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

Yuhao Ding, Junzi Zhang, and Javad Lavaei

***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. We then propose a two-phase stochastic policy gradient (PG) algorithm that uses a large batch size in the first phase to overcome the challenge of the stochastic approximation due to the non-coercive landscape, and uses a small batch size in the second phase by leveraging the curvature information around the optimal policy. We establish a global optimality convergence result and a sample complexity of $\widetilde{\mathcal{O}}(\frac{1}{\epsilon^2})$ for the proposed algorithm. Our result is the first global convergence and sample complexity results for the stochastic entropy-regularized vanilla PG method.**

***Index Terms*—Reinforcement learning, policy gradient, stochastic approximation**

I. Introduction

Entropy regularization is a popular technique to encourage exploration and prevent premature convergence for reinforcement learning (RL) algorithms. It was originally proposed in [1] to improve the performance of REINFORCE, a classical family of vanilla policy gradient (PG) 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 [2, 3], Q-learning [4, 5] and trust-region policy optimization methods [6]. It has been shown that the entropy regularization works satisfactorily with deep learning approximations for achieving an impressive empirical performance boost, provides a substantial improvement in exploration and robustness [3, 5, 7], and connects the policy gradient with Q-learning under a one-step entropy regularization [4] or a trajectory-level KL regularization[1] [8].

Recently, there has been considerable interest in the theoretical understanding of how the entropy regularization exploits the geometry of the optimization landscape. In particular, it has been shown in [9, 10, 11] that entropy regularization makes the regularized objective behave similar to a local quadratic function and thus accelerates the convergence of entropy-regularized PG algorithms. In the exact gradient setting, a linear convergence rate has been established for the entropy-regularized PG algorithms with the natural PG (NPG) or policy mirror descent [10, 11] or without the NPG [9]. However, these advantages have mostly been established for the true gradient setting and it is not fully understood whether any geometric property can be exploited to accelerate convergence to global optimality in inexact gradient settings. In the inexact gradient settings,it is proven in [11] that the NPG with the entropy regularization has a sample complexity of $\widetilde{\mathcal{O}}(\frac{1}{\epsilon^2})$ where the inexactness of the gradient can be reduced to the inexactness of the state-action value functions. However, the literature on the global optimality convergence and the sample complexity of the most fundamental PG, namely REINFORCE and its variants with regularizations, is still limited, despite its simplicity and popularity in practice. The work [9] has recently developed the first set of global convergence results for PG, which focuses on the soft-max policy parametrization by assuming access to exact PG evaluations. However, their result heavily relies on the access to the exact PG evaluations, and it has been shown that the geometric advantages existing in the exact gradient setting may not be preserved in the stochastic setting [12, 13]. It remains open whether a global optimality convergence result and a low sample complexity can be obtained for the PG with entropy regularization in the practical stochastic gradient setting.

In this paper, we provide an affirmative answer to the above question. In particular, we revisit the classical entropy regularized (vanilla) policy gradient method proposed in the seminal work [1] under the soft-max policy parametrization. We focus on the modern trajectory-level entropy regularization proposed in [5], which is shown to improve over the original one-step entropy regularization adopted in [1, 2] and [4]. Our contributions are summarized below:

- We begin by proposing two new entropy regularized

Y. Ding and J. Lavaei are with the Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94709 USA (e-mail: yuhao_ding@berkeley.edu; lavaei@berkeley.edu).

J. Zhang is with the Amazon Advertising (work done prior to joining or outside of Amazon), Palo Alto, CA 94301 USA (e-mail: junziz@amazon.com).

[1]Note that this is related to but different from the widely-used trajectory-level entropy regularization later introduced in [5].

stochastic PG 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 PG estimators are the first likelihood-ratio-based estimators in the literature with a trajectory-level entropy regularization. We show that although the estimators themselves are unbounded in general due to the entropy-induced logarithmic policy rewards, the variances indeed remain uniformly bounded.

- One main challenge on extending the result in [9] to the stochastic PG setting is the non-coercive landscape of the entropy-regularized RL. To overcome this challenge, we propose a two-phase stochastic PG algorithm that uses a large batch size in the first phase and uses a small batch size in the second phase. We establish a global optimality convergence result and a sample complexity of $\widetilde{\mathcal{O}}(\frac{1}{\epsilon^2})$ for the proposed algorithm under the softmax parameterization. Our result is the first to achieve the sample complexity of $\widetilde{\mathcal{O}}(\frac{1}{\epsilon^2})$ for the stochastic entropy-regularized vanilla PG method and matches the sample complexity of the natural PG [11] in terms of dependence on .

A. Related work

Stochastic policy gradient estimators with the original one-step entropy regularization has been proposed and adopted in [1, 2, 4]. For trajectory-level entropy regularization, an exact (visitation measure-based) policy gradient formula has been derived in [14] and later re-derived in the soft-max policy parametrization setting in [9], while stochastic policy gradient estimators have not been formally proposed or studied in the literature. The only exceptions are [3, 8], where [8] provides stochastic policy gradient estimators with a related but different trajectory-level KL regularization term, while [3] utilizes a reparametrization approach to reduce policy stochasticity to a fixed generating distribution. In contrast, we focus on the more widely used trajectory-level entropy regularization and classical likelihood-ratio-based estimators.

The theoretical understanding of policy-based methods has received considerable attention recently [9, 10, 11, 15, 16, 17, 18, 19, 20, 21]. Several techniques have been developed to improve standard PG and achieve a linear convergence rate, such as adding entropy regularization [9, 10, 11, 15], exploiting natural geometries based on Bregman divergences leading to NPG or policy mirror descent [10, 11, 16], and using a geometry-aware normalized PG (GNPG) approach to exploit the non-uniformity of the value function [22]. For the stochastic policy optimization, the existing results have mostly focused on policy mirror ascent methods with the goal of reducing the stochastic analysis to the estimation of the Q-value function [10, 11], as well as incorporating variance reduction techniques to improve the sample complexity of the vanilla PG [23, 24]. The prior literature still lacks globally optimal convergence results and sample complexity for stochastic (vanilla) PG with the entropy regularization.

B. Notation

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

II. Preliminaries

Markov decision processes. RL is generally modeled as a discounted Markov decision process (MDP) defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r,)$. 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; $\in (0,1)$ 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 $(a|s)$ at the state s is usually represented by a conditional probability distribution $_\theta(a|s)$ associated to the parameter $\in \mathbb{R}^d$. Let $= \{s_0, a_0, s_1, a_1, \ldots\}$ denote the data of a sampled trajectory under policy $_\theta$ with the probability distribution over the trajectory as $p(\,|\,;\,) := (s_0)\prod_{h=1}^{\infty}\mathbb{P}(s_{h+1}|s_h,a_h)\ _\theta(a_h|s_h)$; where $\in (\mathcal{S})$ is the probability distribution of the initial state s_0.

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

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

The state-value function $V^\pi : \mathcal{S}\to\mathbb{R}$ and the advantage function $A^\pi : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ can be defined as $V^\pi(s) := \mathbb{E}_{a\sim\pi(\cdot|s)}[Q^\pi(s,a)]$; $A^\pi(s,a) := Q^\pi(s,a) - V^\pi(s)$. The goal is to find an optimal policy in the underlying policy class that maximizes the expected discounted return, namely, $\max_{\theta\in\mathbb{R}^d} V^\pi(\,) := \mathbb{E}_{s_0\sim\rho}[V^\pi(s_0)]$. For the notional convenience, we will denote $V^\pi(\,)$ by the shorthand notation $V^\theta(\,)$.

Exploratory initial distribution. The discounted state visitation distribution $d^\pi_{s_0}$ is defined as $d^\pi_{s_0}(s) := (1-)\sum_{h=0}^{\infty}\ ^h\mathbb{P}(s_h = s|s_0;\,)$; where $\mathbb{P}(s_h = s|s_0;\,)$ is the state visitation probability that s_h is equal to s under the policy starting from the state s_0. The discounted state visitation distribution under the initial distribution is defined as $d^\pi_\rho(s) := \mathbb{E}_{s_0\sim\rho}[d^\pi_{s_0}(s)]$. Furthermore, the state-action visitation distribution induced by and the initial state distribution is defined as $v^\pi_\rho(s,a) := d^\pi_\rho(s)\ (a|s)$, which can also be written as $v^\pi_\rho(s,a) := (1-)\mathbb{E}_{s_0\sim\rho}\sum_{h=0}^{\infty}\ ^h\mathbb{P}(s_h = s, a_h = a|s_0;\,)$;

where $\mathsf{P}(s_h = s, a_h = a | s_0, \ )$ is the state-action visitation probability that $s_h = s$ and $a_h = a$ under $\ $ starting from the state s_0. To facilitate the presentation of the main results of the paper, we assume that the state distribution $\ $ for the performance measure is exploratory [9, 18], i.e., $\ (\cdot)$ adequately covers the entire state distribution:

Assumption 1. The state distribution $\ $ satisfies $\ (s) > 0$ for all $s \in \mathcal{S}$.

In practice, when the above assumption is not satisfied, we can optimize under another initial distribution $\ $, i.e., the gradient is taken with respect to the optimization measure $\ $, where $\ $ is usually chosen as an exploratory initial distribution that adequately covers the state distribution of some optimal policy. It is shown in [15] that the difficulty of the exploration problem faced by PG algorithms can be captured through the distribution mismatch coefficient defined as $\left\| \frac{d}{\mu} \right\|_\infty$, where $\frac{d}{\mu}$ denotes component-wise division.

Soft-max policy parameterization. In this work, we consider the soft-max parameterization – a widely adopted scheme that naturally ensures that the policy lies in the probability simplex. Specifically, for an unconstrained parameter $\ \in \mathsf{R}^{|\mathcal{S}||\mathcal{A}|}$, $\ _\theta(a|s)$ is chosen to be $\frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}$. The soft-max parameterization is generally used for MDPs with finite state and action spaces. It is complete in the sense that every stochastic policy can be represented by this class. For the soft-max parameterization, it can be shown that the gradient and Hessian of the function $\log \ _\theta(a|s)$ are bounded, i.e., for all $\ \in \mathsf{R}^{|\mathcal{S}||\mathcal{A}|}$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we have: $\|\nabla \log \ _\theta(a|s)\|_2 \le 2$; $\left\|\nabla^2 \log \ _\theta(a|s)\right\|_2 \le 1$.

RL with entropy regularization. Entropy is a commonly used regularization in RL to promote exploration and discourage premature convergence to suboptimal policies [5, 8, 25]. It is far less aggressive in penalizing small probabilities, in comparison to other common regularizations such as log barrier functions [15]. 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

$$V_\lambda^\pi(\ ) = V^\pi(\ ) + \ \mathsf{H}(\ ; \ ); \tag{1}$$

where $\ \ge 0$ determines the strength of the penalty and $\mathsf{H}(\ ; \ )$ stands for the discounted entropy defined as

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

Equivalently, $V_\lambda^\pi(\ )$ can be viewed as the weighted value function of $\ $ by adjusting the instantaneous reward to be policy-dependent regularized version as $r^\lambda(s,a) \coloneqq r(s,a) - \log \ (a|s)$, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. We also define $V_\lambda^\pi(s)$ analogously when the initial state is fixed at a given state $s \in \mathcal{S}$. The regularized Q-function Q_λ^π of a policy $\ $, also known as the soft Q-function, is related to V_λ^π as (for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$)

$$Q_\lambda^\pi(s,a) = r(s,a) + \ \mathsf{E}_{s' \sim P(\cdot|s,a)} \left[V_\lambda^\pi(s') \right];$$
$$V_\lambda^\pi(s) = \mathsf{E}_{a \sim \pi(\cdot|s)} \left[- \log \ (a|s) + Q_\lambda^\pi(s,a) \right].$$

Bias due to entropy regularization. Due to the presence of regularization, the optimal solution will be biased with the bias disappearing as $\ \to 0$. More precisely, the optimal policy $\ _\lambda^*$ of the entropy-regularized problem could also be nearly optimal in terms of the unregularized objective function, as long as the regularization parameter $\ $ is chosen to be small. Denote by $\ ^*$ and $\ _\lambda^*$ the policies that maximize the objective function and the entropy-regularized objective function with the regularization parameter $\ $, respectively. Let V^* and V_λ^* represent the resulting optimal objective value function and the optimal regularized objective value function. [11] shows a simple but crucial connection between $\ ^*$ and $\ _\lambda^*$ via the following sandwich bound:

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

which holds for all initial distribution $\ $.

III. Stochastic PG methods for entropy regularized RL

A. Review: Exact PG methods

The PG method is one of the most popular approaches for a direct policy search in RL [26]. The vanilla PG with exact gradient information and the entropy regularization is summarized in Algorithm 1.

Algorithm 1 Exact PG method

1: **Inputs**: $\{ \ _t \}_{t=1}^T$, $\ _1$.
2: **for** $t = 1, 2, \ldots, T-1$ **do**
3: $\quad \ _{t+1} = \ _t + \ _t \nabla V_\lambda^{\theta_t}(\ )$.
4: **end for**
5: **Outputs**: $\ _T$.

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 1 ([9]): $V_\lambda^\theta(s) \le \frac{\bar{r} + \lambda \log|\mathcal{A}|}{1-\gamma}$ and $Q_\lambda^\pi(s,a) \le \frac{\bar{r} + \lambda \log|\mathcal{A}|}{1-\gamma}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $\ \in \mathsf{R}^{|\mathcal{S}||\mathcal{A}|}$.

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

Lemma 2 (Proposition 2 in [27]): The entropy regularized PG with respect to $\ $ is

$$\nabla V_\lambda^\theta(\ ) = \tag{2}$$
$$\frac{1}{1 - \ } \mathsf{E}_{s,a \sim v} \left[\nabla_\theta \log \ _\theta(a|s) \left(Q_\lambda^\theta(s,a) - \ \log \ _\theta(a|s) \right) \right];$$

where

$$\frac{\partial \log \ _\theta(a|s)}{\partial \ _{s',a'}} = \begin{cases} - \ _\theta(a'|s') \ _\theta(a|s); & (s',a') \ne (s,a); \\ \ _\theta(a|s) - \ _\theta(a|s) \ _\theta(a|s); & (s',a') = (s,a). \end{cases}$$

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

In addition, it is shown that the PG $\nabla V_\lambda^\theta(\ )$ is Lipschitz continuous.

Lemma 3 (Lemmas 7 and 14 in [9]): The PG $\nabla V_\lambda^\theta(\ )$ is Lipschitz continuous with some constant $L > 0$, i.e.,

$\|\nabla V(\theta_1) - \nabla V(\theta_2)\| \le L\|\theta_1 - \theta_2\|$, for all $\theta_1, \theta_2 \in \mathbb{R}^d$, where the value of the Lipschitz constant L is defined as $L = \frac{8r}{(1-\gamma)^4} + \frac{8\log|\mathcal{A}|}{(1-\gamma)^3}$.

Challenges for designing entropy regularized PG estimators. Existing works either consider one-step entropy regularization [2, 28], KL divergence [8], or the re-parametrization technique [3, 5] (which introduces approximation errors that are difficult to quantify exactly). In general, the regularized reward $r - \tau\log\pi$ 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 A of the supplemental materials.

B. Sampling the unbiased PG

It results from (2) that in order to obtain an unbiased sample of $\nabla V(\theta)$, we need to first draw a state-action pair (s, a) from the distribution $\nu(\theta)$ and then obtain an unbiased estimate of the action-value function $Q(\theta; s, a)$. For the standard discounted infinite-horizon RL setting with bounded reward functions, [29] proposes an unbiased estimate of the PG 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 $r(s,a) - \tau\log\pi(a|s)$ could be unbounded when $\pi(a|s) \to 0$. Fortunately, we can still show that an unbiased PG estimator with the bounded variance for the entropy regularized RL can be obtained in a similar fashion as in [29]. 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 $Q(\theta; \cdot, \cdot)$ is evaluated, where the horizon $H \sim \text{Geom}(1-\gamma)$. It can be shown that $(s_H, a_H) \sim \nu(s, a)$. Moreover, given (s_H, a_H), we will perform Monte-Carlo rollouts for another trajectory with the horizon $H' \sim \text{Geom}(1-\gamma^{1/2})$ independent of H, and estimate the advantage function value $Q(s, a)$ along the trajectory $(s'_0, a'_0, \ldots, s'_{H'})$ with $s'_0 = s, a'_0 = a$ as follows:

$$\hat{Q}(s,a) = r(s'_0, a'_0) + \sum_{t=1}^{H'} \gamma^{t/2}\left(r(s'_t, a'_t) - \tau\log\pi(a'_t|s'_t)\right). \tag{3}$$

The subroutines of sampling one pair (s, a) from $\nu(\theta)$, estimating $Q(\theta; s, a)$, and estimating $V(\theta; s)$ are summarized as Sam-SA and Est-EntQ in Algorithms 2 and 3, respectively.

Algorithm 2 Sam-SA: Sample for $(s, a) \sim \nu(\theta)$

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

Algorithm 3 Est-EntQ: Unbiasedly estimating entropy-regularized Q function

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

Motivated by the form of PG in (2), we propose the following stochastic estimator:

$$\widehat{\nabla V}(\theta) = \frac{1}{1-\gamma}\nabla\log\pi(a_H|s_H)\hat{Q}(s_H, a_H) + \nabla\log\pi(a_H|s_H), \tag{4}$$

where $(s_H, a_H) \sim$ Sam-SA(θ, γ, τ) and $\hat{Q}$ is defined in (3). The following lemma shows that the stochastic PG (4) is an unbiased estimator of $\nabla V(\theta)$.

Lemma 4: For $\widehat{\nabla V}(\theta)$ defined in (4), we have $\mathbb{E}\,\widehat{\nabla V}(\theta) = \nabla V(\theta)$.

The next lemma shows that the proposed PG estimator $\widehat{\nabla V}(\theta)$ has a bounded variance even if it is unbounded when π approaches a deterministic policy.

Lemma 5: For $\widehat{\nabla V}(\theta)$ defined in (4), we have $\text{Var}\,\widehat{\nabla V}(\theta) \le W^2$, where $W^2 = \frac{8}{(1-\gamma)^2}\left(r^2 + \frac{\log^2|\mathcal{A}|}{(1-\gamma^{1/2})^2}\right)$.

C. Sampling the trajectory-based PG

Compared to the unbiased PG with a random horizon (4), a more practical PG estimator is the trajectory-based PG. To derive the trajectory-based PG for the entropy-regularized RL, we first notice that the gradient $\nabla V(\theta)$ can also be written as

$$\nabla V(\theta) = \mathbb{E}_{s_0\sim\rho}\left(\nabla\log\pi(a_0|s_0)\cdot\mathbb{E}_{a_0\sim\pi}Q(\theta; s_0, a_0) + \sum_{t=1}^{\infty}\gamma^t\,\mathbb{E}_{s_t\sim p_t}\left(\nabla\log\pi(a_t|s_t)\cdot\mathbb{E}_{a_t\sim\pi}Q(\theta; s_t, a_t)\right)\right),$$

where the expectation is taken over the trajectory distribution, i.e., $p(\tau)$.

Since the distribution $p(\tau)$ is unknown, $\nabla V(\theta)$ needs to be estimated from samples. The trajectory-based estimators include REINFORCE [28], PGT [30] and GPOMDP [31]. In

practice, the truncated versions of these trajectory-based PG estimators are used to approximate the in nite 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 of length H. Then, with the commonly used truncated GPOMDP, the truncated PG estimator for ∇V can be written as:

$$\widehat{\nabla V}^{;H}(\theta) = \sum_{h=0}^{H-1} \sum_{j=0}^{h} \nabla \log \pi_\theta(a_j|s_j) \left(\gamma^h r(s_h, a_h) - b \log \pi_\theta(a_h|s_h) \right). \quad (5)$$

Due to the horizon truncation, the PG estimator (5) may no longer be unbiased, but its bias can be very small with a large horizon H.

Lemma 6: For $\widehat{\nabla V}^{;H}(\theta)$ de ned in (5), we have

$$\left\| \mathbb{E}\widehat{\nabla V}^{;H}(\theta) - \nabla V(\theta) \right\|_2 \le B \frac{2r}{1-\gamma} \log(|A|) H \gamma^{H-1} + \frac{1}{1-\gamma}.$$

From Lemma 6, we can observe that the bias is proportional to γ^H and thus can be controlled to be arbitrarily small with a constant horizon up to some logarithmic term. We then show that the truncated PG estimator $\widehat{\nabla V}^{;H}$ has a bounded variance even if it may be unbounded when θ approaches a deterministic policy.

Lemma 7: For $\widehat{\nabla V}^{;H}(\theta)$ de ned in (5), we have

$$\mathrm{Var}\left[\widehat{\nabla V}^{;H}(\theta)\right] \le B \frac{12 r^2}{1-\gamma} + \frac{24}{(1-\gamma)^4} \log(|A|).$$

D. Batched PG algorithms

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

Algorithm 4 Ent-RPG: Random-horizon PG for Entropy-regularized RL

1: Inputs: θ, θ_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: $s^i_{H_t}, a^i_{H_t} \sim \mathrm{Sam}(\pi_\theta, \tau_t, \cdot)$.
5: $\widehat{Q}^{t,i} \leftarrow$ Est-EntQ$(s^i_{H_t}, a^i_{H_t}, \tau_t, \theta, \cdot)$.
6: end for
7: $\theta_{t+1} = \theta_t + \frac{\eta_t}{1-\gamma} \frac{1}{B} \sum_{i=1}^B \nabla \log \pi_\theta(a^i_{H_t}|s^i_{H_t}) \cdot \left(\widehat{Q}^{t,i} - \log \pi_\theta(s^i_{H_t}, a^i_{H_t}) \right)$
8: end for
9: Outputs: θ_T.

Remark 1: For the simplicity of the presentation, we focus on deriving the stochastic PG estimator for the soft-max policy parameterization. However, our results in this section (and also the stationary point convergence result in Section IV-C below) can be easily extended to the general parameterization as long as $\nabla \log \pi_\theta(s,a) \cdot Y_2$ and $\nabla^2 \log \pi_\theta(s,a) \cdot Z_2$ are bounded for all $(s,a) \in S \times A$.

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 PG estimator (4) for the rest of the paper. Similar results hold for the trajectory-based PG estimator in (5) since its bias is exponentially small with respect to the horizon (see Lemma 6). We leave the formal discussion of these results as future work.

IV. Non-coercive landscape

In this section, we first review some key results for the entropy-regularized RL with the exact PG and highlight the difficulty of generalizing these results to the stochastic PG setting, due to the non-coercive landscape.

A. Review: Linear convergence with exact PG

A key result from [9] shows that, under the soft-max parameterization, the entropy-regularized value function $V(\theta)$ in (1) satisfies a non-uniform Łojasiewicz inequality as follows:

Lemma 8 (Lemma 15 in [9]): It holds that

$$\|\nabla V(\theta)\|_2^2 \ge C(\theta) \cdot (V^* - V(\theta)),$$

where

$$C(\theta) = \frac{2}{|S|} \min_s \pi_\theta(s) \min_{s,a} \pi_\theta(a|s)^2 \left(\sum_s d_{\rho,\pi_\theta}(s) \right)^2.$$

Furthermore, it is shown in [9] that the action probabilities under the soft-max parameterization are uniformly bounded away from zero if the exact PG is available.

Lemma 9 (Lemma 16 in [9]): Using the exact PG (Algorithm 1) with $\eta_t = \frac{2}{L}$ for the entropy regularized objective, it holds that $\inf_{t \ge 1} \min_{s,a} \pi_t(a|s) > 0$.

Remark 2: Note that by Algorithm 1, $\inf_{t \ge 1} \min_{s,a} \pi_t(a|s)$ is only dependent on the initialization θ_1 and step-size (apart from problem dependent constants). Hence hereafter we denote $c(\theta_1) := \inf_{t \ge 1} \min_{s,a} \pi_t(a|s)$.

With Lemmas 3, 8 and 9, it is shown in Theorem 6 of [9] that the convergence rate for the entropy regularized PG is $V^* - V(\theta_t) \le C e^{-Ct}$, where the value of C depends on $\inf_{t \ge 1} \min_{s,a} \pi_t(a|s) > 0$ and $\{\pi_t(a|s)\}_{t=1}^\infty$ is generated by Algorithm 1. With a bad initialization θ_1, $\min_{s,a} \pi_1(a|s)$ could be very small and result in a slow convergence rate. When studying the stochastic PG, 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 PG. The iterates of stochastic gradient methods may indeed escape to in nity in general, rendering the entire scheme of stochastic approximation useless [32, 33]. In particular, when using the stochastic truncated PG for the entropy regularized RL, the key result of Lemma 9 may no longer hold true. This in turn results in the loss of gradient domination condition in guaranteeing the global convergence.

B. 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 0) with 2 actions, 2 parameters (θ_1, θ_2), the reward vector $r = (2, 1)$ and the regularization parameter $\tau = 1$. Then, the entropy-regularized value function can be written as $\tilde{r} = r - \tau \log \pi$.

Fig. 1. Landscape of $\tilde{r} = r - \tau \log \pi$.

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

When the stochastic PG 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 globally optimal convergence for the stochastic PG algorithm if the initial point is at the flat region.

C. Convergence to the first-order stationary point

Before presenting our main result, we first show that the stochastic PG proposed in Algorithm 4 asymptotically converges to a region where the PG vanishes almost surely if a specific adaptive step-size sequence is used.

Lemma 10: Suppose that the sequence $\{\theta_t\}_{t=1}^{\infty}$ is generated by Algorithm 4 for the entropy regularized objective with the step-sizes satisfying $\sum_{t=1}^{\infty} \alpha_t = \infty$, $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$ and $\alpha_t \le \frac{2}{L}$ for all $t = 1, 2, \ldots$. It holds that $\lim_{t\to\infty} \|\nabla V_\tau(\theta_t)\|_2 = 0$ with probability 1.

This result follows from classic results for the Robbins-Monro algorithm [34, 35, 36] when an unbiased PG estimator with the bounded variance, as in Algorithm 4, is used in the update rule. No requirement on the batch size is needed in Lemma 10. We now provide the proof of Lemma 10 below.

Proof.

To prove Lemma 10, it suffices to check the conditions in Proposition 1 (see Appendix B) for the objective function $V_\tau(\theta)$ and the update rule $\theta_{t+1} = \theta_t + \alpha_t (u_t + w_t)$, where $u_t = \nabla V_\tau(\theta_t)$ and $w_t = \widehat{\nabla} V_\tau(\theta_t) - \nabla V_\tau(\theta_t)$.

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

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

□

However, since the entropy-regularized value function $V_\tau(\theta)$ is not coercive in θ and it may be the case that the gradient $\nabla V_\tau(\theta_t)$ diminishing to 0 corresponds to θ_t going to infinity instead of converging to a stationary point. In addition, the existing results [32, 35, 36] on the almost surely stationary point convergence rely on the assumption that the trajectories of the process are bounded, i.e., $\sup_{t\ge 0} \|\theta_t\|_2 < \infty$, almost surely. This assumption is proven to hold when the function is coercive [37]. 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.

V. Main result

To overcome the non-coercive landscape challenge, we propose a two-phase stochastic PG algorithm (Algorithm 5). In the first phase, we will use a large batch size to control the estimation error to guarantee that the stochastic PG is informative even in the regime where the landscape is almost flat. After a certain number of iterations, which is a constant with respect to the optimality gap, the iteration will reach a region where the landscape has enough curvature information. Then, in the second phase, a small batch size is enough to guarantee a fast convergence to the optimal policy.

Before presenting the main result, we first introduce some helpful definitions. Let $D_t = V_\tau^* - V_\tau(\theta_t)$ denote the sub-optimality gap. Since the optimal policy of (1) is unique [11], there must exist a continuum of optimal solutions

$$\pi^* = \{\pi^* : \pi^*(a|s) = \frac{\exp(\theta^*_{s,a})}{\sum_{a'\in\mathcal{A}} \exp(\theta^*_{s,a'})}, \theta^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}\}.$$

In addition, we use π_τ and π^* interchangeably to denote the optimal policy of the entropy-regularized RL. Let $\{\theta_t\}_{t=1}^{T}$ denote the iterates of the algorithm with the exact PG (Algorithm 1) with $\alpha_t = \frac{1}{2L}$ starting from the initial point θ_1. For the soft-max parameterization, we have $\theta_{s,a} = \log \pi(a|s) + C_s$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, where $\{C_s\}_{s=1}^{|\mathcal{S}|}$ are some constants. Then, we have

$$\min_{\pi^* \in \Pi^*} \|\theta_t - \theta^*\|_2^2 = \|\log \pi_t - \log \pi^*\|_2^2, \quad \text{for all } t = 1, 2, \ldots.$$

Furthermore, by Lemma 9, we can define $Y = \log c_{\theta_1, \pi^*} - \log \pi^* Y_2$, where $c_{\theta_1,\pi^*} > 0$ is defined in Remark 2. Note that Y is only dependent on θ_1 and π^* (apart from problem dependent constants), and $\|\log \pi_t - \log \pi^*\|_2 \le B$ for any θ_1 and $\alpha \le \frac{1}{2L}$.

In addition, with a fair degree of hindsight and for some $\epsilon > 0$, we define the stopping time for the iterates $\{\pi_t\}_{t\geq 1}$ as

$$\min\{t \geq 0 \mid \min_{s\in\mathcal{S}} V_t(s) \geq \tfrac{1}{2}(1-\gamma)\epsilon\}, \tag{6}$$

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

$$\mathcal{G}_0 := \{\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|} \mid \min_{s\in\mathcal{S}} V_\pi(s) \leq \tfrac{1}{2}(1-\gamma)\epsilon\}.$$

Finally, we define $\hat{\tau}_t := \min_{s\in\mathcal{S}} V_t(s)$. We are now ready to present the main result.

Theorem 1: Consider an arbitrary tolerance level $\epsilon > 0$ and a small enough tolerance level $\epsilon_0 > 0$. For every initial point π_1, if π_{T+1} is generated by Algorithm 5 with

$$T_1 := \left\lceil \frac{6\hat{\tau}_1}{\epsilon_0} \right\rceil \frac{8L}{C_0 \ln 2}, \quad T_2 := \left\lceil \frac{t_0}{6} \right\rceil t_0, \quad T := T_1 + T_2,$$

$$B_1 := \max\left\{\frac{30}{C_0\epsilon_0^2}, \frac{6T_1 \log T_1}{L}\right\}, \quad B_2 := \frac{2\ln^2 T_2 \, t_0}{6C},$$

$$\eta_t = \min\left\{\frac{\log T_1}{T_1 L}, \frac{8}{C_0}, \frac{1}{2L}\right\} \quad \text{for } t \leq T_1,$$

$$\eta_t = \frac{1}{t - T_1 + t_0} \quad \text{for } t > T_1$$

where

$$\epsilon_0 := \min\left\{\frac{1}{2}, \frac{\min_s \pi_1(s)^2}{6\ln 2}, \exp\left(\frac{r}{1-\gamma}\right)^{-4}\right\}; \tag{7}$$

$$t_0 := \left\lceil \frac{3}{2C_0} \right\rceil; \tag{8}$$

$$C := \frac{2}{|\mathcal{S}|}\min_s \pi_1(s)(1-\gamma)^2 \min_{s,a} \pi_1(a|s)^2 \left\lfloor \frac{d_\mu}{|\mathcal{S}|} \right\rfloor^{-1} > 0, \tag{9}$$

$$C_0 := \frac{2}{|\mathcal{S}|}\left\lfloor \frac{d_\mu}{|\mathcal{S}|} \right\rfloor^{-1} \min_s \pi_1(s) \min_{\pi\in\mathcal{G}_0}\min_{s,a} \pi(a|s)^2; \tag{10}$$

then we have $\mathbb{P}(\hat{\tau}_{T+1} \leq \epsilon) \geq 1-\delta$. In total, it requires $\mathcal{O}(\epsilon^{-2}\delta^{-2})$ samples to obtain an ϵ-optimal policy with high probability.

Algorithm 5 Two-phase stochastic PG for entropy regularized RL

1: Inputs: η, π_1, B_1, B_2, T_1, T, $\{\tilde{\tau}_t\}_{t=1}^T$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: **if** $t \leq T_1$ **then**
4: $B = B_1$
5: **else**
6: $B = B_2$
7: **end if**
8: Run lines 3-7 in Algorithm 4
9: **end for**
10: Outputs: π_T.

A. Discussion

In Theorem 1, we have derived strong last-iterate complexity bounds (in contrast to the predominant running-min and ergodic complexity bounds in the reinforcement learning literature), with the desirable $\mathcal{O}(\epsilon^{-2})$ dependency on the targeting tolerance ϵ. That being said, the polynomial dependency on $\frac{1}{1-\gamma}$ and exponential dependencies on other problem- and algorithm-dependent constants also indicate that our bounds may not be tight in general.

The convergence analysis of the stochastic softmax PG with the entropy regularization is challenging[2] due to the weaker regularization effect of the entropy regularization (compared to the log-barrier regularization adopted in previous works on global optimality convergence of policy gradient methods [15, 38]), as well as the "softmax gravity well" induced by the softmax parameterization which has also been observed in the exact gradient setting [9, 39]. In particular, it only entails uniform gradient domination properties for policies that are bounded below uniformly (cf. Lemma 8). We thus need to control the trajectory to ensure that π_t remains in the region where it is uniformly bounded from below for all t. However, even with large batches, it is generally difficult to control stochastic trajectories, which eventually leads to the polynomial dependency on $\frac{1}{1-\gamma}$ and the exponential dependencies on some constants. If large batches are not used, then the trajectories would be even harder to control and no guarantees may be attained unless additional structural assumptions are enforced on the underlying MDP.

In the next three sections, we provide the proof of Theorem 1. We begin by showing that the iterates will converge to a neighborhood of the optimal solution with high probability in Section VI, and then utilize the curvature information around the optimal policy to guarantee that the action probabilities will still remain uniformly bounded with high probability in Section VII. We then combine the two steps to prove Theorem 1 in Section VIII.

VI. Global convergence with arbitrary initialization

In this section, we provide the first step towards the proof of Theorem 1. In particular, we will prove that after the first phase of Algorithm 5, the iterates will converge to a neighborhood of the optimal solution with high probability due to the use of a large batch size.

With a large batch size, we can show that if the iterations with the exact PG are bounded, then the iterations with the unbiased stochastic PG will remain bounded with high probability. This will further imply that the unbiased stochastic PG will converge to the neighborhood of the globally optimal policy with high probability. This is a non-trivial result involving the stopping/hitting time analysis, as presented below.

Lemma 11: Consider arbitrary tolerance levels $\epsilon > 0$ and $\epsilon_0 > 0$. For every initial point π_1, if π_T is generated by Algorithm 4

[2] Note that similar difficulties in generalization from exact policy gradients to stochastic policy gradients have been observed in [13] which states that "unlike the true gradient setting, geometric information cannot be easily exploited in the stochastic case for accelerating policy optimization without detrimental consequences or impractical assumptions".

with $\eta_t = B\min\left\{\frac{\log T_1}{T_1 L}, \frac{8}{C'}, \frac{1}{2L}\right\}$, $T_1 = \left\lceil \frac{6D(x_1) - \Phi^*}{\varepsilon_0} \cdot \frac{8L}{C'\ln 2} \right\rceil$, and $B_1 = \max\left\{\frac{30\sigma^2}{C'\varepsilon_0}, \frac{6}{L}\right\} T_1 \log T_1$, then we have $\mathbb{P}(D(x_{T_1}) \le B\varepsilon_0) \ge 1 - 2\delta$.

A. Helpful lemmas

Lemma 12: Suppose that $f(x)$ is L-smooth. Given $0 < \eta_t \le \frac{1}{2L}$ for all $t \ge 1$, let $\{x_t\}_{t=1}^{T+1}$ 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

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

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

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

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

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

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

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

This completes the proof. ∎

To prove Lemma 11, we first consider the case when A_{T_1}, where is defined in (6). Conditioning on this event, we can use Lemma 8 to show that $D(x_t)$ is linearly convergent up to some aggregated estimation error.

Lemma 13: If $\eta_t = B\frac{1}{2L}$, then $\mathbb{E}[D(x_{T_1+1}) \mid A_{T_1}] \le \left(1 - \frac{C'}{8}\eta\right)^{T_1+1} D(x_1) + \frac{5\sigma^2}{8C'B_1}$.

Proof. Let $e_t = \nabla V(x_t) - u_t$, where $u_t = \frac{1}{B_1}\sum_{i=1}^{B_1} \widehat{\nabla} V(x_t; i)$ and $\widehat{\nabla} V(x_t; i)$ is an unbiased estimator of $\nabla V(x_t)$. Let $\mathcal{F}_t$ denote the sigma field generated by the randomness up to iteration t. We define $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$ as the expectation operator conditioned on the sigma field $\mathcal{F}_t$. Since $\nabla V(\cdot)$ is L-smooth due to Lemma 3, it follows from Lemma 12 in the supplementary material:

$$\begin{aligned}
&\mathbb{E}_t[D(x_{t+1}) - D(x_t)] \le 1_{A_t} \\
&\mathbb{E}_t[V(x_{t+1}) - V(x_{t+1})] \le 1_{A_t} \\
&B\mathbb{E}_t\left[-\frac{\eta_t}{8}\|u_t\|_2^2 + \frac{3}{4}\eta_t\|e_t\|_2^2\right] \le 1_{A_t} \\
&B\mathbb{E}_t\left[-\frac{\eta_t}{8}\|\nabla V(x_t) - \nabla V(x_t)\|_2^2 + \frac{3}{4}\eta_t\|e_t\|_2^2\right] \le 1_{A_t} \\
&\mathbb{E}_t\left[-\frac{\eta_t}{8}\|\nabla V(x_t) - \nabla V(x_t)\|_2^2 + \frac{5}{8}\eta_t\|e_t\|_2^2\right] \le 1_{A_t} \\
&B\mathbb{E}_t\left[-\frac{C'\eta_t}{8}D(x_t) + \frac{5}{8}\eta_t\|e_t\|_2^2\right] \le 1_{A_t},
\end{aligned}$$

for every $\eta = B\frac{1}{2L}$, where the second inequality uses the fact that u_t is an unbiased estimator of $\nabla V(x_t)$ and the last inequality is due to Lemma 8. We now consider two cases:

- Case 1: Assume that A_t, which implies that $\eta_t > G'$ and $C'(x_t) \ge C'$. Then, we have $\mathbb{E}[D(x_{t+1}) \mid \mathcal{F}_t] \le B\left(1 - \frac{C'}{8}\eta\right) D(x_t) + \frac{5}{8}\mathbb{E}[\eta_t\|e_t\|_2^2 \mid \mathcal{F}_t]$.
- Case 2: Assume that B t which leads to $\mathbb{E}[D(x_{t+1}) \mid \mathcal{F}_t] \le 1_{A_t} \le 0$.

Now combining the above two cases yields the inequality

$$\begin{aligned}
&\mathbb{E}[D(x_{t+1}) \mid \mathcal{F}_t] \le 1_{A_t} \\
&B\left(1 - \frac{C'}{8}\eta\right) D(x_t) + \frac{5}{8}\mathbb{E}[\eta_t\|e_t\|_2^2 \mid \mathcal{F}_t] + 1_{A_t} \\
&B\left(1 - \frac{C'}{8}\eta\right) D(x_t) + 1_{A_t} + \frac{5}{8}\mathbb{E}[\eta_t\|e_t\|_2^2 \mid \mathcal{F}_t].
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[D(x_{t+1}) 1_{A_{t+1}} \mid \mathcal{F}_t] &\le B\mathbb{E}[D(x_{t+1}) 1_{A_t} \mid \mathcal{F}_t] \\
&\le \mathbb{E}[D(x_{t+1}) \mid \mathcal{F}_t] 1_{A_t},
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}[D(x_{t+1}) 1_{A_{t+1}}] \\
&B\sum_{i=0}^{t}\left(1 - \frac{C'}{8}\eta\right)^i \frac{5}{8}\mathbb{E}[\eta_i\|e_i\|_2^2] + \left(1 - \frac{C'}{8}\eta\right)^{t+1} D(x_1) \\
&B\left(1 - \frac{C'}{8}\eta\right)^{t+1} D(x_1) + \frac{5\sigma^2}{C'B_1}.
\end{aligned}$$

By setting $t = T_1 - 1$, we obtain that $\mathbb{E}[D(x_{T_1+1}) 1_{A_{T_1}}] \le B\left(1 - \frac{C'}{8}\eta\right)^{T_1} D(x_1) + \frac{5\sigma^2}{C'B_1}$. This completes the proof. ∎

We now establish that $\{x_t\}_{t=1}^{T+1}$ will be bounded with high probability if the large batch size is used.

Lemma 14: It holds that $\mathbb{P}(\cap_{t=1}^{T_1+1}\{\|x_t - x_1\| \le B\}) \ge 1 - \frac{T_1(T_1+1)L\eta^2}{B_1}$.

Proof. By the triangle inequality and the fact that the iterations of the algorithm with the exact PG are bounded by γ, we have

$$\|x_t - x_1\| \le \|x_t - x_1\| \le \min_{x^* \in X^*}\|x_t - x^*\| + \|x_1 - x^*\| \le 2\|x_1 - x^*\|.$$

Using the update rule of the algorithm with the exact PG ©V ¡ˆ • and the stochastic PG G_i $\frac{1}{B_1} P_{j\,1}^{B_1}$ ©V ¡;j ˆ •, one can write

$$
\begin{aligned}
d\hat{}\;{}_i\bullet \;\;]\hat{}\;{}_1 \; \prod_{i\,1}^{t\,1} {}_i u_i \bullet \;\; \hat{}\;{}_1 \prod_{i\,1}^{t\,1} {}_i ©V\;{}^{i\,\hat{}}\;\bullet\bullet]_2 \\
B \prod_{i\,1}^{t\,1} {}_i [u_i \;\; ©V\;{}^{i\,\hat{}}\;\bullet[_2 \\
\prod_{i\,1}^{t\,1} {}_i [u_i \;\; ©V\;{}^{i\,\hat{}}\;\bullet \;\; ©V\;{}^{i\,\hat{}}\;\bullet \;\; ©V\;{}^{i\,\hat{}}\;\bullet[_2 \\
B \prod_{i\,1}^{t\,1} {}_i Ye_i Y_2 \;\; \prod_{i\,1}^{t\,1} {}_i L\, Z_i \;\; {}_i Z_2 \quad :
\end{aligned}
$$

By expanding Z_i ${}_i Z_2$ recursively, it can be concluded that

$$
\begin{aligned}
&d\hat{}\;{}_t\bullet \\
&B \prod_{i\,1}^{t\,1} {}_i Ye_i Y_2 \;\; {}_{t\,1} L\, Z_{t\,1} \;\; {}_{t\,1} Z_2 \;\; \prod_{i\,1}^{t\,2} {}_i L\, Z_i \;\; {}_i Z_2 \\
&B \prod_{i\,1}^{t\,1} {}_i Ye_i Y_2 \;\; {}_{t\,1} L \prod_{i\,1}^{t\,2} {}_i Ye_i Y_2 \;\; {}_{t\,1} L^2 \prod_{i\,1}^{t\,2} {}_i Z_i \;\; {}_i Z_2 \\
&\quad \prod_{i\,1}^{t\,2} {}_i L\, Z_i \;\; {}_i Z_2 \\
&\prod_{i\,1}^{t\,1} {}_i Ye_i Y_2 \;\; {}_{t\,1} L \prod_{i\,1}^{t\,2} {}_i Ye_i Y_2 \\
&\quad \prod_{i\,1}^{t\,2} ‰_i L \;\; {}_{t\,1}\,{}_i L^2 \check{Z} Z_i \;\; {}_i Z_2 \\
&B \prod_{i\,1}^{t\,1} {}_i Ye_i Y_2 \;\; {}_{t\,1} L \prod_{i\,1}^{t\,2} {}_i Ye_i Y_2 \\
&\quad ‰_{t\,2} L \;\; {}_{t\,1}\,{}_{t\,2} L^2 \check{Z} \prod_{i\,1}^{t\,3} {}_i Ye_i Y_2 \\
&\quad \prod_{i\,1}^{t\,3} ‰‰_{t\,2} L \;\; {}_{t\,1}\,{}_{t\,2} L^2 \check{Z} \;\; {}_i L \\
&\qquad ‰_i L \;\; {}_{t\,1}\,{}_i L^2 \check{Z}\check{Z} Z_i \;\; {}_i Z_2 \\
&B \;\; \prod_{i\,1}^{t\,1} {}_i Ye_i Y_2 \;\; {}_{t\,1} L \prod_{i\,1}^{t\,2} {}_i Ye_i Y_2 \\
&\quad ‰_{t\,2} L \;\; {}_{t\,1}\,{}_{t\,2} L^2 \check{Z} \prod_{i\,1}^{t\,3} {}_i Ye_i Y_2 \\
&\quad \prod_{i\,1}^{t\,4} ‰‰_{t\,2} L \;\; {}_{t\,1}\,{}_{t\,2} L^2 \check{Z} \;\; {}_{t\,3} L \\
&\qquad ‰_{t\,3} L \;\; {}_{t\,1}\,{}_{t\,3} L^2 \check{Z}\check{Z} Ye_i Y_2 \\
&\quad \prod_{i\,1}^{t\,4} ‰‰_{t\,2}\,{}_{t\,3} L^2 \;\; {}_{t\,1}\,{}_{t\,2}\,{}_{t\,3} L^3 \check{Z} \;\; {}_i L \\
&\qquad ‰_{t\,3} L \;\; {}_{t\,1}\,{}_{t\,3} L^2 \check{Z} \;\; {}_i \check{Z} Z_i \;\; {}_i Z_2 \\
&\prod_{i\,1}^{t\,1} {}_i \prod_{j\,i\,1}^{t\,1} M \;\hat{}1 \;\; {}_j L \bullet Ye_i Y_2 \quad :
\end{aligned}
$$

Then, by the de nition of in (6) and Markov inequality, we obtain

$$
\begin{aligned}
P\hat{}\;\; B\,T_1\bullet \;\; P\hat{}\;\; \max_{t > \tilde{}\,1;:::;T\,1\bullet} d\hat{}\;{}_t\bullet\; C\hat{}1 \;\; \frac{1}{\;}\bullet \;\; \bullet \\
B P\hat{}\; \prod_{i\,1}^{T_1\,1} {}_i \prod_{j\,i\,1}^{T_1\,1} M \;\hat{}1 \;\; {}_j L \bullet Ye_i Y_2 \;\;\; C\hat{}1 \;\; \frac{1}{\;}\bullet \;\; \bullet \\
B \frac{P_{i\,1}^{T_1\,1}\;{}_i L_{j\,i\,1}^{T_1\,1}\hat{}1 \;\; {}_j L \bullet E\, Ye_i Y_2}{1} \\
B \frac{\hat{}1 \;\; L\bullet^{T_1\,1} P_{i\,1}^{T_1\,1} E\, Ye_i Y_2}{\;} ;
\end{aligned}
$$

where we use the fact that ${}_t$ ¼ for all t > ˜1; 2; : : :•. Furthermore, since $E\, Ye_i Y_2$ B $E\, Ye_i Y_2^2$ B $\frac{\;}{B_1}$, we have $P\hat{}\;\; B\,T_1\bullet$ B $\frac{T_1\,\hat{}1\;\; L\,\bullet^{T_1\,1}}{B_1}$: This completes the proof. j

B. Proof of Lemma 11

Proof. By combining Lemmas 13 and 14, we obtain that

$$
\begin{aligned}
&P\hat{}D\hat{}\;{}_{T_1}\bullet\; C\;{}_0\bullet \\
&B P\hat{}\;\; A\,T_1; D\hat{}\;{}_{T_1}\bullet\; C\;{}_0\bullet \;\;\; P\hat{}\;\; B\,T_1; D\hat{}\;{}_{T_1}\bullet\; C\;{}_0\bullet \\
&B \frac{E\;1\;{}_{A T_1} D\hat{}\;{}_{T_1}\bullet}{0} \;\;\; P\hat{}\;\; B\,T_1\bullet \\
&B Œ1 \;\; \frac{C\,0}{8}‘^{T_1\,1} \frac{D\hat{}\;{}_1\bullet}{0} \;\; \frac{5\;\;2}{C^0 B_1\;0} \\
&\qquad \frac{T_1\;\;\hat{}1 \;\;\; L\bullet^{T_1\,1}}{B_1} \\
&B Œ1 \;\; \frac{C\,0}{8}‘^{\frac{8}{C\,0}\frac{C\,0 T_1}{8}} \frac{D\hat{}\;{}_1\bullet}{0} \\
&\qquad \frac{5\;\;2}{C^0 B_1\;0} \;\;\; \frac{T_1\;\;\hat{}1 \;\;\; L\bullet^{T_1\,1}}{B_1} \\
&B \frac{1}{2}^{\frac{C\,0 T_1}{8}} \frac{D\hat{}\;{}_1\bullet}{0} \;\; \frac{5\;\;2}{C^0 B_1\;0} \;\;\; \frac{T_1\;\;\hat{}1 \;\;\; L\bullet^{T_1\,1}}{B_1} ;
\end{aligned}
$$

where the second inequality holds due to the Markov inequality, and the last inequality holds because ˆ1 $\frac{1}{m}\bullet^m$ B $\frac{1}{2}$ for all m C1 and $\frac{8}{C\,0}$ C1. By taking B min š$\frac{\log T_1}{T_1 L}$; $\frac{8}{C^0}$; $\frac{1}{2L}$Ÿ, we obtain

$$
\begin{aligned}
&P\hat{}D\hat{}\;{}_{T_1}\bullet\; C\;{}_0\bullet \\
&B \frac{1}{2}^{\frac{C\,0 \log T_1}{8L}} \frac{D\hat{}\;{}_1\bullet}{0} \;\; \frac{5\;\;2}{C^0 B_1\;0} \;\;\; \frac{\log T_1\;\;\hat{}1 \;\; \frac{\log T_1}{T_1}\bullet^{T_1\;1}}{B_1 L} \\
&B \frac{1}{2}^{\frac{C\,0 \log T_1}{8L}} \frac{D\hat{}\;{}_1\bullet}{0} \;\; \frac{5\;\;2}{C^0 B_1\;0} \\
&\qquad \frac{\log T_1\;\;\hat{}1 \;\; \frac{\log T_1}{T_1}\bullet^{\frac{T_1}{\log T_1}\log T_1}}{B_1 L} \\
&B \frac{1}{2}^{\frac{C\,0 \log T_1}{8L}} \frac{D\hat{}\;{}_1\bullet}{0} \;\; \frac{5\;\;2}{C^0 B_1\;0} \;\;\; \frac{\log T_1\;\; T_1}{B_1 L} \\
&B \frac{1}{T_1^{\frac{\ln 2\; C\,0}{8L}}} \frac{D\hat{}\;{}_1\bullet}{0} \;\; \frac{5\;\;2}{C^0 B_1\;0} \;\;\; \frac{\log T_1\;\; T_1}{B_1 L} ;
\end{aligned}
$$

where we have used $(1-x)^{1-x} \geq e^{-x}$ in the third inequality and $a^{\ln b} = b^{\ln a}$ in the last inequality. To guarantee $D(\theta_{T_1}) \leq C_0 B^2/2$, it suffices to have

$$T_1 \geq \frac{6D(\theta_1)}{\frac{\epsilon^2 (1-\gamma)^2 \ln 2}{8L}}, \quad B_1 \geq \max\left\{\frac{30}{C_0 (1-\gamma)^2}, \frac{6}{L}\right\} T_1 \log T_1 .$$

This completes the proof. □

VII. Uniformly bounded action probabilities given a good initialization

In this section, we will show how to utilize the curvature information around the optimal policy to guarantee that the action probabilities will still remain uniformly bounded with high probability, which serves as the second step towards the proof of Theorem 1.

Lemma 15: Given a tolerance level $\epsilon > 0$, let θ^* be the optimal policy of $V(\cdot)$. Assume further that Algorithm 4 is run for T_2 iterates with a step-size sequence of the form $\alpha_t = 1/(t+t_0)$ and a batch-size sequence $B_t = C/\alpha_t$ for all $t = 1, 2, \ldots, T_2$. If $t_0 \geq \frac{3\epsilon^2}{2C_0}$, and θ_1 is initialized in a neighborhood U_1 such that

$$U_1 \supseteq \{\theta : D(\theta) \leq B_0\}, \tag{11}$$

where $B_0 = \min\left\{\frac{\min_{s,a} \pi(s,a)}{6\ln 2}\epsilon^2, \exp\left(\frac{r}{1-\gamma}\right)\epsilon^4, 1\right\}$ and the constant $C > (0, 1]$, then the event

$$\mathcal{T}_1 = \{\min_{s,a} \pi_t(a|s) \geq (1-C)\min_{s,a} \pi(a|s), \ t = 1, 2, \ldots, T_2\} \tag{12}$$

occurs with probability at least $1 - 6\delta$.

A. Helpful lemmas

To prove Lemma 15, we first characterize the maximum amount by which $D(\theta_t)$ can grow at each step.

Lemma 16: Suppose that $\{\theta_t\}$ is generated by Algorithm 4 with $0 < \alpha_t \leq \frac{(1-\gamma)^3}{16\gamma^2 + 16\log|A|}$ for all $t \geq 1$. We have

$$D(\theta_{t+1}) \leq D(\theta_t) + \frac{\alpha_t C(\theta_t)}{4} \nabla D(\theta_t) + \frac{\alpha_t}{2}\|\epsilon_t\|_2^2, \tag{13}$$

where $\epsilon_t = \{\epsilon_t(s,a)\}$ and $\epsilon_t(s,a) = Q(\theta_t, s, a) - Q^{\theta_t}(s,a)$.

Proof. Since $V(\theta)$ is L-smooth in light of Lemma 3, it follows from Lemma 12 that

$$\begin{aligned}
&D(\theta_{t+1}) - D(\theta_t) \\
&\leq \frac{\alpha_t}{4}\|\nabla V(\theta_t)\|_2^2 + \frac{\alpha_t}{2}\|\epsilon_t\|_2^2 \\
&\leq \frac{\alpha_t}{4}\|\nabla V(\theta_t) - \nabla V(\theta^*) + \nabla V(\theta^*)\|_2^2 + \frac{\alpha_t}{2}\|\epsilon_t\|_2^2 \\
&\leq \frac{\alpha_t}{4}\|\nabla V(\theta_t) - \nabla V(\theta^*)\|_2^2 + \frac{\alpha_t}{4}\|\nabla V(\theta^*)\|_2^2 \\
&\quad + \frac{\alpha_t}{2}\langle \nabla V(\theta_t) - \nabla V(\theta^*), \epsilon_t\rangle + \frac{\alpha_t}{2}\|\epsilon_t\|_2^2 \\
&\leq \frac{\alpha_t}{4}\|\nabla V(\theta_t) - \nabla V(\theta^*)\|_2^2 + \frac{\alpha_t}{4}\|\nabla V(\theta^*)\|_2^2 \\
&\quad + \frac{\alpha_t}{2}\|\nabla V(\theta_t) - \nabla V(\theta^*)\|_2 \|\epsilon_t\|_2 + \frac{\alpha_t}{2}\|\epsilon_t\|_2^2 \\
&\leq \frac{\alpha_t}{4}\|\nabla V(\theta_t) - \nabla V(\theta^*)\|_2^2 + \frac{\alpha_t}{4}\|\nabla V(\theta^*)\|_2^2 \\
&\quad + \frac{\alpha_t}{2}\|\theta_t - \theta^*\|_2 \|\epsilon_t\|_2 + \frac{\alpha_t}{4}\|\epsilon_t\|_2^2 \\
&\leq \frac{\alpha_t C(\theta_t)}{4} D(\theta_t) + \frac{\alpha_t}{2}\|\theta_t - \theta^*\|_2 \|\epsilon_t\|_2 + \frac{\alpha_t}{4}\|\epsilon_t\|_2^2,
\end{aligned}$$

for every $\alpha_t \leq \frac{1}{2L}$, where the last inequality is due to Lemma 8. □

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

Similar as the techniques used in [22, 37, 40, 41, 42], we now encode the error terms in (13) as $M_n = \sum_{t=1}^{n} \alpha_t \epsilon_t$ and $S_n = \sum_{t=1}^{n} \frac{\alpha_t}{4}\|\epsilon_t\|_2^2$. Since $E_n = 0$, we have $E[M_n | \mathcal{F}_{n-1}] = M_{n-1}$. Therefore, $\{M_n\}$ is a zero-mean martingale; likewise, $E[S_n | \mathcal{F}_{n-1}] \geq S_{n-1}$, and therefore $\{S_n\}$ is a submartingale. The difficulty of controlling the errors in M_n and S_n lies in the fact that the estimation error ϵ_n may be unbounded. Because of this, we need to take a less direct, step-by-step approach to bound the total error increments conditioned on the event that $D(\theta_n)$ remains close to $D(\theta^*)$. We begin by introducing the "cumulative mean square error" $R_n = M_n^2 - S_n$. By construction, we have

$$\begin{aligned}
R_n &\leq (M_{n-1} + \alpha_n \epsilon_n)^2 - S_{n-1} - \frac{\alpha_n}{4}\|\epsilon_n\|_2^2 \\
R_{n-1} &= 2M_{n-1}\alpha_n \epsilon_n + \alpha_n^2 \|\epsilon_n\|_2^2 - \frac{\alpha_n}{4}\|\epsilon_n\|_2^2 .
\end{aligned}$$

Hence, $E[R_n - R_{n-1}] = 2M_{n-1}\alpha_n E[\epsilon_n] + \alpha_n^2 E[\|\epsilon_n\|_2^2] - \frac{\alpha_n}{4}E[\|\epsilon_n\|_2^2] \leq \frac{\alpha_n}{4}E[\|\epsilon_n\|_2^2] \leq C R_{n-1}$; i.e., $\{R_n\}$ is a submartingale. With a fair degree of hindsight, we define U as:

$$U = \{\theta : D(\theta) \leq 2B_0\} \cup U_0. \tag{14}$$

To condition it further, we also define the events

$$\begin{aligned}
\mathcal{E}_n &= \{\theta_t \in U \text{ for all } t = 1, 2, \ldots, n\} \\
\mathcal{E}_n' &= \{R_t \leq B_0 \text{ for all } t = 1, 2, \ldots, n\}
\end{aligned}$$

By definition, we also have $\mathcal{E}_0 = \mathcal{E}_0' = \Omega$ (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 θ_t has escaped the vicinity of θ^*.

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 17: Let θ^* be the optimal policy. Then, for all $n \geq 1, 2, \ldots$, the following statements hold:

1) $\mathcal{E}_{n+1} \subseteq \mathcal{E}_n$ and $\mathcal{E}_{n+1}' \subseteq \mathcal{E}_n'$.
2) $\mathcal{E}_{n+1}' \subseteq \mathcal{E}_n$.
3) Consider the "large noise" event

$$\begin{aligned}
\mathcal{E}_n'' = \mathcal{E}_n \setminus \mathcal{E}_{n+1}' = \{&\mathcal{E}_n \text{ and } \mathcal{E}_{n+1}' \text{ fail}\} \\
= \{&R_n > B_0 \text{ and } R_t \leq B_0 \text{ for all } t = 1, 2, \ldots, n-1 \text{ and } R_n > B_0\}
\end{aligned}$$

and let $R_n'' = R_n 1_{\mathcal{E}_{n+1}'}$ denote the cumulative error subject to the noise being "small" until time n. Then,

$$E[R_n''] \leq E[R_{n-1}''] + G^2 \alpha_n^2 + \frac{\alpha_n^2}{4B_0} P(\mathcal{E}_{n-1}'' | \mathcal{E}_{n-1}'). \tag{15}$$

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

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

- For the base case $n = 1$, we have $\theta_1 \in U$ because θ_1 is initialized in $U_1 \subseteq U$. Since $E_0 = \emptyset$, our claim follows.
- For the inductive step, assume that $E_{n-1} \subseteq \Theta_n$ for some $n \geq 1$. To show that $E_n \subseteq \Theta_{n+1}$, we fix a realization in E_n such that $R_t \leq \epsilon$ for all $t = 1, 2, \ldots, n$. Since $E_n \subseteq E_{n-1}$, the inductive hypothesis posits that Θ_n also occurs, i.e., $\theta_t \in U$ for all $t = 1, 2, \ldots, n$; hence, it suffices to show that $\theta_{n+1} \in U$. To that end, given that $\theta_t \in U$ for all $t = 1, 2, \ldots, n$, the distance estimate (13) readily gives $D(\theta_{t+1}) \leq D(\theta_t) + \eta_t \xi_t + \frac{\eta_t}{4} \|e_t\|_2^2$ for all $t = 1, 2, \ldots, n$. Therefore, after telescoping, we obtain

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

by the inductive hypothesis. This completes the induction.

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

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

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

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

and therefore

$$E[(R_n - R_{n-1}) 1_{E_{n-1}}] \tag{16}$$
$$= 2\eta_n E[M_{n-1} \xi_n 1_{E_{n-1}}] + \eta_n^2 E[\psi_n^2 1_{E_{n-1}}] + \frac{1}{4} \eta_n E[\|e_n\|^2 1_{E_{n-1}}].$$

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

- For the term in (16), by the unbiasedness of the gradient estimator shown in Lemma 4, we have $E[M_{n-1} \xi_n 1_{E_{n-1}}] = E[M_{n-1} 1_{E_{n-1}} E[\xi_n \mid \mathcal{F}_n]] = 0$.
- The second term in (16) is where the conditioning on E_{n-1} plays the most important role. It holds that:

$$\begin{aligned}
E[\psi_n^2 1_{E_{n-1}}] &= E[1_{E_{n-1}} E[\langle \hat{e}_n, \nabla V^{\pi}(\theta_n) \rangle^2 \mid \mathcal{F}_n]] \\
&\leq E[1_{E_{n-1}} \|\nabla V^{\pi}(\theta_n)\|^2 E[\|\hat{e}_n\|^2 \mid \mathcal{F}_n]] \\
&\leq E[1_{\Theta_n} \|\nabla V^{\pi}(\theta_n)\|^2 E[\|\hat{e}_n\|^2 \mid \mathcal{F}_n]] \\
&\leq G^2 \sigma^2
\end{aligned}$$

  where the first inequality is due to the Cauchy-Schwarz inequality, the second inequality follows from $E_{n-1} \subseteq \Theta_n$ and the last inequality results from Lemmas 2 and 5.
- Finally, for the third term in (16), we have:

$$\frac{\eta_n}{4} E[\|e_n\|_2^2 1_{E_{n-1}}] \leq \frac{\eta_n \sigma^2}{4B}. \tag{17}$$

Thus, putting together all of the above, we obtain $E[(R_n - R_{n-1}) 1_{E_{n-1}}] \leq G^2 \sigma^2 \eta_n^2 + \frac{\eta_n \sigma^2}{4B}$. Since $R_{n-1} \geq \epsilon$ if $\bar{E}_{n-1}$ occurs, we obtain $E[R_{n-1} 1_{\bar{E}_{n-1}}] \geq \epsilon E[1_{\bar{E}_{n-1}}] = \epsilon P(\bar{E}_{n-1})$. This completes the proof of Statement 3. ∎

With the above results, we can show that the cumulative mean square error R_n is small with high probability at all times.

Lemma 18: Consider an arbitrary tolerance level $\delta > 0$. If Algorithm 4 is run with a step-size schedule of the form $\eta_t = \gamma / (t + t_0)$ where $t_0 \geq \frac{3\gamma^2}{2\epsilon_0}$ and a batch size schedule $B_t \geq \frac{1}{t}$, we have $P(E_n) \geq 1 - \delta/6$ for all $n = 1, 2, \ldots$

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

$$\begin{aligned}
P(\bar{E}_n) &= P(E_{n-1} / E_n) = P(E_{n-1} \cap \{R_n > \epsilon\}) \\
&= E[1_{E_{n-1}} 1_{\{R_n > \epsilon\}}] \\
&\leq E[1_{E_{n-1}} (R_n / \epsilon)] = E[\tilde{R}_n / \epsilon],
\end{aligned}$$

which is derived by using the fact that $R_n \geq 0$ (so $1_{\{R_n > \epsilon\}} \leq R_n / \epsilon$). Now, by summing up (15), we conclude that $E[\tilde{R}_n] \leq E[\tilde{R}_0] + \frac{\sigma^2}{4B} \sum_{t=1}^{n} \eta_t - \epsilon \sum_{t=1}^{n} P(\bar{E}_{t-1})$. Hence, combining the above results, we obtain the estimate

$$\sum_{t=1}^{n} P(\bar{E}_k) \leq \frac{\sigma^2}{4B\epsilon_0} \sum_{t=1}^{n} \eta_t \leq \frac{\sigma^2}{4\epsilon_0} \sum_{t=1}^{n} \eta_t^2 \leq \frac{\Gamma \sigma^2}{4\epsilon_0};$$

where $\Gamma = \sum_{t=1}^{\infty} \eta_t^2 = \sum_{t=1}^{\infty} \gamma^2 / (t + t_0)^2$, and we have used the relations that $\tilde{R}_0 = 0$ and $\bar{E}_0 = \emptyset$ (by convention). By choosing $t_0 \geq \frac{3\gamma^2}{2\epsilon_0}$, we ensure that $\frac{\Gamma\sigma^2}{4\epsilon_0} < \delta/6$; moreover, since the events $\bar{E}_t$ are disjoint for all $t = 1, 2, \ldots$, we obtain $P(\cup_{t=1}^{n} \bar{E}_t) = \sum_{t=1}^{n} P(\bar{E}_t) \leq \delta/6$. Hence, $P(E_n) = P(\cap_{t=1}^{n} E_t^c) \geq 1 - \delta/6$ as claimed. ∎

Furthermore, we can show that the entropy-regularized value function $V^{\pi}(\theta)$ is locally quadratic around the optimal policy π^{*}.

Lemma 19: For every policy π, we have

$$D(\pi) \geq \frac{\min_s \rho(s)}{2 \ln 2} \sum \|\pi(a \mid s) - \pi^{*}(a \mid s)\|^2, \quad \forall s \in S, a \in A.$$

Proof. It follows from the soft sub-optimality difference lemma (Lemma 26 in [43]) that

$$\begin{aligned}
& V^{*}(\rho) - V^{\pi}(\rho) \\
&= \frac{1}{1-\gamma} \sum_s d(s) \cdot D_{KL}(\pi(\cdot \mid s) \| \pi^{*}(\cdot \mid s)) \\
&\geq \frac{1}{1-\gamma} \sum_s d(s) \cdot \frac{1}{2 \ln 2} \|\pi(\cdot \mid s) - \pi^{*}(\cdot \mid s)\|_1^2 \\
&\geq \frac{\lambda}{2 \ln 2} \sum_s \rho(s) \|\pi(\cdot \mid s) - \pi^{*}(\cdot \mid s)\|_1^2 \\
&\geq \frac{\lambda}{2 \ln 2} \sum_s \rho(s) \|\pi(\cdot \mid s) - \pi^{*}(\cdot \mid s)\|_2^2 \\
&\geq \frac{\lambda}{2 \ln 2} \rho(s) \|\pi(\cdot \mid s) - \pi^{*}(\cdot \mid s)\|_2^2, \quad \forall s \in S \\
&\geq \frac{\lambda \min_s \rho(s)}{2 \ln 2} \sum \|\pi(a \mid s) - \pi^{*}(a \mid s)\|^2, \quad \forall s \in S, a \in A;
\end{aligned}$$

where the first inequality is due to Theorem 11.6 in [41] stating that

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

for every two discrete distributions $P(\cdot)$ and $Q(\cdot)$. Moreover, the second inequality is due to $\|x(\cdot)\|_2 \leq \|x(\cdot)\|_1$ and the third inequality is due to the equivalence between ℓ_1-norm and ℓ_2-norm. This completes the proof. ∎

B. Proof of Lemma 15

Proof. Since the sequence $\{\epsilon_n\}$ is decreasing and $\epsilon_n \in (0, 1-\gamma)$ (by the second part of Lemma 17), Lemma 18 yields that $V(\pi_{T_2}) \geq \inf_n V(\pi_n) \geq \inf_n V(\pi_{\epsilon_n}) \geq V^* - \epsilon$ provided that t_0 is chosen large enough.

Now, it remains to show that $\mathbb{P}(\mathcal{A}_{T_2}) \geq 1-\delta$. We fix a realization in $\mathcal{A}_{T_2}$ such that $\|Q(\cdot,t)\|_2 \leq 2^{t_0}\sqrt{|\mathcal{A}|}$ for all $t = 1, 2, \ldots, T_2$. By Lemma 19, we have

$$\frac{\|\pi(\cdot,s)-\pi_{\epsilon_t}(\cdot,s)\|_2}{\sqrt{\frac{2\|Q(\cdot,t)\|_2^2\ln 2}{\min_a \pi(a,s)}}} \leq \sqrt{\frac{2^{2t_0}|\mathcal{A}|\ln 2}{\min_a \pi(a,s)}} \leq \sqrt{\frac{6^{t_0}|\mathcal{A}|\ln 2}{\min_a \pi(a,s)}} \leq \exp\left(\frac{r}{2(1-\gamma)}\right)\sqrt{\frac{|\mathcal{A}|}{\min_{s,a}\pi(a,s)}},$$

where the second inequality is due to the condition that the event $\mathcal{A}_{T_2}$ occurs, the third inequality is due to $t_0 \geq t_0$ when $t_0 \geq 1$, the forth inequality is due to the definition of t_0, and the last inequality is due to Theorem 1 in [45] where it holds that $\log\left(\pi(a,s)\right) \geq -\left(\frac{1}{2}+\frac{r}{1-\gamma}\right)\left(Q(\pi,s,a)+V(\pi,s)\right)$, $\forall(\pi,s,a)$.

Now, it can be easily verified that $\|\pi(a,s)-\pi_{\epsilon_t}(a,s)\|_2 \leq \|\pi(a,s)-\min_{s,a}\pi(a,s)\|_2$. For every $t > 1, 2, \ldots, T_2$, let $(s,a) \in \arg\min_{s,a}\pi_{\epsilon_t}(a,s)$. One can write

$$\min_{s,a}\pi_{\epsilon_t}(a,s) \geq \pi_{\epsilon_t}(a,s) - \|\pi(a,s)-\pi_{\epsilon_t}(a,s)\|_2 \geq \min_{s,a}\pi(a,s) - \|\pi(a,s)-\pi_{\epsilon_t}(a,s)\|_2 \geq (1-\gamma)\min_{s,a}\pi(a,s),$$

where the last inequality is due to $\pi(a,s) \geq \min_{s,a}\pi(a,s)$ for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Thus, we obtain $\mathbb{P}(\mathcal{A}_{T_2}) \geq \mathbb{P}(\mathcal{A}_{T_2}) \geq 1-\delta$. This completes the proof. ∎

VIII. Proof of Theorem 1

From Lemma 11, we conclude that, with a large batch size, the iteration will converge to a neighborhood of the optimal solution with high probability. From Lemma 15, we know that, with a good initialization, the policies will remain in the interior of the probability simplex with high probability. By combining the above two results, we are now ready to prove the sample complexity of the stochastic PG for entropy-regularized RL.

Proof. From Lemma 11, we can conclude that $\mathbb{P}(D(\pi^*,\pi_{T_1}) \leq \epsilon_0) \geq 1-\delta_1$ after the first phase. We then establish the algorithm's sample complexity when the initial policy of the second phase satisfies the good initialization condition $\mathbb{P}(D(\pi^*,\pi_{T_1}) \leq \epsilon_0) \geq 1-\delta_1$. It follows from Lemma 16 that

$$\mathbb{E}\left[D(\pi^*,\pi_{t+1})\mathbb{1}_{\mathcal{A}_{T_1}}\right] \leq \left(1-\frac{(1-\gamma)\alpha}{4}\right)\mathbb{E}\left[D(\pi^*,\pi_t)\mathbb{1}_{\mathcal{A}_{T_1}}\right] + \frac{\alpha}{2}\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_{T_1}}\right] + \frac{\alpha}{4}\mathbb{E}\left[\|e_t\|_2^2\mathbb{1}_{\mathcal{A}_{T_1}}\right],$$

for all $t \in T_1$, where $e_t := \widehat{\nabla V}(\pi_t) - \nabla V(\pi_t)$ and $\mathcal{A}_{T_1}$ is defined in (12). When the event $\mathcal{A}_{T_1}$ occurs, we have $\|\pi_t\|_2 \leq C$, where C is defined in (9). By taking the expectation, we can obtain

$$\begin{aligned}
&\mathbb{E}\left[\frac{\alpha}{2}\mathbb{1}_{\mathcal{A}_{T_1}}\right] \leq \frac{\alpha}{4}\mathbb{E}\left[\|e_t\|_2^2\mathbb{1}_{\mathcal{A}_{T_1}}\right] \\
&\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_{T_1}}\right] \leq \mathbb{E}\left[\frac{\alpha}{2}\mathbb{1}_{\mathcal{A}_{T_1}}\right] \leq \frac{\alpha}{4}\mathbb{E}\left[\|e_t\|_2^2|\mathcal{F}_t\right] \\
&\mathbb{E}\left[\mathbb{1}_{\mathcal{A}_{T_1}}\right] \leq \mathbb{E}\left[\frac{\alpha}{4}\|e_t\|_2^2|\mathcal{F}_t\right] \leq \frac{\alpha}{4B},
\end{aligned}$$

where the first equality is due to the fact that $\mathcal{A}_{T_1}$ is deterministic conditioning on $\mathcal{F}_t$, the second equality is due to the unbiasedness of π_t conditioning on $\mathcal{F}_t$, and the first inequality is due to (17). Therefore, $\mathbb{E}\left[D(\pi^*,\pi_{t+1})\mathbb{1}_{\mathcal{A}_{T_1}}\right] \leq \left(1-\frac{(1-\gamma)\alpha}{4}\right)\mathbb{E}\left[D(\pi^*,\pi_t)\mathbb{1}_{\mathcal{A}_{T_1}}\right] + \frac{\alpha^2}{4B}$. Arguing inductively yields that

$$\begin{aligned}
\mathbb{E}\left[D(\pi^*,\pi_{T_1})\mathbb{1}_{\mathcal{A}_{T_1}}\right] &\leq \prod_{i=1}^{T_2}\left(1-\frac{(1-\gamma)\alpha}{4}\right)D(\pi^*,\pi_1) + \sum_{i=1}^{T_2}\left(1-\frac{(1-\gamma)\alpha}{4}\right)^i\frac{\alpha^2}{4B} \\
&\leq \prod_{i=1}^{T_2}\left(1-\frac{(1-\gamma)\alpha}{4}\right)D(\pi^*,\pi_1) + \sum_{i=1}^{T_2}\frac{\alpha^2}{4B}.
\end{aligned}$$

By taking $\alpha_{T_1+i} = \frac{4}{C(i+t_0)}$, we obtain that

$$\begin{aligned}
\mathbb{E}\left[D(\pi^*,\pi_{T_1+T_2})\mathbb{1}_{\mathcal{A}_{T_1}}\right] &\leq \prod_{i=1}^{T_2}\left(\frac{i+t_0-1}{i+t_0}\right)D(\pi^*,\pi_{T_1}) + \frac{2}{C}\sum_{i=1}^{T_2}\frac{1}{i+t_0} \\
&\leq \frac{t_0}{T_2+t_0}D(\pi^*,\pi_{T_1}) + \frac{2\ln(T_2+t_0)}{C}.
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}(D(\pi^*,\pi_{T_1+T_2}) \geq \epsilon) \\
&\leq \mathbb{P}(D(\pi^*,\pi_{T_1+T_2}) \geq \epsilon, \mathcal{A}_{T_1}) + \mathbb{P}(D(\pi^*,\pi_{T_1+T_2}) \geq \epsilon, \mathcal{A}_{T_1}^c) \\
&\leq \mathbb{P}(D(\pi^*,\pi_{T_1+T_2}) \geq \epsilon, \mathcal{A}_{T_1})\mathbb{P}(\mathcal{A}_{T_1}) + \mathbb{P}(\mathcal{A}_{T_1}^c) \\
&\leq \frac{\mathbb{E}\left[D(\pi^*,\pi_{T_1+T_2})\mathbb{1}_{\mathcal{A}_{T_1}}\right]}{\epsilon}\mathbb{P}(\mathcal{A}_{T_1}) + \mathbb{P}(\mathcal{A}_{T_1}^c) \\
&\leq \frac{\mathbb{E}\left[D(\pi^*,\pi_{T_1})\mathbb{1}_{\mathcal{A}_{T_1}}\right]}{\epsilon} + \mathbb{P}(\mathcal{A}_{T_1}^c) \\
&\leq \frac{t_0}{\epsilon(T_2+t_0)}D(\pi^*,\pi_{T_1}) + \frac{2\ln(T_2+t_0)}{\epsilon C} + \delta_1,
\end{aligned}$$

where the second inequality follows from Lemma 15. To guarantee that $\mathbb{P}(D(\pi^*,\pi_{T_1+T_2}) \geq \epsilon) \leq \delta_1 + \frac{\delta}{2}$, it suffices to have $T_2 \geq \frac{t_0 D(\pi^*,\pi_{T_1})}{6\epsilon} - t_0$, $\frac{2\ln(T_2+t_0)}{\epsilon C} \leq \frac{\delta}{6}$. This completes the proof. ∎

IX. Conclusion

In this work, we studied the global convergence and the sample complexity of stochastic PG methods for the entropy-regularized RL with the soft-max parameterization.

We proposed two new (nearly) unbiased PG estimators for the entropy-regularized RL and proved that they have a bounded variance even though they could be unbounded. In addition, we developed a two-phase stochastic PG algorithm to overcome the non-coercive landscape challenge. This work provided the first global convergence result for stochastic PG methods for the entropy-regularized RL and obtained the sample complexity of $\widetilde{\mathcal{O}}(\epsilon^{-\frac{1}{2}})$, where ϵ is the optimality threshold. This work paves the way for a deeper understanding of other stochastic PG methods with entropy-related regularization, including those with trajectory-level KL regularization and policy reparameterization.

Acknowledgment

This work was funded by grants from AFOSR, ARO, ONR, NSF and C3.ai Digital Transformation Institute.

References

[1] R. J. Williams and J. Peng, "Function optimization using connectionist reinforcement learning algorithms," Connection Science, vol. 3, no. 3, pp. 241–268, 1991.

[2] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in International conference on machine learning, PMLR, 2016, pp. 1928–1937.

[3] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in International conference on machine learning, PMLR, 2018, pp. 1861–1870.

[4] B. O'Donoghue, R. Munos, K. Kavukcuoglu, and V. Mnih, "Combining policy gradient and q-learning," arXiv preprint arXiv:1611.01626, 2016.

[5] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in International Conference on Machine Learning, PMLR, 2017, pp. 1352–1361.

[6] H. Zang, X. Li, L. Zhang, P. Zhao, and M. Wang, "Teac: Intergrating trust region and max entropy actor critic for continuous control," https://openreview.net/references/pdf?id=bzTQQZQ6ix 2020.

[7] B. D. Ziebart, Modeling purposeful adaptive behavior with the principle of maximum causal entropy. Carnegie Mellon University, 2010.

[8] J. Schulman, P. Abbeel, and X. Chen, "Equivalence between policy gradients and soft q-learning," CoRR, vol. abs/1704.06440, 2017.

[9] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans, "On the global convergence rates of softmax policy gradient methods," in International Conference on Machine Learning, PMLR, 2020, pp. 6820–6829.

[10] G. Lan, "Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes," arXiv preprint arXiv:2102.00135, 2021.

[11] S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi, "Fast global convergence of natural policy gradient methods with entropy regularization," Operations Research, 2021.

[12] W. Chung, V. Thomas, M. C. Machado, and N. Le Roux, "Beyond variance reduction: Understanding the true impact of baselines on policy optimization," in International Conference on Machine Learning, PMLR, 2021, pp. 1999–2009.

[13] J. Mei, B. Dai, C. Xiao, C. Szepesvari, and D. Schuurmans, "Understanding the effect of stochasticity in policy optimization," Advances in Neural Information Processing Systems, vol. 34, 2021.

[14] Z. Ahmed, N. Le Roux, M. Norouzi, and D. Schuurmans, "Understanding the impact of entropy on policy optimization," in International Conference on Machine Learning, PMLR, 2019, pp. 151–160.

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

[16] L. Xiao, "On the convergence rates of policy gradient methods," arXiv preprint arXiv:2201.07443, 2022.

[17] L. Shani, Y. Efroni, and S. Mannor, "Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 5668–5675.

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

[19] J. Zhang, C. Ni, Z. Yu, C. Szepesvari, and M. Wang, "On the convergence and sample efficiency of variance-reduced policy gradient method," arXiv preprint arXiv:2102.08607, 2021.

[20] J. Zhang, A. Koppel, A. S. Bedi, C. Szepesvari, and M. Wang, "Variational policy gradient method for reinforcement learning with general utilities," in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 4572–4583.

[21] G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen, "Softmax policy gradient methods can take exponential time to converge," in Conference on Learning Theory, PMLR, 2021, pp. 3107–3110.

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

[23] Y. Liu, K. Zhang, T. Basar, and W. Yin, "An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods," Advances in Neural Information Processing Systems, vol. 33, 2020.

[24] Y. Ding, J. Zhang, and J. Lavaei, "On the global convergence of momentum-based policy gradient," arXiv preprint arXiv:2110.10116, 2021.

[25] B. Eysenbach and S. Levine, "If MaxEnt RL is the answer, what is the question?" arXiv preprint arXiv:1910.01913, 2019.

[26] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
[27] S. Cayci, N. He, and R. Srikant, "Linear convergence of entropy-regularized natural policy gradient with linear function approximation," arXiv preprint arXiv:2106.04096, 2021.
[28] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.
[29] K. Zhang, A. Koppel, H. Zhu, and T. Basar, "Global convergence of policy gradient methods to (almost) locally optimal policies," *SIAM Journal on Control and Optimization*, vol. 58, no. 6, pp. 3586–3612, 2020.
[30] R. S. Sutton, D. A. McAllester, S. P. Singh, Y. Mansour et al., "Policy gradient methods for reinforcement learning with function approximation." in *NIPs*, vol. 99. Citeseer, 1999, pp. 1057–1063.
[31] J. Baxter and P. L. Bartlett, "In nite-horizon policy-gradient estimation," *Journal of Arti cial Intelligence Research*, vol. 15, pp. 319–350, 2001.
[32] M. Benaïm, "Dynamics of stochastic approximation algorithms," in Seminaire de probabilites XXXIII Springer, 1999, pp. 1–68.
[33] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint* Springer, 2009, vol. 48.
[34] D. P. Bertsekas and J. N. Tsitsiklis, "Gradient convergence in gradient methods with errors," *SIAM Journal on Optimization*, vol. 10, no. 3, pp. 627–642, 2000.
[35] M. Benaïm and M. W. Hirsch, "Asymptotic pseudotrajectories and chain recurrent ows, with applications," *Journal of Dynamics and Differential Equations*, vol. 8, no. 1, pp. 141–176, 1996.
[36] H. J. Kushner and D. S. Clark, *Stochastic approximation methods for constrained and unconstrained systems* Springer Science & Business Media, 2012, vol. 26.
[37] P. Mertikopoulos, N. Hallak, A. Kavis, and V. Cevher, "On the almost sure convergence of stochastic gradient descent in non-convex problems," arXiv preprint arXiv:2006.11144, 2020.
[38] J. Zhang, J. Kim, B. O'Donoghue, and S. Boyd, "Sample ef cient reinforcement learning with REINFORCE," *35th AAAI Conference on Arti cial Intelligence*, 2021.
[39] J. Mei, C. Xiao, B. Dai, L. Li, C. Szepesvári, and D. Schuurmans, "Escaping the gravitational pull of softmax," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 130–21 140, 2020.
[40] Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos, "On the convergence of single-call stochastic extra-gradient methods," arXiv preprint arXiv:1908.08465, 2019.
[41] ——, "Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling," arXiv preprint arXiv:2003.10162, 2020.
[42] P. Mertikopoulos and Z. Zhou, "Learning in games with continuous action sets and unknown payoff functions," *Mathematical Programming*, vol. 173, no. 1, pp. 465–507, 2019.
[43] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans, "On the global convergence rates of softmax policy gradient methods," in *International Conference on Machine Learning* PMLR, 2020, pp. 6820–6829.
[44] T. M. Cover, *Elements of information theory* John Wiley & Sons, 1999.
[45] O. Nachum, M. Norouzi, K. Xu, and D. Schuurmans, "Bridging the gap between value and policy based reinforcement learning," arXiv preprint arXiv:1702.08892, 2017.
[46] Y. Ding, J. Zhang, and J. Lavaei, "On the global convergence of momentum-based policy gradient," arXiv preprint arXiv:2110.10116, 2021.
[47] R. Yuan, R. M. Gower, and A. Lazaric, "A general sample complexity analysis of vanilla policy gradient," arXiv preprint arXiv:2107.11433, 2021.
[48] D. P. Bertsekas and J. N. Tsitsiklis, "Gradient convergence in gradient methods with errors," *SIAM Journal on Optimization*, vol. 10, no. 3, pp. 627–642, 2000.

APPENDIX A
PROPERTIES OF STOCHASTIC POLICY GRADIENT

A. Proof of Lemma 4

Proof. We rst show the unbiasedness of the Q-estimate, i.e., $\mathbb{E}\,\hat{Q}(\hat{s},\hat{a}\,|\,s,a) = Q(\hat{s},\hat{a})$ for all $(\hat{s},\hat{a}) \in \mathcal{S}\times\mathcal{A}$ and $(s,a)\in\mathcal{S}\times\mathcal{A}$. In particular, from the de nition of $\hat{Q}(\hat{s},\hat{a})$, we have

$$\begin{aligned}
&\mathbb{E}\,\hat{Q}(\hat{s},\hat{a}\,|\,s,a)\\
&\quad\overset{(a)}{=}\mathbb{E}\Big(r(\hat{s},\hat{a}) + \sum_{h=1}^{H-1}\gamma^{h}\, r(\hat{s}_h,\hat{a}_h)\,\nabla\log\pi_\theta(\hat{a}_h|\hat{s}_h)\,\Big|\,s_0=s,a_0=a\Big)\\
&\quad\overset{(b)}{=}\mathbb{E}\Big(r(\hat{s},\hat{a}) + \sum_{h=1}^{a-1}\gamma^{h}\, r(\hat{s}_h,\hat{a}_h)\,\nabla\log\pi_\theta(\hat{a}_h|\hat{s}_h)\,\Big|\,s_0=s,a_0=a\Big),
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}\,\hat{Q}(\hat{s},\hat{a}\,|\,s,a) \qquad\qquad (18)\\
&\quad=\mathbb{E}_{H-1}\,\mathbb{E}\Big(r(\hat{s},\hat{a}) + \sum_{h=1}^{a-1}\gamma^{h}\, r(\hat{s}_h,\hat{a}_h)\,\nabla\log\pi_\theta(\hat{a}_h|\hat{s}_h)\,\Big|\,s_0=s,a_0=a,H-1\Big),
\end{aligned}$$

where the trajectory $(\hat{s},\hat{a},\hat{s}_1,\hat{a}_1,\ldots)$ equal to $(s_0,a_0,s_1,a_1,\ldots)$. The inner expectation over τ can be written as

$$\begin{aligned}
&\mathbb{E}\Big(r(\hat{s},\hat{a}) + \sum_{h=1}^{a-1}\gamma^{h}\, r(\hat{s}_h,\hat{a}_h)\,\nabla\log\pi_\theta(\hat{a}_h|\hat{s}_h)\,\Big|\,s_0=s,a_0=a,H-1\Big)\\
&\quad=\sum_{\tau}\Big(r(\hat{s},\hat{a}) + \sum_{h=1}^{a-1}\gamma^{h}\, r(\hat{s}_h,\hat{a}_h)\,\nabla\log\pi_\theta(\hat{a}_h|\hat{s}_h)\Big)\,P(\tau\,|\,s_0=s,a_0=a,H-1)\\
&\quad=r(\hat{s},\hat{a}) + \sum_{\tau}\sum_{h=1}^{a-1}\gamma^{h}\, r(\hat{s}_h,\hat{a}_h)\,\nabla\log\pi_\theta(\hat{a}_h|\hat{s}_h)\,P(\tau\,|\,s_0=s,a_0=a,H-1). \qquad (19)
\end{aligned}$$

By the de nition of the probability over the sample trajectory $P(\cdot)$, for every $h = 0, 1, 2, \ldots$, it holds that

$$\begin{aligned}
&S(s_h, a_h) \quad \log \pi(a_h|s_h) \quad P(\cdot) \; S \\
&S(s_h, a_h) \quad \log \pi(a_h|s_h) \quad \pi(a_h|s_h) \; P(s_1|s_0, a_0) \\
&\quad \pi(a_1|s_1) \cdots P(s_h|s_{h-1}, a_{h-1}) \; P(s_{h+1}|s_h, a_h) \cdots \\
&P(s_{H+\infty-1}|s_{H+\infty-2}, a_{H+\infty-2}) \quad \pi(a_{H+\infty-1}|s_{H+\infty-1}) \; S \\
&B r \quad \frac{}{e}.
\end{aligned}$$

where the last inequality follows from $P(s_{+\infty}|s, a) \; B \; 1$, $\pi(a|s) \; B \; 1$ for all $s, s_{+\infty} > S$ and $a > A$ together with $x \log x \; S \; B \; \frac{1}{e}$ for $x > 0, 1$. Thus, for each trajectory and $N \; A \; 0$, we have

$$\begin{aligned}
&\sum_{h=1}^{N} 1_{H+\infty \text{ ch } 0}^{h-2} \, r(s_h, a_h) \quad \log \pi(a_h|s_h) \quad P(\cdot) \\
&B \frac{1}{1-\gamma} < r \quad \frac{}{e} \cdot .
\end{aligned} \tag{20}$$

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

$$\begin{aligned}
&E \quad r(s_0, a_0) \cdot \sum_{h=1}^{a} 1_{H+\infty \text{ ch } 0}^{h-2} \, r(s_h, a_h) \cdot \\
&\qquad \log \pi(a_h|s_h) \cdot S \; s_0 \quad s, a_0 \quad a, H+\infty \\
&r(s_0, a_0) \cdot \sum_{h=1}^{a} \sum 1_{H+\infty \text{ ch } 0}^{h-2} \, r(s_h, a_h) \cdot \\
&\qquad \log \pi(a_h|s_h) \cdot P(\cdot) \cdot S \; s_0 \quad s, a_0 \quad a, H+\infty \\
&r(s_0, a_0) \cdot \sum_{h=1}^{a} E \; 1_{H+\infty \text{ ch } 0}^{h-2} r(s_h, a_h) \cdot \\
&\qquad \log \pi(a_h|s_h) \cdot S \; s_0 \quad s, a_0 \quad a, H+\infty .
\end{aligned}$$

In addition, for every $N \; A \; 0$, we have

$$\begin{aligned}
&\sum_{h=1}^{N} E \; 1_{H+\infty \text{ ch } 0}^{h-2} r(s_h, a_h) \cdot \quad \log \pi(a_h|s_h) \cdot \\
&S \; s_0 \quad s, a_0 \quad a, H+\infty \quad r(s_0, a_0) \cdot \\
&B \quad {}^{1-\gamma} \sum_{h=0}^{a} E \quad {}^{h} \, {}^{1-\gamma} r(s_h, a_h) \cdot \quad \log \pi(a_h|s_h) \cdot \\
&S \; s_0 \quad s, a_0 \quad a, H+\infty \quad r(s_0, a_0) \cdot \\
&B r(s_0, a_0) \cdot \quad {}^{1-\gamma} E_{s_1} \, V_{;, -2}(s_1) \cdot S s_0, a_0 \\
&B r \quad \frac{-2 r \quad \log S A S \cdot}{1 \quad -2} \\
&B \frac{r \quad \log S A S}{1 \quad -2}, ;
\end{aligned} \tag{21}$$

where the third inequality is due to the boundedness of the enrtopy-regularized value function in Lemma 1. Furthermore, since (21) is non-decreasing and the limit as $N \to a$ exists, by the Monotone Convergence Theorem, we can interchange the limit with the outer-expectation over $H+\infty$ in (18) as follows:

$$\begin{aligned}
&E \; \hat{Q}(s, a) \cdot S \; s, a \\
&r(s_0, a_0) \cdot \lim_{N \to a} E_{H+\infty} E \sum_{h=1}^{N} 1_{H+\infty \text{ ch } 0}^{h-2} \, r(s_h, a_h) \cdot \\
&\qquad \log \pi(a_h|s_h) \cdot S \; s, s_0 \quad s, a_0 \quad a, H+\infty \\
&r(s_0, a_0) \cdot \lim_{N \to a} \sum_{h=1}^{N} E \quad E_{H+\infty} 1_{H+\infty \text{ ch } 0}^{h-2} \, r(s_h, a_h) \cdot \\
&\qquad \log \pi(a_h|s_h) \cdot S \; s, s_0 \quad s, a_0 \quad a \\
&r(s_0, a_0) \cdot \lim_{N \to a} \sum_{h=1}^{N} E \quad {}^{h} \, r(s_h, a_h) \cdot \\
&\qquad \log \pi(a_h|s_h) \cdot S \; s, s_0 \quad s, a_0 \quad a \\
&r(s_0, a_0) \cdot E \sum_{h=1}^{a} {}^{h} \, r(s_h, a_h) \cdot \\
&\qquad \log \pi(a_h|s_h) \cdot S \; s, s_0 \quad s, a_0 \quad a \\
&Q(s, a) \cdot
\end{aligned} \tag{22}$$

Where we have also used the fact that $H+\infty$ is drawn independently from the trajectory in the first equality, the fact that $H+\infty$ Geom$(1 \quad {}^{1-\gamma})$ and thus $E_{H+\infty} 1_{H+\infty \text{ ch } 0}^{t-2} \quad {}^{t-2}$ in the second equality, and the interchangeability between the limit and the summation over the trajectory in the third equality. This completes the proof of the unbiasedness of $\hat{Q}(s, a)$.

Now, we are ready to show unbiasedness of the stochastic gradients $\hat{\nabla} V(\cdot)$. It follows from Lemma 2 that

$$\begin{aligned}
&E \; \hat{\nabla} V(\cdot) \; S \\
&E_{H, s_H, a_H} \cdot \rangle E_{H+\infty, s_{1:H+\infty}, a_{1:H+\infty}} \cdot \quad \hat{\nabla} V(\cdot) \\
&\qquad S \; s_0^{+\infty} \quad s_H, a_0^{+\infty} \quad a_H \quad S \cdot \\
&E_{H, s_H, a_H} \cdot < E_{H+\infty, s_{1:H+\infty}, a_{1:H+\infty}} \cdot \rangle \frac{1}{1} \nabla \log \pi(a_0^{+\infty}|s_0^{+\infty}) \\
&\%(s_0^{+\infty}, a_0^{+\infty}) \cdot \quad \log \pi(a_0^{+\infty}|s_0^{+\infty}) \cdot \check{Z} S \; s_0^{+\infty} \quad s_H, a_0^{+\infty} \quad a_H, \check{z} S \check{Z} \\
&E_{H, s_H, a_H} \cdot < \frac{1}{1} \nabla \log \pi(a_H|s_H) \cdot E_{H+\infty, s_{1:H+\infty}, a_{1:H+\infty}} \cdot \\
&\text{™}(s_0^{+\infty}, a_0^{+\infty}) \cdot \quad \log \pi(a_0^{+\infty}|s_0^{+\infty}) \cdot \check{Z} S \; s_0^{+\infty} \quad s_H, a_0^{+\infty} \quad a_H, \check{z} S \check{Z} \\
&E_{H, s_H, a_H} \cdot \rangle \frac{1}{1} \nabla \log \pi(a_H|s_H) \cdot \\
&\%(s_H, a_H) \cdot \quad \log \pi(a_H|s_H) \cdot \check{Z} S \check{z} .
\end{aligned}$$

where we have used (22) in the last equality. By using the identity function $1_{h \quad H}$, the above expression can be further written as

$$\begin{aligned}
&E \; \hat{\nabla} V(\cdot) \; S \\
&\frac{1}{1} E_{H, s_H, a_H} \cdot \text{œ} \sum_{h=0}^{a} 1_{h \quad H} \nabla \log \pi(a_H|s_H) \cdot \\
&\%(s_H, a_H) \cdot \quad \log \pi(a_H|s_H) \cdot \check{Z} S \check{z} .
\end{aligned} \tag{23}$$

Since for the softmax parameterization, we have

$$\frac{\partial \log \pi_\theta(s_H, a_H)}{\partial \theta_{s,a}} = \begin{cases} 1 - \pi_\theta(s_H, a_H), & (s, a) = (s_H, a_H), \\ -\pi_\theta(s_H, a_H), & (s, a) \neq (s_H, a_H). \end{cases}$$

Thus, the term $\sum_{h=0}^{H} \mathbb{1}_{h \le H} \cdot \log \pi_\theta(s_H, a_H) \cdot \left(Q_\omega(s_H, a_H) - \log \pi_\theta(s_H, a_H) \right)$ is uniformly bounded for every ω and non-decreasing with respect to H, we can interchange the limit and the expectation in (23) by the Monotone Convergence Theorem to obtain

$$\begin{aligned}
&\mathbb{E}\left[\widehat{\nabla} V(\theta)\right] \\
&= \sum_{h=0}^{\infty} \frac{P(H \ge h)}{1-\gamma} \mathbb{E}_{H, (s_H, a_H) \sim \tau(\theta)} \left[\nabla \log \pi_\theta(s_H, a_H) \cdot \left(Q_\omega(s_H, a_H) - \log \pi_\theta(s_H, a_H) \right) \right] \\
&= \sum_{h=0}^{\infty} \mathbb{E}_{(s_h, a_h) \sim \tau(\theta)} \left[\nabla \log \pi_\theta(s_H, a_H) \cdot \left(Q_\omega(s_H, a_H) - \log \pi_\theta(s_H, a_H) \right) \right] \\
&= \sum_{h=0}^{\infty} \sum_{s,a} P(s_h = s, a_h = a \mid s_0) \cdot \nabla \log \pi_\theta(s, a) \cdot \left(Q_\omega(s, a) - \log \pi_\theta(s, a) \right) \\
&= \sum_{s,a} \nabla \log \pi_\theta(s, a) \cdot \left(Q_\omega(s, a) - \log \pi_\theta(s, a) \right) \\
&= \frac{1}{1-\gamma} \mathbb{E}_{(s,a) \sim \nu(\cdot)} \left[\nabla \log \pi_\theta(s, a) \cdot \left(Q_\omega(s, a) - \log \pi_\theta(s, a) \right) \right],
\end{aligned}$$

where the second equality is due to the fact that $\mathrm{Geom}(1-\gamma)$ and thus $P(h \le H) = (1-\gamma)^h$, 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 $\widehat{\nabla} V(\theta)$. ∎

B. Proof of Lemma 5

Proof. We first note that the policy gradient estimator $\widehat{\nabla} V(\theta)$ can be decomposed as:

$$\begin{aligned}
&\nabla \log \pi_\theta(s_H, a_H) \cdot \left(Q_\omega(s_H, a_H) - \log \pi_\theta(s_H, a_H) \right) \\
&\nabla \log \pi_\theta(s_H, a_H) \cdot \left(\sum_{i=0}^{H-1} \gamma^i r(s_i, a_i) - \log \pi_\theta(s_H, a_H) \right),
\end{aligned}$$

where $H \sim \mathrm{Geom}(1-\gamma)$, $(s_H, a_H) \sim (s, a)$, $(s_0, a_0) \sim (s, a)$. To streamline the presentation, we introduce the following notations:

$$\begin{aligned}
g_1(s_H, a_H) &= \sum_{i=0}^{H-1} \gamma^i r(s_i, a_i), \\
g_2(s_H, a_H) &= \sum_{i=0}^{H-1} \nabla \log \pi_\theta(s_i, a_i),
\end{aligned}$$

Then, the policy gradient estimator $\widehat{\nabla} V(\theta)$ can be decomposed as:

$$\widehat{\nabla} V(\theta) = \frac{1}{1-\gamma} \nabla \log \pi_\theta(s_H, a_H) \left(g_1(s_H, a_H) - g_2(s_H, a_H) \right).$$

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

$$\begin{aligned}
&\mathrm{Var}\left[\widehat{\nabla} V(\theta)\right] \\
&\frac{\left\| \nabla \log \pi_\theta(s_H, a_H) \right\|^2}{(1-\gamma)^2} \mathrm{Var}\left[g_1(s_H, a_H) - g_2(s_H, a_H) \right] \\
&\frac{2 \left\| \nabla \log \pi_\theta(s_H, a_H) \right\|^2}{(1-\gamma)^2} \left(\mathrm{Var}\left[g_1(s_H, a_H) \right] + \mathrm{Var}\left[g_2(s_H, a_H) \right] \right) \\
&\frac{8}{(1-\gamma)^2} \left(\mathrm{Var}\left[g_1(s_H, a_H) \right] + \mathrm{Var}\left[g_2(s_H, a_H) \right] \right),
\end{aligned}$$

where the last inequality follows from $\left\| \nabla \log \pi_\theta(s_H, a_H) \right\| \le 2$ [46]. Since $g_1(s, a)$ is uniformly bounded, i.e., $|g_1(s, a)| \le \frac{r}{1-\gamma}$ for all $s \in S, a \in A$, we must have

$$\mathrm{Var}\left[g_1(s_H, a_H) \right] \le \frac{r^2}{(1-\gamma)^2 (1-\gamma^2)}.$$

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

$$\begin{aligned}
\|g_2\|^2 &\le \left(\sum_{i=0}^{H-1} \left\| \nabla \log \pi_\theta(s_i, a_i) \right\| \right)^2 \\
&\le \left(\sum_{i=0}^{H-1} \gamma^{i/4} \gamma^{-i/4} \left\| \nabla \log \pi_\theta(s_i, a_i) \right\| \right)^2 \\
&\le \left(\sum_{i=0}^{H-1} \gamma^{i/2} \right) \left(\sum_{i=0}^{H-1} \gamma^{-i/2} \left\| \nabla \log \pi_\theta(s_i, a_i) \right\|^2 \right) \\
&\le \frac{2}{1-\gamma^{1/2}} \left(\sum_{i=0}^{H-1} \gamma^{-i/2} \left\| \nabla \log \pi_\theta(s_i, a_i) \right\|^2 \right),
\end{aligned}$$

where the second inequality is due to the Cauchy-Schwarz inequality. By fixing the state action pairs (s_H, a_H) and the horizon H for now and taking expectation of $\|g_2\|^2$ only over the sample trajectory $\tau \sim p_\tau(\cdot)$, it holds that

$$\begin{aligned}
&\mathbb{E}_{\tau \sim p_\tau(\cdot)} \|g_2\|^2 \\
&\le \frac{2}{1-\gamma^{1/2}} \sum_{i=0}^{H-1} \mathbb{E}_{\tau \sim p_\tau(\cdot)} \left[\left\| \nabla \log \pi_\theta(s_i, a_i) \right\|^2 \right]. \qquad (24)
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}_{\tau \sim p_\tau(\cdot)} \left[\left\| \nabla \log \pi_\theta(s_i, a_i) \right\|^2 \right] \\
&\mathbb{E}_{s_1 \sim p_1(\cdot), a_0 \sim \pi_\theta(s_0, \cdot), \ldots, a_{H-1} \sim \pi_\theta(s_{H-1}, \cdot)} \left[\left\| \nabla \log \pi_\theta(s_i, a_i) \right\|^2 \right] \\
&\mathbb{E}_{s_1 \sim p_1(\cdot), a_0 \sim \pi_\theta(s_0, \cdot), \ldots, a_i \sim \pi_\theta(s_i, \cdot)} \left[\left\| \nabla \log \pi_\theta(s_i, a_i) \right\|^2 \right] \\
&\mathbb{E}_{s_1 \sim p_1(\cdot), a_0 \sim \pi_\theta(s_0, \cdot), \ldots, s_i \sim p(\cdot \mid s_{i-1}, a_{i-1}), a_i \sim \pi_\theta(s_i, \cdot)} \left[\sum_{a_i \in A} \pi_\theta(s_i, a_i) \left\| \nabla \log \pi_\theta(s_i, a_i) \right\|^2 \right].
\end{aligned}$$

By checking the optimality conditions for the optimization problem

$$\max \sum_{i=1}^{n} x_i \log x_i \quad \text{such that} \sum_{i=1}^{n} x_i = 1, \qquad (25)$$

it can be concluded that the maximizer for the constrained problem (25) is $x_1 = x_2 = \cdots = x_n = \frac{1}{n}$ and the maximum solution is $\log n$.

Thus, we have $\sum_{a_h \in \mathcal{A}} \hat{a}_h(s_h) \log \hat{a}_h^i(s_h) \geq \log |\mathcal{A}|$ and

$$\mathbb{E}_{\hat{p}_\theta}\Big[\log \hat{a}_h^i(s_h)\Big] \geq \mathbb{E}_{s_0, a_0 \sim \hat{s}_0, s_1 \sim \hat{p}_\theta(\cdot|s_0,a_0), \ldots, s_h \sim \hat{p}_\theta(\cdot|s_{h-1},a_{h-1})} \log \hat{a}_h^i(s_h) \geq -\log |\mathcal{A}|.$$

By substituting the above inequality into (24), we obtain that

$$\mathbb{E}_{\hat{p}_\theta}\Big[\|\nabla g_2\|^2\Big] \leq \frac{2}{(1-\gamma)^2} \sum_{i=0}^{H-1} \gamma^i \mathbb{E}_{\hat{p}_\theta}\Big[\big\|\nabla \log \hat{a}_i(s_i)\big\|^2\Big] \leq \frac{\log^2 |\mathcal{A}|}{(1-\gamma)^2} \sum_{i=0}^{H-1} \gamma^i \leq \frac{\log^2 |\mathcal{A}|}{(1-\gamma)^3},$$

for every $H \in \mathbb{N}$. By taking expectation of $\|\nabla g_2\|^2$ over the state action pair (s_H, a_H) and the horizon $H \sim \mathrm{Geom}(1-\gamma)$, it yields that

$$\mathbb{E}\Big[\|\nabla g_2\|^2\Big] \leq \frac{\log^2 |\mathcal{A}|}{(1-\gamma)^3},$$

which further implies that $\mathrm{Var}\big[\|\nabla g_2\|^2\big] \leq \frac{\log^2 |\mathcal{A}|}{(1-\gamma)^3}$. This completes the proof. ∎

C. Proof of Lemma 6

Proof. To simplify the notation, we denote $r_{j,\theta} = r_j(s_j, a_j) \log \hat{a}_j(s_j)$. By definition, we have

$$\mathbb{E}\big[\|\hat{V}^{\pi,H}(\theta) - V^{\pi}(\theta)\|_2\big] = \mathbb{E}\left[\left\|\sum_{h=0}^{H-1} \nabla \log \hat{a}_h(s_h) \sum_{j=0}^{h} r_{j,\theta} - \sum_{h=H}^{\infty} \nabla \log \hat{a}_h(s_h) \sum_{j=0}^{h} r_{j,\theta}\right\|\right].$$

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

$$\big[\mathbb{E}\big[\|\hat{V}^{\pi,H}(\theta) - V^{\pi}(\theta)\|_2\big]\big]^2 \leq \mathbb{E}\left[\left\|\sum_{h=0}^{H-1} \nabla \log \hat{a}_h(s_h) \sum_{j=0}^{h} r_{j,\theta}\right\|^2\right] \mathbb{E}\left[\left\|\sum_{h=H}^{\infty} \nabla \log \hat{a}_h(s_h) \sum_{j=0}^{h} r_{j,\theta}\right\|^2\right].$$

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

$$\mathbb{E}\big[\|\hat{V}^{\pi,H}(\theta) - V^{\pi}(\theta)\|_2\big] \leq 2\mathbb{E}\left[\left\|\sum_{h=0}^{H-1} \nabla \log \hat{a}_h(s_h) \sum_{j=0}^{h} r_{j,\theta}\right\|\right] \qquad (26)$$

$$2\mathbb{E}\left[\left\|\sum_{h=H}^{\infty} \nabla \log \hat{a}_h(s_h) \sum_{j=0}^{h} r_{j,\theta}\right\|\right]. \qquad (27)$$

For the term in (26), we can rewrite it as

$$2\mathbb{E}\left[\left\|\sum_{h=0}^{H-1} \nabla \log \hat{a}_h(s_h) \sum_{j=0}^{h} r_{j,\theta}\right\|\right] \sum_{h=0}^{\infty} \hat{p}_\theta(s_h, a_h) \geq 2\mathbb{E}\left[\left\|\sum_{h=0}^{H-1} \nabla \log \hat{a}_h(s_h) \sum_{j=0}^{h} r_{j,\theta}\right\|\right].$$

Since left-hand side of the above equation is non-decreasing and the limit as $N \to \infty$ exists, by the the Monotone Convergence Theorem, we can interchange the limit with the summation over the trajectory in (26) as follows:

$$2\mathbb{E}\left[\left\|\sum_{h=0}^{H-1} \nabla \log \hat{a}_h(s_h) \sum_{j=0}^{h} r_{j,\theta}\right\|\right] \leq 2\sum_{h=0}^{H-1} \sum_{j=0}^{h} \mathbb{E}\big[\|r_j(s_j, a_j) \nabla \log \hat{a}_j(s_j)\|\big].$$

Due to $\sum_a \hat{a}(s) \log \hat{a}(s) \geq -\log |\mathcal{A}|$, the term in (26) can be upper bounded as

$$2\mathbb{E}\left[\left\|\sum_{h=0}^{H-1} \nabla \log \hat{a}_h(s_h) \sum_{j=0}^{h} r_{j,\theta}\right\|\right] \leq 2r_{\max} \log |\mathcal{A}| \sum_{h=0}^{H-1} \sum_{j=0}^{h} 1 \leq \frac{2r_{\max} \log |\mathcal{A}| H^2}{1-\gamma}.$$

Similarly, we can interchange the limit with the summation over the trajectory in (27) and upper bound it as

$$2\mathbb{E}\left[\left\|\sum_{h=H}^{\infty} \nabla \log \hat{a}_h(s_h) \sum_{j=0}^{h} r_{j,\theta}\right\|\right] \leq 2r_{\max} \log |\mathcal{A}| \sum_{h=H}^{\infty} \sum_{j=0}^{h} \gamma^j \leq \frac{2r_{\max} \log |\mathcal{A}| H \gamma^H}{(1-\gamma)^2}.$$

Combining the above two inequalities, we have

$$\big[\mathbb{E}\big[\|\hat{V}^{\pi,H}(\theta) - V^{\pi}(\theta)\|_2\big]\big]^2 \leq \frac{2r_{\max} \log |\mathcal{A}| H \gamma^H}{(1-\gamma)^2} < H \left(\frac{1}{1-\gamma}\right)^2.$$

This completes the proof. ∎

D. Proof of Lemma 7

Proof. For the simplicity of the notation, we first denote:

$$g_1(\theta, \theta') = \sum_{h=0}^{H-1} \sum_{j=0}^{h} \nabla \log \hat{a}_j(s_j) \gamma^h r_h(s_h, a_h) \qquad (28)$$

$$g_2(\theta, \theta') = \sum_{h=0}^{H-1} \sum_{j=0}^{h} \nabla \log \hat{a}_j(s_j) \gamma^h \log \hat{a}_h(s_h). \qquad (29)$$

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

$$\mathrm{Var}\hat{\ }\copyright V\ ;H\ \hat{\ }\ \bullet\bullet$$
$$E\ \%g_1\hat{\ }\ {}^{H}S;\ \bullet\ g_2\hat{\ }\ {}^{H}S;\ \bullet\bullet \tag{30}$$
$$E\ g_1\hat{\ }\ {}^{H}S;\ \bullet\ E\ g_2\hat{\ }\ {}^{H}S;\ \bullet\ \check{Z}^2$$
$$B3E\hat{\ }g_1\hat{\ }\ {}^{H}S;\ \bullet\ E\ g_1\hat{\ }\ {}^{H}S;\ \bullet\bullet^2$$
$$3E\hat{\ }g_2\hat{\ }\ {}^{H}S;\ \bullet\bullet\ E\ g_2\hat{\ }\ {}^{H}S;\ \bullet\bullet\bullet^2 \tag{31}$$
$$3\%\mathrm{Var}\hat{\ }g_1\hat{\ }\ {}^{H}S;\ \bullet\bullet\ \mathrm{Var}\hat{\ }g_2\hat{\ }\ {}^{H}S;\ \bullet\bullet\check{Z}: \tag{32}$$

As shown in Lemma 4.2 of [7], the fact that $Y\copyright \log\ \hat{a}S\bullet Y_2\ B\ 2$ for all $\ >R^{SSSSS}$ directly implies that $\mathrm{Var}\hat{\ }g_1\hat{\ }\ {}^{H}S;\ \bullet\bullet\ B\ \frac{4r^2}{\hat{1}\ \bullet^4}$ for all $\ >R^{SSSSS}$.

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

$$Yg_2Y\ \left\|\sum_{h\ 0}^{H\ 1}\Big(\sum_{j\ 0}^{h}\copyright\log\ \hat{a}a_j^iSs_j^i\bullet\Big)\bullet\ {}^{h}\log\ \hat{a}a_h^iSs_h^i\bullet\check{Z}\right\|$$
$$B\ \sum_{h\ 0}^{H\ 1}\Big(\sum_{j\ 0}^{h}\copyright\log\ \hat{a}a_j^iSs_j^i\bullet Z\Big)\bullet\ {}^{h}\log\ \hat{a}a_h^iSs_h^i\bullet$$
$$B\ \ 2\sum_{h\ 0}^{H\ 1}\hat{h\ 1\bullet}\ {}^{h}\log\ \hat{a}a_h^iSs_h^i\bullet:$$

where the first inequality is due to the triangle inequality and the second inequality is due to $\copyright\log\ \hat{a}a_j^iSs_j^i\bullet Z\ B\ 2$. Then, by taking the suqre of Yg_2Y, we obtain

$$Yg_2Y^2\ B4\ {}^{2}\Big(\sum_{h\ 0}^{H\ 1}\hat{h\ 1\bullet}\ {}^{h}\log\ \hat{a}a_h^iSs_h^i\bullet\Big)^2$$
$$4\ {}^{2}\Big(\sum_{h\ 0}^{H\ 1}\hat{h\ 1\bullet}\ {}^{h}\log\ \hat{a}a_h^iSs_h^i\bullet\Big)^2$$
$$B4\ {}^{2}\Big(\sum_{h\ 0}^{H\ 1}\hat{h\ 1\bullet}^2\ {}^{h}\Big)\Big(\sum_{h\ 0}^{H\ 1}\ {}^{h}\%\log\ \hat{a}a_h^iSs_h^i\bullet\check{Z}^2\Big)$$
$$4\ {}^{2}\Big(\sum_{h\ 0}^{H\ 1}\hat{h^2\ \ 2h\ 1\bullet}\ {}^{h}\Big)\Big(\sum_{h\ 0}^{H\ 1}\ {}^{h}\%\log\ \hat{a}a_h^iSs_h^i\bullet\check{Z}^2\Big)$$
$$B4\ {}^{2}\Big(\frac{2}{\hat{1}\ \bullet^3}\ \ \frac{2}{\hat{1}\ \bullet^2}\ \ \frac{1}{1}\Big)$$
$$\Big(\sum_{h\ 0}^{H\ 1}\ {}^{h}\%\log\ \hat{a}a_h^iSs_h^i\bullet\check{Z}^2\Big)$$
$$4\ {}^{2}\Big(\frac{1}{\hat{1}\ \bullet^3}\Big)\Big(\sum_{h\ 0}^{H\ 1}\ {}^{h}\%\log\ \hat{a}a_h^iSs_h^i\bullet\check{Z}^2\Big)$$

where the second inequality is due to the Cauchy-Schwarz inequality and the last inequality is due to $\sum_{h\ 0}^{a}h^2\ {}^{h}\ \frac{2}{\hat{1}\ \bullet^3}$, $\sum_{h\ 0}^{a}h\ {}^{h}\ \frac{1}{\hat{1}\ \bullet^2}$ and $\sum_{h\ 0}^{a}\ {}^{h}\ \frac{1}{1}$.

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

$$E\ {}_{p\hat{\ }S\bullet}\ Yg_2Y^2$$
$$B4\ {}^{2}\Big(\frac{1}{\hat{1}\ \bullet^3}\Big)\sum_{h\ 0}^{H\ 1}\ {}^{h}E\ {}_{p\hat{\ }S\bullet}\ \%\log\ \hat{a}a_h^iSs_h^i\bullet\check{Z}^2 \tag{33}$$

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

$$E\ {}_{p\hat{\ }S\bullet}\ \%\log\ \hat{a}a_h^iSs_h^i\bullet\check{Z}^2$$
$$E_{s_0\ ;a_0\ \hat{\ }s_0\bullet;s_1\ p\hat{\ }s_0;s_0\bullet:::a_{H\ 1}\ \hat{\ }s_{H\ 1}\bullet;s_H\ p\hat{\ }s_{H\ 1};a_{H\ 1}\bullet}$$
$$\%\log\ \hat{a}a_h^iSs_h^i\bullet\check{Z}^2$$
$$E_{s_0\ ;a_0\ \hat{\ }s_0\bullet;s_1\ p\hat{\ }s_0;s_0\bullet:::a_h\ \hat{\ }s_h\bullet}\ \%\log\ \hat{a}a_h^iSs_h^i\bullet\check{Z}^2$$
$$E_{s_0\ ;a_0\ \hat{\ }s_0\bullet;s_1\ p\hat{\ }s_0;s_0\bullet:::s_h\ p\hat{\ }s_{h\ 1};s_{h\ 1}\bullet}$$
$$\left\langle\frac{\partial}{\partial Q\ \hat{\ }a_hSs_h\bullet}\%\log\ \hat{a}a_h^iSs_h^i\bullet\check{Z}^2\right\rangle_{a_h>A}.$$

Since the maximizer for the constrained problem

$$\max\ \sum_{i\ 1}^{n}x_i\hat{\log x_i\bullet}^2\quad\text{such that}\sum_{i\ 1}^{n}x_i\ \ 1;$$

is $x_1\ \ x_2\ \ \ldots\ \ x_n\ \ \frac{1}{n}$ and the maximum solution is $\hat{\log n\bullet}^2$. Thus, we have $P_{a_h>A}\ \hat{a}a_hSs_h\bullet\%\log\ \hat{a}a_h^iSs_h^i\bullet\check{Z}^2\ B\ \hat{\log SAS\bullet}^2$ and

$$E\ {}_{p\hat{\ }S\bullet}\ \%\log\ \hat{a}a_h^iSs_h^i\bullet\check{Z}^2$$
$$BE_{s_0\ ;a_0\ \hat{\ }s_0\bullet;s_1\ p\hat{\ }s_0;s_0\bullet:::s_h\ p\hat{\ }s_{h\ 1};s_{h\ 1}\bullet}\ \hat{\log SAS\bullet}^2$$
$$\hat{\log SAS\bullet}^2: \tag{34}$$

By combining (33) and (34), we have

$$\mathrm{Var}\hat{\ }g_2\bullet\ BE\ {}_{p\hat{\ }S\bullet}\ Yg_2Y^2$$
$$B4\ {}^{2}\Big(\frac{1}{\hat{1}\ \bullet^3}\Big)\sum_{h\ 0}^{H\ 1}\ {}^{h}E\ {}_{p\hat{\ }S\bullet}\ \%\log\ \hat{a}a_h^iSs_h^i\bullet\check{Z}^2$$
$$B4\ {}^{2}\Big(\frac{1}{\hat{1}\ \bullet^3}\Big)\sum_{h\ 0}^{H\ 1}\ {}^{h}\hat{\log SAS\bullet}^2$$
$$B\frac{8\ {}^{2}\hat{\log SAS\bullet}^2}{\hat{1}\ \bullet^4}:$$

Finally, by substituting $\mathrm{Var}\hat{\ }g_1\bullet;\mathrm{Var}\hat{\ }g_2\bullet$ and $\mathrm{Var}\hat{\ }g_3\bullet$ into (30), it holds that

$$\mathrm{Var}\hat{\ }\copyright V\ ;H\ \hat{\ }\ \bullet\bullet\ B\ \frac{12r^2\ \ 24\ {}^{2}\hat{\log SAS\bullet}^2}{\hat{1}\ \bullet^4}:$$

This completes the proof. j

APPENDIX B
HELPFUL RESULTS FOR THE PROOF OF LEMMA 10

Proposition 1 (Proposition 3 in [48]): Consider the problem $\max_{x>R^d}\ f\ \hat{x}\bullet$, where R^d denotes the d-dimensional Euclidean space. Let $\hat{x_t\bullet}_{t\ 0}^{a}$ be a sequence generated by the iterative method $x_{t\ 1}\ \ x_t\ \ {}_t\hat{u_t\ \ w_t\bullet}$, where $\ _t$ 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 fields. Assume that

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

$$Y\copyright f\ \hat{x}\bullet\ \copyright\ f\ \hat{x}\bullet Y\ B\ L\ Yx\ \ xY;\quad \forall\ x;x>R^d:$$

2) x_t and u_t are F_t measurable.