partial \theta} = \frac{1}{1 - \gamma} \cdot \sum_s d_{\rho}^{\pi_\theta}(s) \sum_a \pi_\theta(a | s) \cdot e_{s,a} \cdot A^\theta_\lambda(s, a)
\]

Finally, from [17] and Jensen’s inequality, we have

\[
\left\| \frac{\partial V^\theta_\lambda(\rho)}{\partial \theta} \right\| \leq \frac{1}{1 - \gamma} \max_{a, s} \| \nabla \log \pi_\theta(a | s) \| \cdot \max_s \left\| \sum_a \pi_\theta(a | s) \cdot [Q^\theta_\lambda(s, a) - \lambda \log \pi_\theta(a | s)] \right\|
\]

\[
= \frac{1}{1 - \gamma} \max_{a, s} \| \nabla \log \pi_\theta(a | s) \| \cdot \max_s \| V^\theta_\lambda(s) \|
\]

\[
\leq \frac{2(\bar{r} + \lambda \log |A|)}{(1 - \gamma)^2}.
\]

This completes the proof.

\[\square\]

8.3 Subroutines for Algorithm 2

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

**Algorithm 4 Sam-SA:** Sample for \( s, a \sim \nu^\rho_\pi(\cdot, \cdot) \)

1: Inputs: \( \rho, \theta, \gamma \).
2: Draw \( H \sim \text{Geom}(1 - \gamma) \).
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: \hspace{1em} Simulate the next state \( s_{h+1} \sim \mathbb{P}(\cdot | s_h, a_h) \) and action \( a_{h+1} \sim \pi_\theta(\cdot | s_{h+1}) \).
6: end for
7: Outputs: \( s_H, a_H \).
Algorithm 5 Est-EntQ: Unbiasedly estimating entropy-regularized Q function

1: **Inputs**: $s, a, \gamma, \lambda$ and $\theta$.
2: Initialize $s_0 \leftarrow s, a_0 \leftarrow a, \hat{Q} \leftarrow r(s_0, a_0)$.
3: Draw $H \sim \text{Geom}(1 - \gamma^{1/2})$.
4: for $h = 0, 1, \dots, H - 1$ do
5: Simulate the next state $s_{h+1} \sim \mathbb{P}(\cdot | s_h, a_h)$ and action $a_{h+1} \sim \pi_\theta(\cdot | s_{h+1})$.
6: 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 $\hat{Q}$: $\hat{Q} \leftarrow \hat{Q} + \gamma^{(h+1)/2} (r(s_{h+1}, a_{h+1}) - \lambda \log \pi_\theta(a_{h+1}|s_{h+1}))$.
7: end for
8: Outputs: $\hat{Q}$.

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

1: **Inputs**: $s, \gamma, \lambda$ and $\theta$.
2: Initialize $s_0 \leftarrow s$, draw $a_0 \sim \pi_\theta(\cdot | s_0)$ and let $\hat{V} \leftarrow r(s_0, a_0) - \lambda \log \pi_\theta(a_0|s_0)$.
3: Draw $H \sim \text{Geom}(1 - \gamma^{1/2})$.
4: for $h = 0, 1, \dots, H - 1$ do
5: Simulate the next state $s_{h+1} \sim \mathbb{P}(\cdot | s_h, a_h)$ and action $a_{h+1} \sim \pi_\theta(\cdot | s_{h+1})$.
6: 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 $\hat{V}$: $\hat{V} \leftarrow \hat{V} + \gamma^{(h+1)/2} (r(s_{h+1}, a_{h+1}) - \lambda \log \pi_\theta(a_{h+1}|s_{h+1}))$.
7: end for
8: Outputs: $\hat{V}$.

8.4 Proof of Lemma 3.4

Proof. We first show the unbiasedness of the Q-estimate, i.e., $\mathbb{E}[\hat{Q}_\theta^0(s, a) | \theta, s, a] = Q_\theta^0(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $\theta \in \mathbb{R}^d$. In particular, from the definition of $Q_\theta^0(s, a)$, we have

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

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

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

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

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

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

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

$$
= r(s_0, a_0) + \sum_{\tau} \sum_{h=1}^{\infty} \mathbbm{1}_{H' \geq h \geq 0} \gamma^{h/2} \cdot (r(s_h, a_h) - \lambda \log \pi_\theta(a_h|s_h)) \cdot \mathbb{P}(\tau) | \theta, s_0 = s, a_0 = a, H'.
$$
By the definition of the probability over the sample trajectory \( P(\tau) \), for every \( h \in \{0,1,2,\ldots\} \), it holds that

\[
|E \left[ (r(s_h, a_h) - \lambda \log \pi_\theta(a_h|s_h)) \cdot P(\tau) \right] = \left( \sum_{s=0}^{N-1} P(s_h) \sum_{a} \pi_\theta(a_h|s_h) \right) \cdot \sum_{s=0}^{N-1} P(s_h | s_{h-1}) \cdot \sum_{a} \pi_\theta(a_h | s_h) \cdot \sum_{s'} \pi_\theta(a_{h'} | s_{h'}) |
\]

where the last inequality follows from \( P(s'|a, s) \leq 1 \), \( \pi_\theta(a|s) \leq 1 \) for all \( s, s' \in S \) and \( a \in A \) together with \( x \log x \leq \frac{1}{\epsilon} x^2 \) for \( x \in [0,1] \). Thus, for each trajectory \( \tau \) and \( N > 0 \), we have

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

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 \( \tau \) in (19) as follows:

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

In addition, for every \( N > 0 \), we have

\[
r(s_0, a_0) + \sum_{h=1}^{N} \mathbb{E}_r \left[ \gamma^{h/2} \cdot (r(s_h, a_h) - \lambda \log \pi_\theta(a_h|s_h)) \right] \mid \theta, s_0 = s, a_0 = a, H'
\]

where the third inequality is due to the boundedness of the entropy-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-expected value function over \( H' \) in (18) as follows:

\[
\mathbb{E} \left[ Q^\lambda(s, a) \mid \theta, s, a \right] = r(s_0, a_0) + \lim_{N \to \infty} \mathbb{E}_H [E_r \left[ \sum_{h=1}^{N} \gamma^{h/2} \cdot (r(s_h, a_h) - \lambda \log \pi_\theta(a_h|s_h)) \right] \mid \theta, s_0 = s, a_0 = a, H']
\]

where we have also used the fact that \( H' \) is drawn independently from the trajectory \( \tau \) in the first equality, the fact that \( H' \sim \text{Geom}(1 - \gamma^{1/2}) \) and thus \( \mathbb{E}_{H'} [1_{H' > h \geq 0}] = \gamma^{h/2} \) in the second equality, and the interchangeability
between the limit and the summation over the trajectory \( \tau \) in the third equality. This completes the proof of the unbiasedness of \( \hat{Q}_\lambda^\theta(s,a) \).

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

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

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 \( s' \sim \mathcal{P}(\cdot | s,a) \) and \( a' \sim \pi_\theta(\cdot | s') \), we have

\[
\mathbb{E} \left[ \hat{A}_\lambda^\theta(s,a) \right] = \mathbb{E} \left[ \hat{Q}_\lambda^\theta(s,a) - \lambda \log \pi_\theta(s | a) - \hat{V}_\lambda^\theta(s) \right] = A_\lambda^\theta(s,a)
\]

(28)

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

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

\[
\mathbb{E} \left[ \hat{V}_\lambda^\theta(\rho) \mid \theta \right] = \mathbb{E}_{H,(s_H,a_H)} \left\{ \mathbb{E}_{H',(s'_{i_H},a'_{i_H})} \left[ \hat{V}_\lambda^\theta(\rho) \mid \theta, s'_{0} = s_H, a'_{0} = a_H \right] \mid \theta \right\}
\]

\[
= \mathbb{E}_{H,(s_H,a_H)} \left\{ \frac{1}{1 - \gamma} \mathbb{E}_{s'_{i_H},a'_{i_H}} \left[ \hat{A}_\lambda^\theta(s'_{i_H},a'_{i_H}) \mid \theta, s'_{0} = s_H, a'_{0} = a_H \right] \mid \theta \right\}
\]

\[
= \mathbb{E}_{H,(s_H,a_H)} \left\{ \frac{1}{1 - \gamma} e_{s_H,a_H} \cdot A_\lambda^\theta(s_H,a_H) \mid \theta \right\}
\]

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

\[
\mathbb{E} \left[ \hat{V}_\lambda^\theta(\rho) \mid \theta \right] = \frac{1}{1 - \gamma} \mathbb{E}_{H,(s_H,a_H)} \left\{ \sum_{h=0}^\infty 1_{h=H} \cdot e_{s_H,a_H} \cdot A_\lambda^\theta(s_H,a_H) \mid \theta \right\}
\]

(29)

Since \( \sum_{h=0}^\infty 1_{h=H} e_{s_H,a_H} \cdot A_\lambda^\theta(s_H,a_H) \) is uniformly bounded by the boundedness of \( A_\lambda^\theta \) 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

\[
\mathbb{E} \left[ \hat{V}_\lambda^\theta(\rho) \mid \theta \right] = \frac{1}{1 - \gamma} \mathbb{E}_{H,(s_H,a_H)} \left\{ \sum_{h=0}^\infty 1_{h=H} \cdot e_{s_H,a_H} \cdot A_\lambda^\theta(s_H,a_H) \mid \theta \right\}
\]

where the second equality is due to the fact that \( H \sim \text{Geom}(1 - \gamma) \) and thus \( \mathbb{P}(h = H) = (1 - \gamma)^h \), 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 \( \hat{V}_\lambda^\theta(\rho) \).

\[\square\]

8.5 Proof of Lemma 3.5

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

\[
\hat{A}_\lambda^\theta(s_H,a_H) = \sum_{j=0}^{H'} \gamma^{j/2} \left( r(s_j',a_j') - \lambda \log \pi_\theta(a_j'|s_j') \right) - \sum_{j=0}^{H'} \gamma^{j/2} \left( r(s_j'',a_j'') - \lambda \log \pi_\theta(a_j''|s_j'') \right)
\]

(30)

\[
= \sum_{j=0}^{H'} \gamma^{j/2} r(s_j',a_j') - \sum_{j=0}^{H'} \gamma^{j/2} \lambda \log \pi_\theta(a_j'|s_j') - \sum_{j=0}^{H'} \gamma^{j/2} r(s_j'',a_j'') + \sum_{j=0}^{H'} \gamma^{j/2} \lambda \log \pi_\theta(a_j''|s_j'')
\]

(31)
Then, the policy gradient estimator \( \hat{V}^\theta(\rho) \) can be decomposed as:

\[
\hat{V}^\theta(\rho) = \frac{1}{1-\gamma} e_{s_H, a_H} \cdot (g_1(s_H, a_H) - g_2(s_H) - g_3(s_H, a_H) + g_4(s_H)).
\]  (34)

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

\[
\text{Var}(V^\theta(\rho)) = \frac{1}{(1-\gamma)^2} \text{Var}(g_1(s_H, a_H) - g_2(s_H) - g_3(s_H, a_H) + g_4(s_H)) \leq \frac{4}{(1-\gamma)^2} (\text{Var}(g_1(s_H, a_H)) + \text{Var}(g_2(s_H)) + \text{Var}(g_3(s_H, a_H)) + \text{Var}(g_4(s_H))).
\]  (35)

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

\[
\text{Var}(g_1(s_H, a_H)) \leq \frac{\bar{r}^2}{(1-\gamma)^2}, \quad \text{Var}(g_2(s_H)) \leq \frac{\bar{r}^2}{(1-\gamma)^2}.
\]

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

\[
\|g_3\|^2 \leq \lambda^2 \left( \sum_{i=0}^{H'} \gamma^{i/2} \log \pi_\theta(a'_i|s'_i) \right)^2 = \lambda^2 \left( \sum_{i=0}^{H'} \gamma^{i/2} \gamma^{i/4} \log \pi_\theta(a'_i|s'_i) \right)^2 \leq \lambda^2 \left( \sum_{i=0}^{H'} \gamma^{i/2} \left( \sum_{i=0}^{H'} \gamma^{i/2} (\log \pi_\theta(a'_i|s'_i)) \right)^2 \right) \leq \frac{\lambda^2}{1-\gamma^{1/2}} \left( \sum_{i=0}^{H'} \gamma^{i/2} (\log \pi_\theta(a'_i|s'_i)) \right)^2,
\]

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

\[
\mathbb{E}_{\tau'\sim p(\tau|\theta)} \left[ \|g_3\|^2 \right] \leq \frac{\lambda^2}{1-\gamma^{1/2}} \sum_{i=0}^{H'} \gamma^{i/2} \mathbb{E}_{\tau'\sim p(\tau|\theta)} \left[ (\log \pi_\theta(a'_i|s'_i))^2 \right].
\]  (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

\[
\mathbb{E}_{\tau'\sim p(\tau|\theta)} \left[ (\log \pi_\theta(a'_i|s'_i))^2 \right] = \mathbb{E}_{s'_l\sim p(s'_l|s'_{l-1}, a'_{l-1})} \cdots s'_{H-1}\sim p(s'_{H-1}|a'_{H-1}) \mathbb{E}_{a'_H\sim \pi_\theta(a'_H|s'_{H-1})} \left[ (\log \pi_\theta(a'_i|s'_i))^2 \right] = \mathbb{E}_{s'_l\sim p(s'_l|s'_{l-1}, a'_{l-1})} \cdots s'_{i-1}\sim p(s'_{i-1}|a'_{i-1}) \mathbb{E}_{a'_i\sim \pi_\theta(a'_i|s'_i)} \left[ \sum_{a'_i\in A} \pi_\theta(a'_i|s'_i) (\log \pi_\theta(a'_i|s'_i))^2 \right].
\]
By checking the optimality conditions for the optimization problem
\[
\max \sum_{i=1}^{n} x_i (\log x_i)^2 \quad \text{such that} \quad \sum_{i=1}^{n} x_i = 1, \tag{39}
\]
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 \(\sum_{h \in A} \pi_\theta(a_h | s_h) (\log \pi_\theta(a'_h | s'_h))^2 \leq (\log |A|)^2\) and
\[
\mathbb{E}_{\tau' \sim p(\tau | \theta)} \left[ (\log \pi_\theta(a'_h | s'_h))^2 \right] 
\leq \mathbb{E}_{s_0 \sim p(s_0), s_1 \sim p(s_1 | s_0), \ldots, s_{H-1} \sim p(s_{H-1} | s_{H-2})} \left[ (\log |A|)^2 \right] 
= (\log |A|)^2.
\]

By substituting the above inequality into \([38]\), we obtain that
\[
\mathbb{E}_{\tau' \sim p(\tau | \theta)} \left[ l_{g_3}^2 \right] \leq \frac{\lambda^2}{1 - \gamma^{1/2}} \sum_{i=0}^{H'} \gamma^{i/2} \mathbb{E}_{\tau' \sim p(\tau | \theta)} \left[ (\log \pi_\theta(a'_h | s'_h))^2 \right] 
\leq \frac{(\lambda \log |A|)^2}{1 - \gamma^{1/2}} \sum_{i=0}^{H'} \gamma^{i/2}
\leq \frac{(\lambda \log |A|)^2}{(1 - \gamma^{1/2})^2},
\]
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
\[
\mathbb{E} \left[ l_{g_3}^2 \right] \leq \frac{(\lambda \log |A|)^2}{(1 - \gamma^{1/2})^2},
\]
which further implies that \(\text{Var} \left[ l_{g_3}^2 \right] \leq \frac{(\lambda \log |A|)^2}{(1 - \gamma^{1/2})^2} \). Similarly, we can bound the variance of \(g_4\) as \(\text{Var} \left[ l_{g_4}^2 \right] \leq \frac{(\lambda \log |A|)^2}{(1 - \gamma^{1/2})^2} \). Finally, through \([35]\), we obtain
\[
\text{Var}(V_{\lambda}^\theta(\rho)) \leq \frac{8}{(1 - \gamma)^2} \left( \frac{\bar{r}^2 + (\lambda \log |A|)^2}{(1 - \gamma^{1/2})^2} \right).
\]
This completes the proof. \(\square\)

### 8.6 Proof of Lemma 8.6

**Proof.** By definition, we have
\[
\mathbb{E}[\hat{V}_{\lambda}^\theta(H)(\rho)] - \nabla V_{\lambda}^\theta(\rho) = \mathbb{E} \left[ \sum_{h=0}^{H-1} \nabla \log \pi_\theta(a_h | s_h) \left( \sum_{j=h}^{H-1} \gamma^{j} (r_j(s_j, a_j) - \lambda \log \pi_\theta(a_j | s_j)) \right) + \lambda \sum_{h=H}^{H-1} -\gamma^{j} \nabla \log \pi_\theta(a_h, s_h) \right] 
= \mathbb{E} \left[ \sum_{h=0}^{H-1} \nabla \log \pi_\theta(a_h | s_h) \left( \sum_{j=h}^{H-1} \gamma^{j} (r_j(s_j, a_j) - \lambda \log \pi_\theta(a_j | s_j)) \right) + \lambda \sum_{h=H}^{H-1} -\gamma^{j} \nabla \log \pi_\theta(a_h, s_h) \right].
\]
Then, by the Cauchy-Schwarz inequality and the triangle inequality, we obtain

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

Since \( \| \nabla \log \pi_\theta(\alpha|s) \|_2 \leq 2 \) for all \( \theta \in \mathbb{R}^{[S]\cdot[A]} \), it holds that

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

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

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

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

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

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

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

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

\[
2 \mathbb{E} \left[ \sum_{h=H}^{\infty} \left( \sum_{j=H}^{\infty} \gamma^j (r_j(s_j, a_j) - \lambda \log \pi_\theta(a_j|s_j)) \right) \right] \leq 2(\bar{r} + \lambda \log |A|) \sum_{h=H}^{\infty} \sum_{j=H}^{\infty} \gamma^j \\
\leq \frac{2(\bar{r} + \lambda \log |A|) \gamma^H}{(1 - \gamma)^2}.
\]
For the term in \([42]\), it can be easily bounded as

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

This completes the proof. \(\square\)

### 8.7 Proof of Lemma 3.7

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

\[
\text{Var}(\nabla V_{\lambda, H}(\rho)) = \mathbb{E} \left[ (g_1(\tau^H|\theta, \rho) + g_2(\tau^H|\theta, \rho)) + g_3(\tau^H|\theta, \rho) \right]
\]

As shown in Lemma 4.2 of Yuan et al. (2021), the fact that

\[
\text{Var}(\nabla V_{\lambda, H}(\rho)) \leq 3 \mathbb{E} \left[ (g_1(\tau^H|\theta, \rho) + g_2(\tau^H|\theta, \rho)) \right] + 3 \mathbb{E} \left[ (g_3(\tau^H|\theta, \rho)) \right]
\]

This completes the proof.

As shown in Lemma 4.2 of [Yuan et al., 2021], the fact that \(\|\nabla \log \pi_{\theta}(a|s)\|_1 \leq 2\) for all \(\theta \in \mathbb{R}^{|S||A|}\) directly implies that \(\text{Var}(g_1(\tau^H|\theta, \rho)) \leq \frac{4\lambda}{(1-\gamma)^2}\) for all \(\theta \in \mathbb{R}^{|S||A|}\).

Similar, due to \(\|\nabla \log \pi_{\theta}(a|s)\|_1 \leq 2\) for all \(\theta \in \mathbb{R}^{|S||A|}\), it can be verified that \(\text{Var}(g_3(\tau^H|\theta, \rho)) \leq \frac{4\lambda}{(1-\gamma)^2}\) for all \(\theta \in \mathbb{R}^{|S||A|}\).

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

\[
\left\| g_2 \right\| = \lambda \left\| \sum_{h=1}^{H-1} \left( \sum_{j=0}^{h} \nabla \log \pi_{\theta}(a_j^i|s_j^i) \right) (-\gamma^j \log \pi_{\theta}(a_j^i|s_j^i)) \right\|
\]

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

\[
\|g_2\|^2 \leq 4\lambda^2 \left( \sum_{h=0}^{H-1} (h+1) \gamma^h \log \pi_{\theta}(a_h^i|s_h^i) \right)^2
\]

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 $\tau$, it holds that
\[
\mathbb{E}_{\tau \sim p(\tau|\theta)} \left[ |g_2|^2 \right] \leq 4\lambda^2 \left( \frac{\gamma + 1}{(1-\gamma)^3} \right)^{H+1} \sum_{h=0}^{H} \gamma^h \mathbb{E}_{\tau \sim p(\tau|\theta)} \left[ \left( \log \pi_{\theta}(a_h^i|s_h^i) \right)^2 \right].
\]  
(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
\[
\mathbb{E}_{\tau \sim p(\tau|\theta)} \left[ \left( \log \pi_{\theta}(a_h^i|s_h^i) \right)^2 \right] = \mathbb{E}_{x_0 \sim \rho, a_0 \sim \pi_\theta(s_0), a_1 \sim \rho(a_0, s_0), \ldots, a_{H-1} \sim \pi_\theta(s_{H-1}, s_H-1)} \left[ \left( \log \pi_{\theta}(a_h^i|s_h^i) \right)^2 \right] = \ldots = \mathbb{E}_{x_0 \sim \rho, a_0 \sim \pi_\theta(s_0), a_1 \sim \rho(a_0, s_0), \ldots, a_{H-1} \sim \pi_\theta(s_{H-1}, s_H-1)} \left[ \sum_{a_h \in A} \pi_{\theta}(a_h|s_h^i) \left( \log \pi_{\theta}(a_h^i|s_h^i) \right)^2 \right].
\]

As proved earlier in Lemma 3.5, we know 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 $\sum_{a_h \in A} \pi_{\theta}(a_h|s_h^i) \left( \log \pi_{\theta}(a_h^i|s_h^i) \right)^2 \leq (\log |A|)^2$ and
\[
\mathbb{E}_{\tau \sim p(\tau|\theta)} \left[ \left( \log \pi_{\theta}(a_h^i|s_h^i) \right)^2 \right] \leq \mathbb{E}_{x_0 \sim \rho, a_0 \sim \pi_\theta(s_0), a_1 \sim \rho(a_0, s_0), \ldots, a_{H-1} \sim \pi_\theta(s_{H-1}, s_H-1)} \left[ \left( \log |A| \right)^2 \right] = (\log |A|)^2.
\]  
(45)

By combining (44) and (45), we have
\[
\text{Var}(g_2) \leq \mathbb{E}_{\tau \sim p(\tau|\theta)} \left[ ||g_2||^2 \right] \leq 4\lambda^2 \left( \frac{\gamma + 1}{(1-\gamma)^3} \right)^{H+1} \sum_{h=0}^{H} \gamma^h \mathbb{E}_{\tau \sim p(\tau|\theta)} \left[ \left( \log \pi_{\theta}(a_h^i|s_h^i) \right)^2 \right] \leq 4\lambda^2 \left( \frac{\gamma + 1}{(1-\gamma)^3} \right)^{H+1} \sum_{h=0}^{H} \gamma^h \left( \log |A| \right)^2 \leq \frac{8\lambda^2 (\log |A|)^2}{(1-\gamma)^4}.
\]

Finally, by substituting $\text{Var}(g_1), \text{Var}(g_2)$ and $\text{Var}(g_3)$ into (43), it holds that
\[
\text{Var}(\nabla V_{\lambda, H}^\theta (\rho)) \leq \frac{12\lambda}{(1-\gamma)^2} \left( \frac{12\lambda^2}{(1-\gamma)^4} + \frac{24\lambda^2 (\log |A|)^2}{(1-\gamma)^4} \right).
\]

This completes the proof.

\section{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.

\textbf{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 $x_{t+1} = x_t + \eta_t(u_t + w_t)$, where $\eta_t$ is a deterministic positive step-size, $u_t$ is an update direction, and $w_t$ is a random noise term. Let $F_t$ be an increasing sequence of $\sigma$-fields. Assume that

1. $f$ is a continuously differentiable function and there exists a constant $L$ such that $|\nabla f(x) - \nabla f(\bar{x})| \leq L\|x - \bar{x}\|$, $\forall x, \bar{x} \in \mathbb{R}^d$. 


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

3. There exist positive scalars $c_1$ and $c_2$ such that
   
   \[ c_1 \| \nabla f(x_t) \|^2 \leq \nabla f(x_t)^T u_t, \quad \| u_t \| \leq c_2 (1 + \| \nabla f(x_t) \|), \quad \forall t \in \{1, 2, \ldots\}. \]

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 $V^\rho_\lambda(x)$ and the update rule $\theta_{t+1} = \theta_t + \eta_t (u_t + w_t)$, where $u_t = \nabla V^\rho_\lambda(\rho)$ and $w_t = \hat{V} V^\rho_\lambda(\rho) - \nabla V^\rho_\lambda(\rho)$.

1. From Lemma 3.3, we know that Condition 1 in 9.1 is satisfied with $L = \frac{8r + 4\eta (\lambda + 8 \log |A|)}{1 - \gamma^2}$.

2. Condition 1 in 9.1 is satisfied by the definition of $\theta_t$ and $\nabla V^\rho_\lambda(\rho)$.

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 $A = \frac{8}{(1 - \gamma^2)} \left( \frac{r^2 + (\lambda \log |A|)^2}{(1 - \gamma^2)^2} \right)$.

5. Condition 1 in 9.1 is satisfied by the definition of $\eta_t$.

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

\[ \square \]

### 10 Proof of Theorem 4.5

We begin by introducing some helpful definitions. Let $\{\hat{\theta}_t\}_{t=1}^T$ denote the iterates of the algorithm with the exact PG (Algorithm 1) with $\eta_t \leq \frac{1}{2T}$ starting from the initial point $\theta_1$. Let $\theta^*$, which depends on the initial point $\theta_1$, be the optimal solution that the Algorithm 1 will converge to. Then, by Lemma 4.2, there must exist a bounded constant $\Delta$ such that $\| \theta_t - \theta^* \|_2 \leq \Delta$ for all $t \in \{1, 2, \ldots\}$. 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

\[ \tau := \min \left\{ t : \| \theta_t - \theta^* \|_2 > \left( 1 + \frac{1}{\delta} \right) \Delta \right\}, \]

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

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

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 \( 0 < \eta_t \leq \frac{1}{2L} \) for all \( t \geq 1 \), let \( \{ x_t \} \) 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

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

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

\[
f(x_{t+1}) - f(x_t) = f(x_{t+1}) - (f(x_t) - \langle u_t, x_{t+1} - x_t \rangle)
\]

\[
= f(x_{t+1}) - f(x_t) - \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \langle \sqrt{\eta_t} \nabla f(x_t) - u_t, \frac{1}{\sqrt{\eta_t}} (x_{t+1} - x_t) \rangle
\]

\[
\geq -\frac{L}{2} \| x_{t+1} - x_t \|^2 - \frac{b\eta_t}{2} \| \nabla f(x_t) - u_t \|^2 - \frac{1}{2b\eta_t} \| x_{t+1} - x_t \|^2
\]

\[
= -\left( \frac{1}{2b\eta_t} - \frac{L}{2} \right) \| x_{t+1} - x_t \|^2 - \frac{b\eta_t}{2} \| e_t \|^2,
\]

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

\[
f(x_{t+1}) \geq f(x_t) + \langle u_t, x_{t+1} - x_t \rangle - \left( \frac{1}{2b\eta_t} + \frac{L}{2} \right) \| x_{t+1} - x_t \|^2 - \frac{b\eta_t}{2} \| e_t \|^2
\]

\[
= f(x_t) + \eta_t \| u_t \|^2 - \left( \frac{\eta_t}{2b} + \frac{L\eta_t^2}{2} \right) \| u_t \|^2 - \frac{b\eta_t}{2} \| e_t \|^2
\]

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

\[
f(x_{t+1}) \geq f(x_t) + \frac{\eta_t}{4} \| u_t \|^2 - \frac{\eta_t}{2} \| e_t \|^2.
\]

This completes the proof. \( \square \)

Lemma 10.2 Let \( e_t = \nabla V^{\theta_t}_\lambda(\rho) - u_t \), where \( u_t = \frac{1}{B} \sum_{i=1}^B \nabla V^{\delta_{t,i}}_\lambda(\rho) \) and \( \nabla V^{\delta_{t,i}}_\lambda(\rho) \) is an unbiased estimator of \( \nabla V^{\theta_t}_\lambda(\rho) \). If \( \eta_t = \frac{\eta}{2} \), then

\[
\mathbb{E}[D(\theta_{T+1})] \leq \left( 1 - \frac{\eta^2}{8} \right)^{T-1} D(\theta_1) + \frac{5\eta^2}{8C^2 B}.
\]

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

\[
\mathbb{E}^t[D(\theta_{t+1}) - D(\theta_t)] = \mathbb{E}^t \left[ V^{\theta_t}_\lambda(\rho) - V^{\theta_{t+1}}_\lambda(\rho) \right] 1_{\tau > t}
\]

\[
\leq \mathbb{E}^t \left[ -\frac{\eta}{8} \| u_t \|^2 + \frac{3\eta}{4} \| e_t \|^2 \right] 1_{\tau > t}
\]

\[
\leq \mathbb{E}^t \left[ -\frac{\eta}{8} \| u_t - \nabla V^{\theta_t}_\lambda(\rho) + \nabla V^{\theta_{t+1}}_\lambda(\rho) \|^2 + \frac{3\eta}{4} \| e_t \|^2 \right] 1_{\tau > t}
\]

\[
= \mathbb{E}^t \left[ -\frac{\eta}{8} \| u_t - \nabla V^{\theta_t}_\lambda(\rho) \|^2 + \frac{8}{3} \| \nabla V^{\theta_{t+1}}_\lambda(\rho) \|^2 \right] 1_{\tau > t}
\]

\[
\leq \mathbb{E}^t \left[ -\frac{\eta}{8} \| \nabla V^{\theta_{t+1}}_\lambda(\rho) \|^2 + \frac{5\eta}{8} \| e_t \|^2 \right] 1_{\tau > t}
\]

\[
\leq \mathbb{E}^t \left[ -\frac{\eta \| \nabla \|^2}{8} + \frac{5\eta}{8} \| e_t \|^2 \right] 1_{\tau > t},
\]

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

Using the update rule of the algorithm with the exact policy gradient are bounded by
\[ \nabla \]

By setting \( t + 1 = T \), we obtain that
\[ \mathbb{E}[D(\theta_{t+1})\mathcal{F}_t]\mathbf{1}_{\tau \geq t} \leq \left(1 - \frac{\eta C_0^\delta}{8}\right) D(\theta_1) + \frac{5\eta}{8} \mathbb{E}\left[\|e_i\|_2^2 |\mathcal{F}_t]\right]. \]

In addition, conditioning on \( \mathcal{F}_t \) yields that
\[ \mathbb{E}[D(\theta_{t+1})\mathbf{1}_{\tau \geq t+1} |\mathcal{F}_t] \leq \mathbb{E}[D(\theta_{t+1})\mathbf{1}_{\tau \geq t} |\mathcal{F}_t] = \mathbb{E}[D(\theta_{t+1})|\mathcal{F}_t]\mathbf{1}_{\tau > t}, \]

where the last equality uses the fact that \( \tau \) 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
\[ \mathbb{E}[D(\theta_{t+1})\mathbf{1}_{\tau \geq t}] \leq \prod_{i=0}^{t-1} \left(1 - \frac{\eta C_0^\delta}{8}\right) D(\theta_1) + \sum_{i=0}^{t} \left(1 - \frac{\eta C_0^\delta}{8}\right)^i \frac{5\eta}{8} \mathbb{E}\left[\|e_i\|_2^2\right]. \]

By setting \( t + 1 = T \), we obtain that
\[ \mathbb{E}[D(\theta_T)\mathbf{1}_{\tau \geq T}] \leq \left(1 - \frac{\eta C_0^\delta}{8}\right) D(\theta_1) + \frac{5\eta}{8} \frac{\delta^2}{C_0^\delta B}. \]

This completes the proof. \( \square \)

**Lemma 10.3** Let \( e_t = \nabla V_\lambda^\theta_t(\rho) - u_t \), where \( u_t = \frac{1}{B} \sum_{i=1}^B \hat{\nabla} V_\lambda^\theta_t(\rho) \) and \( \hat{\nabla} V_\lambda^\theta_t(\rho) \) is an unbiased estimator of \( \nabla V_\lambda^\theta_t(\rho) \). Then, it holds that
\[
\Pr(\tau \leq T) \leq \frac{\delta \cdot \eta \cdot T \cdot (1 + \eta L)^{T-1} \cdot \sigma}{\Delta B}.
\]

**Proof.** By the triangle inequality and the fact that the iterations of the algorithm with the exact policy gradient are bounded by \( \Delta \), we have
\[ D(\theta_t) \leq \|\hat{\theta}_t - \bar{\theta}_t\|_2 + \|\theta^* - \bar{\theta}_t\|_2 = \|\theta_t - \bar{\theta}_t\|_2 + \hat{\Delta}. \]

Using the update rule of the algorithm with the exact policy gradient \( \nabla V_\lambda^\theta_t(\rho) \) and the stochastic policy gradient \( u_t = \frac{1}{B} \sum_{j=1}^B \hat{\nabla} V_\lambda^\theta_{t,j}(\rho) \), one can write
\[
D(\theta_t) = \left\| \left( \theta_1 + \sum_{i=1}^{t-1} \eta_i u_i \right) - \left( \theta_1 + \sum_{i=1}^{t-1} \eta_i \nabla V_\lambda^\theta_t(\rho) \right) \right\|_2 + \hat{\Delta} \\
\leq \sum_{i=1}^{t-1} \eta_i \| u_i - \nabla V_\lambda^\theta_t(\rho) \|_2 + \hat{\Delta} \\
= \sum_{i=1}^{t-1} \eta_i \| u_i - \nabla V_\lambda^\theta_t(\rho) + \nabla V_\lambda^\theta_t(\rho) \|_2 + \hat{\Delta} + \| \nabla V_\lambda^\theta_t(\rho) - \nabla V_\lambda^\theta_t(\rho) \|_2 \leq \sum_{i=1}^{t-1} \eta_i \| e_i \|_2 + \sum_{i=1}^{t-1} \eta_i L \| \theta_t - \bar{\theta}_t \|_2 + \hat{\Delta}. \]
By expanding $\|\theta_t - \bar{\theta}_t\|_2$ recursively, it can be concluded that

$$
D(\theta_t) \leq \sum_{i=1}^{t-1} \eta_i \|e_i\|_2 + \eta_{t-1} L \|\theta_{t-1} - \bar{\theta}_{t-1}\|_2 + \sum_{i=1}^{t-2} \eta_i L \|\theta_{i} - \bar{\theta}_{i}\|_2 + \Delta
$$

$$
\leq \sum_{i=1}^{t-1} \eta_i \|e_i\|_2 + \eta_{t-1} L \sum_{i=1}^{t-2} \eta_i \|e_i\|_2 + \eta_{t-1} L^2 \sum_{i=1}^{t-2} \eta_i \|\theta_{i} - \bar{\theta}_{i}\|_2 + \sum_{i=1}^{t-2} \eta_i L \|\theta_{i} - \bar{\theta}_{i}\|_2 + \Delta
$$

$$
= \sum_{i=1}^{t-1} \eta_i \|e_i\|_2 + \eta_{t-1} L \sum_{i=1}^{t-2} \eta_i \|e_i\|_2 + \sum_{i=1}^{t-2} \left( \eta_i L + \eta_{t-1} \eta_i L^2 \right) \|\theta_{i} - \bar{\theta}_{i}\|_2 + \Delta
$$

$$
\leq \sum_{i=1}^{t-1} \eta_i \|e_i\|_2 + \eta_{t-1} L \sum_{i=1}^{t-2} \eta_i \|e_i\|_2 + \left( \eta_{t-2} L + \eta_{t-1} \eta_{t-2} L^2 \right) \sum_{i=1}^{t-3} \eta_i \|e_i\|_2
$$

$$
+ \sum_{i=1}^{t-4} \left( \left( \eta_{t-2} L + \eta_{t-1} \eta_{t-2} L^2 \right) \eta_{i-3} L + \left( \eta_{i-3} L + \eta_{t-1} \eta_{i-3} L^2 \right) \right) \|\theta_{i-1} - \bar{\theta}_{i-1}\|_2 + \Delta
$$

$$
\leq \sum_{i=1}^{t-1} \eta_i \|e_i\|_2 + \eta_{t-1} L \sum_{i=1}^{t-2} \eta_i \|e_i\|_2 + \left( \eta_{t-2} L + \eta_{t-1} \eta_{t-2} L^2 \right) \sum_{i=1}^{t-3} \eta_i \|e_i\|_2
$$

$$
+ \sum_{i=1}^{t-4} \left( \left( \eta_{t-2} \eta_{i-3} L^2 + \eta_{t-1} \eta_{i-2} \eta_{i-3} L^3 \right) \eta_{i-4} L + \left( \eta_{i-4} L + \eta_{t-1} \eta_{i-4} L^2 \right) \eta_i \right) \|\theta_{i-4} - \bar{\theta}_{i-4}\|_2 + \Delta
$$

$$
= \sum_{i=1}^{t-1} \eta_i \prod_{j=i+1}^{t-1} \left( 1 + \eta_j L \right) \|e_i\|_2 + \Delta.
$$

Then, by the definition of $\tau$ and Markov inequality, we obtain

$$
P(\tau \leq T) = \mathbb{P} \left( \max_{t \in \{1, \ldots, T\}} D(\theta_t) \geq (1 + \frac{1}{\delta}) \Delta \right)
$$

$$
\leq \mathbb{P} \left( \sum_{i=1}^{T-1} \eta_i \prod_{j=i+1}^{T-1} \left( 1 + \eta_j L \right) \|e_i\|_2 + \Delta \geq (1 + \frac{1}{\delta}) \Delta \right)
$$

$$
\leq \sum_{i=1}^{T-1} \eta_i \prod_{j=i+1}^{T-1} \mathbb{E}[\|e_i\|_2]
$$

$$
\frac{1}{\delta} \Delta
$$

$$
\leq \frac{\delta \sum_{i=1}^{T-1} \eta_i \prod_{j=i+1}^{T-1} \mathbb{E}[\|e_i\|_2]}{\Delta}
$$

$$
\leq \frac{\delta \eta \left( 1 + \eta L \right)^{T-1} \sum_{i=1}^{T-1} \mathbb{E}[\|e_i\|_2]}{\Delta},
$$

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

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

This completes the proof.
10.1 Proof of Theorem 4.5

Proof. By combining Lemmas 10.2 and 10.3 we obtain that

\[ \mathbb{P}(D(\theta_t) \geq \epsilon) \leq \mathbb{P}(\tau > T, D(\theta_t) \geq \epsilon) + \mathbb{P}(\tau \leq T, D(\theta_t) \geq \epsilon) \leq \mathbb{E}[1_{T > \tau} D(\theta_t)] + \mathbb{P}(\tau \leq T) \]

\[ \leq \left(1 - \frac{\eta \sigma^2}{8}\right) T^{-1} D(\theta_1) \leq \frac{1}{8} \frac{\eta \sigma^2}{C^0 \beta} + \frac{5\sigma^2}{C^0 B} + \frac{\delta \cdot \eta \cdot T \cdot (1 + \eta L)^{T - 1} \cdot \sigma}{\Delta B} \]

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 \(\frac{8}{\pi C^0} \geq 1\). By taking \(\eta \leq \min \left\{ \frac{\log T}{T^2}, \frac{8}{C^0 \beta} \right\}\), we obtain

\[ \mathbb{P}(D(\theta_t) \geq \epsilon) \leq \frac{1}{8} \frac{\eta \sigma^2}{C^0 \beta} + \frac{5\sigma^2}{C^0 B} + \frac{\delta \cdot \log T \cdot (1 + \log T)^{T - 1} \cdot \sigma}{\Delta B} \]

where we have used \((1 + x)^{1/2} \leq e\) 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

\[ T = \mathcal{O}\left(\frac{3D(\theta_1)}{\delta \epsilon}\right)^{\frac{8\sigma^2}{C^0 \beta}} \quad \text{and} \quad B = \mathcal{O}\left(\max\left\{ \frac{15\sigma^2}{C^0 \beta \delta}, \frac{3\sigma}{\Delta \lambda}, \log T \right\}\right). \]

It total, it takes

\[ \bar{\mathcal{O}}\left(\max\left\{ \epsilon^{\frac{8\sigma^2}{C^0 \beta}}, \epsilon^{\frac{16\sigma^2}{C^0 \beta}} \right\}\right) \]

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

\[ \square \]

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 \(V^\theta_\lambda(\rho)\) is locally quadratic around the optimal policy \(\pi_\theta^*\). In particular, for every policy \(\pi_\theta\), we have

\[ D(\theta) \geq \frac{\lambda \min_{a | s} \rho(s)}{2 \ln 2} \left| \pi_\theta(a | s) - \pi_\theta^*(a | s) \right|^2, \quad \forall s \in S, a \in A. \]
Proof. It follows from the soft sub-optimality difference lemma (Lemma 26 in [Mei et al. (2020)]) that

\[ V_\lambda^{\theta^*}(\rho) - V_\lambda^{\theta}(\rho) = \frac{1}{1 - \gamma} \sum_s \left[ d_\rho^{\pi}(s) \cdot \lambda \cdot D_{KL}(\pi_\theta(\cdot | s) \| \pi_{\theta^*}(\cdot | s)) \right] \]

\[ \geq \frac{1}{1 - \gamma} \sum_s \left[ d_\rho^{\pi}(s) \cdot \lambda \cdot \frac{1}{2 \ln 2} |\pi_\theta(\cdot | s) - \pi_{\theta^*}(\cdot | s)|_1^2 \right] \]

\[ \geq \frac{\lambda}{2 \ln 2} \sum_s \left[ \rho(s) \cdot |\pi_\theta(\cdot | s) - \pi_{\theta^*}(\cdot | s)|_1^2 \right] \]

\[ \geq \frac{\lambda}{2 \ln 2} \left[ \rho(s) |\pi_\theta(\cdot | s) - \pi_{\theta^*}(\cdot | s)|_2^2 \right] \quad \forall s \in \mathcal{S} \]

\[ \geq \frac{\lambda \min_s \rho(s)}{2 \ln 2} |\pi_\theta(a | s) - \pi_{\theta^*}(a | s)|_2^2, \quad \forall s \in \mathcal{S}, a \in \mathcal{A}, \]

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

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

for every two discrete distributions \( P(\cdot) \) and \( Q(\cdot) \). Moreover, the second inequality is due to \( d_\rho^{\pi}(s) \geq (1 - \gamma) \rho(s) \) and the third inequality is due to the equivalence between \( \ell_1 \)-norm and \( \ell_2 \)-norm. This completes the proof. \( \square \)

**Lemma 11.2** Suppose that \( \{\theta_t\} \) is generated by Algorithm 2 with \( 0 < \eta_t \leq \frac{(1 - \gamma)^3}{16 \rho + \lambda(8 + 16 \log |\mathcal{A}|)} \) 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 \( e_t = \nabla V_\lambda^{\theta_t}(\rho) - \nabla V_\lambda^{\theta}(\rho) \).

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

\[ D(\theta_{t+1}) - D(\theta_t) \leq -\frac{\eta_t}{4} \left\| \nabla V_\lambda^{\theta_t}(\rho) - \nabla V_\lambda^{\theta_t}(\rho) \right\|_2^2 + \frac{\eta_t}{2} \| e_t \|_2^2 \]

\[ \leq -\frac{\eta_t}{4} \left\| \nabla V_\lambda^{\theta_t}(\rho) - \nabla V_\lambda^{\theta_t}(\rho) \right\|_2^2 + \frac{\eta_t}{2} \| e_t \|_2^2 \]

\[ = -\frac{\eta_t}{4} \left\| \nabla V_\lambda^{\theta_t}(\rho) \right\|_2^2 + \frac{\eta_t}{4} \| e_t \|_2^2 \]

\[ \leq -\frac{\eta_t C(\theta_t)}{4} D(\theta_t) - \frac{\eta_t}{2} \left( \langle e_t, \nabla V_\lambda^{\theta_t}(\rho) \rangle \right) + \frac{\eta_t}{4} \| e_t \|_2^2, \]

for every \( \eta_t \leq \frac{1}{2\rho} \), where the last inequality is due to Lemma 4.1. This completes the proof. \( \square \)

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 \( E[\xi_n] = 0 \), we have \( E[M_n] = M_{n-1} \). Therefore, \( M_n \) is a zero-mean martingale; likewise, \( E[S_n] \geq S_{n-1} \), 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

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

Hence,

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

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:

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

We also assume that $\pi_{\theta_1}$ is initialized in a neighborhood $\mathcal{U}_1 \subseteq \mathcal{U}$ such that

$$\mathcal{U}_1 \subseteq \left\{ \pi \in \Delta(\mathcal{A})^{\mathbb{S}} : D(\pi) \leq \varepsilon \right\}. \quad (50)$$

To condition it further, we also define the events

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

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 $\pi_{\theta_t}$ has escaped the vicinity of $\pi_{\theta_1}$.

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

**Lemma 11.3** Let $\pi_{\theta_1}$ 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

   $$\hat{E}_n \equiv E_{n-1} \setminus E_n = E_{n-1} \cap \{ R_n > \varepsilon \} = \{ R_t \leq \varepsilon \text{ for all } t = 1, 2, \ldots, n-1 \text{ and } R_n > \varepsilon \}$$

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

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

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

**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 $\pi_{\theta_1}$ 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 $\Omega_n$ 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

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

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

$$\tilde{R}_n = R_n \mathbb{1}_{E_{n-1}} = R_{n-1} \mathbb{1}_{E_{n-2}} + (R_n - R_{n-1}) \mathbb{1}_{E_{n-1}} - R_{n-1} \mathbb{1}_{E_{n-1}}$$

where we have used the fact that $E_{n-1} = E_{n-2} \setminus \tilde{E}_{n-1}$ 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

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

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:

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

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

$$\mathbb{E}[\xi_n^2 \mathbb{1}_{E_{n-1}}] \leq \mathbb{E}[\mathbb{1}_{E_{n-1}} \mathbb{E}[(e_n, \nabla V^\theta_n(\rho))^2 | \mathcal{F}_n]]$$

$$\leq \mathbb{E}[\mathbb{1}_{E_{n-1}} \|\nabla V^\theta_n(\rho)\|^2 \mathbb{E}[\|e_n\|^2 | \mathcal{F}_n]]$$

$$\leq \mathbb{E}[\mathbb{1}_{\Omega_n} \|\nabla V^\theta_n(\rho)\|^2 \mathbb{E}[\|e_n\|^2 | \mathcal{F}_n]]$$

$$\leq G^2 \sigma^2$$

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:

$$\frac{\eta_n}{4} \mathbb{E}[\|e_n\|^2 \mathbb{1}_{E_{n-1}}] \leq \frac{\eta_n \sigma^2}{4B}. \quad (53)$$

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

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

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

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

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 \( B_t \geq \eta_t \), we have

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

Proof. We begin by bounding the probability of the “large noise” event \( \tilde{E}_n = E_{n-1} \setminus E_n \) as follows:

\[
P(\tilde{E}_n) = P(E_{n-1} \setminus E_n) = P(E_{n-1} \cap \{ R_n > \varepsilon \})
\]

\[
= \mathbb{E}[1_{E_{n-1}} \times 1_{(R_n > \varepsilon)}]
\]

\[
\leq \mathbb{E}[1_{E_{n-1}} \times (R_n/\varepsilon)]
\]

\[
= \mathbb{E}[R_n] / \varepsilon
\]

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

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

Hence, combining the above results, we obtain the estimate

\[
\sum_{t=1}^n P(\tilde{E}_t) \leq \frac{\sigma^2}{4B\varepsilon} \sum_{t=1}^n \eta_t \leq \frac{\sigma^2}{4\varepsilon} \sum_{t=1}^n \eta_t^2 \leq \frac{\sigma^2 T}{4\varepsilon},
\]

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 \( \tilde{E}_0 = \emptyset \) (by convention).

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

\[
P(\bigcap_{t=1}^n \tilde{E}_t) = \sum_{t=1}^n P(\tilde{E}_t) \leq \delta.
\]

Hence,

\[
P(E_n) = P\left(\bigcap_{t=1}^n \tilde{E}_t\right) \geq 1 - \delta
\]

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

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

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

\[
\mathcal{U}_0 \subseteq \mathcal{U}_1.
\]

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

\[
P(\Omega_T) \geq \inf_n P(\Omega_n) \geq \inf_n P(E_{n-1}) \geq 1 - \delta
\]

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 \( \Omega_T \) such that \( D(\theta_t) \leq 2\epsilon + \sqrt{t} \) for all \( t = 1, 2, \ldots, T \).
By Lemma 11.1, we have
\[
|\pi_{\theta_t}(a \mid s) - \pi_{\theta_t}(a \mid s)| \leq \sqrt{\frac{2D(\theta_t) \ln 2}{\lambda \min_s \rho(s)}} \\
\leq \sqrt{\frac{2(2\epsilon + \sqrt{\epsilon}) \ln 2}{\lambda \min_s \rho(s)}} \\
\leq \sqrt{\frac{6\epsilon \ln 2}{\lambda \min_s \rho(s)}} \\
\leq \alpha \min_{s, a} \pi_{\theta_t}(a \mid s),
\]
where the second inequality is due to the condition that the event $\Omega_T$ 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 $\epsilon$. Now, it can be easily verified that
\[
\pi_{\theta_t}(a \mid s) \geq \pi_{\theta_t}(a \mid s) - \alpha \min_{s, a} \pi_{\theta_t}(a \mid s).
\]
For every $t \in \{1, 2, \ldots, T\}$, let $\bar{s}, \bar{a} = \arg\min_{s, a} \pi_{\theta_t}(a \mid s)$. One can write
\[
\min_{s, a} \pi_{\theta_t}(a \mid s) = \pi_{\theta_t}(\bar{a} \mid \bar{s}) \\
\geq \pi_{\theta_t}(\bar{a} \mid \bar{s}) - \alpha \min_{s, a} \pi_{\theta_t}(a \mid s) \\
\geq (1 - \alpha) \min_{s, a} \pi_{\theta_t}(\bar{a} \mid \bar{s})
\]
where the last inequality is due to $\pi(a \mid s) \geq \min_{s, a} \pi(a \mid s)$ for every $s \in S$ and $a \in A$. Thus, we obtain
\[
P(\Omega_{\alpha, t}) \geq P(\Omega_T) \geq 1 - \delta.
\]
This completes the proof. 

12 Proof of Theorem 4.7

Proof. It follows from Lemma 11.2 that
\[
D(\theta_{t+1}) \mathbb{1}_{\Omega_{\alpha, t}} \leq \left(1 - \frac{\eta}{4} C(\theta_t)\right) D(\theta_t) \mathbb{1}_{\Omega_{\alpha, t}} - \frac{\eta}{2} \xi_t \mathbb{1}_{\Omega_{\alpha, t}} + \frac{\eta}{4} \|\epsilon_t\|_2^2 \mathbb{1}_{\Omega_{\alpha, t}},
\]
where $\xi_t = (\epsilon_t, \nabla V_\lambda^\theta(\rho))$. When the event $\Omega_{\alpha, t}$ occurs, we have $C(\theta_t) \geq C_\alpha$, where
\[
C_\alpha := \frac{2\lambda}{|S|} \min_{s} \rho(s)(1 - \alpha)^2 \min_{s, a} \pi_{\theta_t}(a \mid s)^2 \|d^\lambda_{\theta_t}\|_\infty^{-1} > 0.
\]
By taking the expectation, we can obtain
\[
\mathbb{E} \left[ -\frac{\eta}{2} \xi_t \mathbb{1}_{\Omega_{\alpha, t}} + \frac{\eta}{4} \|\epsilon_t\|_2^2 \mathbb{1}_{\Omega_{\alpha, t}} \right] = \mathbb{E} \left[ \mathbb{1}_{\Omega_{\alpha, t}} \mathbb{E} \left[ -\frac{\eta}{2} \xi_t + \frac{\eta}{4} \|\epsilon_t\|_2^2 \mid \mathcal{F}_t \right] \right] \\
= \mathbb{E} \left[ \mathbb{1}_{\Omega_{\alpha, t}} \mathbb{E} \left[ \frac{\eta}{4} \|\epsilon_t\|_2^2 \mid \mathcal{F}_t \right] \right] \\
\leq \frac{\eta \sigma^2}{4B},
\]
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 $\xi_t$ conditioning on $\mathcal{F}_t$, and the first inequality is due to \[53\]. Therefore,
\[
\mathbb{E} \left[ D(\theta_{t+1}) \mathbb{1}_{\Omega_{\alpha, t}} \right] \leq \left(1 - \frac{\eta C_\alpha}{4}\right) \mathbb{E} \left[ D(\theta_t) \mathbb{1}_{\Omega_{\alpha, t}} \right] + \frac{\eta \sigma^2}{4B}.
\]
Arguing inductively yields that

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

By setting \( t + 1 = T \) and taking \( \eta_t = \frac{4}{c_\alpha(t+1)} \), we obtain that

\[ \mathbb{E}[D(\theta_T) \mathbbm{1}_{\Omega_{\alpha,t-1}}] \leq \prod_{i=1}^{T-1} \left( 1 - \frac{\eta_i C_\alpha}{4} \right) \Delta_1 + \sum_{i=1}^{T-1} \frac{\eta_i \sigma^2}{4B} \]

\[ = \prod_{i=1}^{T} \left( 1 - \frac{\eta_i C_\alpha}{4} \right) \Delta_1 + \sum_{i=1}^{T} \frac{\eta_i \sigma^2}{4B} \]

\[ = \prod_{i=1}^{T} \left( 1 - \frac{\eta_i C_\alpha}{4} \right) \Delta_1 + \frac{\sigma^2}{C_\beta B} \sum_{i=1}^{T} \frac{1}{i + t_0} \]

\[ \leq \frac{t_0}{T + t_0} \Delta_1 + \frac{\sigma^2 \ln(T + t_0)}{BC_\alpha} . \]

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

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

Then, the claim of the theorem follows by noting that

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

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

\[ \mathbb{P}(D(\theta_T) \geq \epsilon) = \mathbb{P}(D(\theta_T) \geq \epsilon \mathbbm{1}_{\Omega_{\alpha,T}}) + \mathbb{P}(D(\theta_T) \geq \epsilon \mathbbm{1}_{\Omega_{\alpha,T}^c}) \]

\[ = \mathbb{P}(D(\theta_T) \geq \epsilon \mid \Omega_{\alpha,T}) \mathbb{P}(\Omega_{\alpha,T}) + \mathbb{P}(D(\theta_T) \geq \epsilon \mid \Omega_{\alpha,T}^c) \mathbb{P}(\Omega_{\alpha,T}^c) \]

\[ \leq \frac{\mathbb{E}[D(\theta_T) \mid \Omega_{\alpha,T}]}{\mathbb{P}(\Omega_{\alpha,T})} + \delta \]

\[ \leq \frac{t_0}{(T + t_0) \epsilon} \Delta_1 + \frac{\sigma^2 \ln(T + t_0)}{BC_\alpha \epsilon} + \delta. \]

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 \( T \cdot B = \mathcal{O}\left(\frac{1}{\epsilon^2}\right) \). This completes the proof. \( \square \)