Local Analysis of Entropy-Regularized Stochastic Soft-Max Policy Gradient Methods

Yuhao Ding[1], Junzi Zhang[2] and Javad Lavaei[1]

***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 by the assumption of exact gradient oracles. To go beyond this limitation, we study the convergence of stochastic soft-max vanilla policy gradient with entropy regularization and prove how to utilize the curvature information around the optimal policy to guarantee that the action probabilities will still remain uniformly bounded with high probability. Moreover, we develop the "last iterate" convergence and sample complexity result for the proposed algorithm given a good initialization.**

I. Introduction

Entropy regularization is a popular technique to encourage exploration and prevent premature convergence for reinforcement learning (RL) algorithms. The idea 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 also been demonstrated to work well with deep learning approximations to achieve an impressive empirical performance boost. Nevertheless, the theoretical understanding of the convergence of these algorithms has been rather limited and mostly restricted to the exact gradient setting.

The theoretical understanding of policy-based methods has received considerable attention recently [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]. Several techniques have been developed to improve standard PG and achieve a linear convergence rate, such as adding entropy regularization [10], [7], [8], [11], exploiting natural geometries based on Bregman divergences leading to NPG or policy mirror descent [9], [8], [11], and using a geometry-aware normalized PG (GNPG) approach to exploit the non-uniformity of the value function [20]. 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. 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 [11], [8], as well as incorporating variance reduction techniques to improve the sample complexity of the vanilla PG [21], [22]. In particular, 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. For NPG without any regularization, it has been shown that the noise introduced by stochastic gradients will incur a positive probability of failure as empirically observed in [23] and later proved in [24]. However, the literature on the 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. It remains open whether a local 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]. In particular, we begin by proposing an unbiased estimator for the new entropy regularized stochastic PG. It is the first likelihood-ratio-based estimators in the literature with a trajectory-level entropy regularization. We show that although the estimator itself is unbounded in general due to the entropy-induced logarithmic policy rewards, the variances indeed remain uniformly bounded. We then establish that with a good initial policy, stochastic entropy-regularized vanilla PG method enjoys a sample complexity of $\widetilde{\mathcal{O}}(\frac{1}{\epsilon^2})$ for the local optimality convergence under the softmax parameterization. We also stress here that this is a "last iterate" convergence guarantee; neither ergodic, nor of a mean-squared gradient norm type. This is crucial for real-world applications because, in practice, stochastic PG training is based on the last generated point.

A. 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 ξ and over all of the randomness. The notion

[1]Yuhao Ding and Javad Lavaei are with the Department of Industrail Engineering and Operations Research, University of California at Berkeley. Email: {yuhao_ding,lavaei}@berkeley.edu

[2]Junzi Zhang is with Citadel Securities (work done prior to joining Citadel Securities), Chicago, IL 60603 USA (e-mail: saslascroyale@gmail.com).

Var[$\cdot$] refers to the variance. $\Delta(\mathcal{X})$ 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 $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma)$. Here, $\mathcal{S}$ and $\mathcal{A}$ denote the finite state and action spaces; $\mathbb{P}(s'|s,a)$ is the probability that the agent transits from the state s to the state s' under the action $a \in \mathcal{A}$; $r(s,a)$ is the reward function, i.e., the agent obtains the reward $r(s_h, a_h)$ after it takes the action a_h at the state s_h at time h; $\gamma \in (0,1)$ is the discount factor. Without loss of generality, we assume that $r(s,a) \in [0, \bar{r}]$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. The policy $\pi(a|s)$ at the state s is usually represented by a conditional probability distribution $\pi_\theta(a|s)$ associated to the parameter $\theta \in \mathbb{R}^d$. Let $\tau = \{s_0, a_0, s_1, a_1, \ldots\}$ denote the data of a sampled trajectory under policy π_θ with the probability distribution over the trajectory as $p(\tau|\theta, \rho) := \rho(s_0) \prod_{h=1}^{\infty} \mathbb{P}(s_{h+1}|s_h, a_h)\pi_\theta(a_h|s_h)$, where $\rho \in \Delta(\mathcal{S})$ is the probability distribution of the initial state s_0.

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

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

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

Exploratory initial distribution. The discounted state visitation distribution $d^\pi_{s_0}$ is defined as $d^\pi_{s_0}(s) := (1-\gamma)\sum_{h=0}^{\infty} \gamma^h \mathbb{P}(s_h = s|s_0, \pi)$, where $\mathbb{P}(s_h = s|s_0, \pi)$ is the state visitation probability that s_h is equal to s under the policy π starting from the state s_0. The discounted state visitation distribution under the initial distribution ρ is defined as $d^\pi_\rho(s) := \mathbb{E}_{s_0 \sim \rho}[d^\pi_{s_0}(s)]$. Furthermore, the state-action visitation distribution induced by π and the initial state distribution ρ is defined as $v^\pi_\rho(s,a) := d^\pi_\rho(s)\pi(a|s)$, which can also be written as $v^\pi_\rho(s,a) := (1-\gamma)\mathbb{E}_{s_0 \sim \rho} \sum_{h=0}^{\infty} \gamma^h \mathbb{P}(s_h = s, a_h = a|s_0, \pi)$, where $\mathbb{P}(s_h = s, a_h = a|s_0, \pi)$ is the state-action visitation probability that $s_h = s$ and $a_h = a$ under π starting from the state s_0. To facilitate the presentation of the main results of the paper, we assume that the state distribution ρ for the performance measure is exploratory [10], [13], i.e., $\rho(\cdot)$ adequately covers the entire state distribution:

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

In practice, when the above assumption is not satisfied, we can optimize under another initial distribution μ, i.e., the gradient is taken with respect to the optimization measure μ, where μ is usually chosen as an exploratory initial distribution that adequately covers the state distribution of some optimal policy. It is shown in [7] that the difficulty of the exploration problem faced by PG algorithms can be captured through the distribution mismatch coefficient defined as $\left\| \frac{d^{\pi^*}_\rho}{\mu} \right\|_\infty$, where $\frac{d^{\pi^*}_\rho}{\mu}$ denotes component-wise division.

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

RL with entropy regularization. Entropy is a commonly used regularization in RL to promote exploration and discourage premature convergence to suboptimal policies [5], [25], [26]. 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

$$V^\pi_\lambda(\rho) = V^\pi(\rho) + \lambda \mathbb{H}(\rho, \pi), \tag{1}$$

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

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

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

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

Bias due to entropy regularization. Due to the presence of regularization, the optimal solution will be biased with the bias disappearing as $\lambda \to 0$. More precisely, the optimal policy π^*_λ of the entropy-regularized problem could also be nearly optimal in terms of the unregularized objective function, as long as the regularization parameter λ is chosen to be small. Denote by π^* and π^*_λ the policies that maximize the objective

function and the entropy-regularized objective function with the regularization parameter λ, respectively. Let V^* and V_λ^* represent the resulting optimal objective value function and the optimal regularized objective value function. [11] shows a simple but crucial connection between π^* and π_λ^* via the following sandwich bound:

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

which holds for all initial distribution ρ.

III. Stochastic PG estimators

The PG method is one of the most popular approaches for a direct policy search in RL [27]. 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 ([10]): $V_\lambda^\theta(s) \le \frac{\bar{r}+\lambda\log|\mathcal{A}|}{1-\gamma}$ and $Q_\lambda^\pi(s,a) \le \frac{\bar{r}+\lambda\log|\mathcal{A}|}{1-\gamma}$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

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

Lemma 2 (Proposition 2 in [28]): The entropy regularized PG with respect to θ is

$$\nabla V_\lambda^\theta(\rho) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim v_\rho^{\pi_\theta}}\left[\nabla_\theta \log\pi_\theta(a|s)\left(Q_\lambda^\theta(s,a) - \lambda\log\pi_\theta(a\mid s)\right)\right], \tag{2}$$

where

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

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

In order to obtain an unbiased sample of $\nabla V_\lambda^\theta(\rho)$, we need to first draw a state-action pair (s,a) from the distribution $\nu_\rho^{\pi_\theta}(\cdot,\cdot)$ and then obtain an unbiased estimate of the action-value function $Q_\lambda^\theta(s,a)$. 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 $r(s,a) - \lambda\log\pi(a|s)$ 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 $Q_\lambda^\theta(\cdot,\cdot)$ is evaluated, where the horizon $H \sim \text{Geom}(1-\gamma)$. It can be shown that $(s_H, a_H) \sim \nu_\rho^{\pi_\theta}(s,a)$. Moreover, given (s_H, a_H), we will perform Monte-Carlo rollouts for another trajectory with the horizon $H' \sim \text{Geom}\left(1-\gamma^{1/2}\right)$ independent of H, and estimate the advantage function value $Q_\lambda^\theta(s,a)$ along the trajectory $(s'_0, a'_0, \ldots, s'_{H'})$ with $s'_0 = s, a'_0 = a$ as follows:

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

The subroutines of sampling one pair (s,a) from $\nu_\rho^{\pi_\theta}(\cdot,\cdot)$, estimating $\hat{Q}_\lambda^\theta(s,a)$, and estimating $\hat{V}_\lambda^\theta(s)$ are summarized as **Sam-SA** and **Est-EntQ** in Algorithms 1 and 2, respectively.

Algorithm 1 Sam-SA: Sample for $s,a \sim \nu_\rho^{\pi_\theta}(\cdot,\cdot)$

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

Algorithm 2 Est-EntQ: Unbiasedly estimating entropy-regularized Q function

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

We then propose the following stochastic estimator:

$$\hat{\nabla} V_\lambda^\theta(\rho) = \frac{1}{1-\gamma}\nabla_\theta\log\pi_\theta(a_H|s_H)\left(\hat{Q}_\lambda^\theta(s_H,a_H) - \lambda\log\pi_\theta(a_H\mid s_H)\right), \tag{4}$$

where $s_H, a_H \leftarrow$ **Sam-SA**(ρ,θ,γ) and $\hat{Q}_\lambda^\theta$ is defined in (3). The following lemma shows that the stochastic PG (4) is an unbiased estimator of $\nabla V_\lambda^\theta(\rho)$.

Lemma 3: For $\hat{\nabla} V_\lambda^\theta(\rho)$ defined in (4), we have $\mathbb{E}[\hat{\nabla} V_\lambda^\theta(\rho)] = \nabla V_\lambda^\theta(\rho)$.

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

Lemma 4: For $\hat{\nabla} V_\lambda^\theta(\rho)$ defined in (4), we have $\text{Var}[\hat{\nabla} V_\lambda^\theta(\rho)] \le \sigma^2$, where $\sigma^2 = \frac{8}{(1-\gamma)^2}\left(\frac{\bar{r}^2 + (\lambda\log|\mathcal{A}|)^2}{(1-\gamma^{1/2})^2}\right)$.

In practice, we can sample and compute a batch of independently and identically distributed PG estimators $\{\hat{\nabla} V_\lambda^{\theta,i}(\rho)\}_{i=1}^B$ where B is the batch size, in order to

reduce the estimation variance. To maximize the entropy-regularized objective function (1), we can then update the policy parameter θ by iteratively running gradient-ascent-based algorithms, i.e., $\theta_{t+1} = \theta_t + \frac{\eta_t}{B}\sum_{i=1}^{B}\hat{\nabla}V_\lambda^{\theta,i}(\rho)$, where $\eta_t > 0$ is the step size. The details of the unbiased PG algorithm with a random horizon for the entropy-regularized RL are provided in Algorithm 3.

Algorithm 3 Ent-RPG: Random-horizon PG for Entropy-regularized RL

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

IV. Review: Linear convergence with exact PG

A key result from [10] shows that, under the soft-max parameterization, the entropy-regularized value function $V_\lambda^\theta(\rho)$ in (1) satisfies a non-uniform Łojasiewicz inequality as follows:

Lemma 5 (Lemma 15 in [10]): It holds that

$$\left\|\nabla V_\lambda^\theta(\rho)\right\|_2^2 \geq C(\theta)(V_\lambda^{\theta^*}(\rho) - V_\lambda^\theta(\rho)),$$

where

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

Furthermore, it is shown in [10] that the action probabilities under the soft-max parameterization are uniformly bounded away from zero if the exact PG is available.

Lemma 6 (Lemma 16 in [10]): Using the exact PG with the learning rate $\eta_t = \eta \leq \frac{2}{L}$ for the entropy regularized objective, it holds that $\inf_{t\geq 1}\min_{s,a}\pi_{\theta_t}(a|s) > 0$.

Remark 1: Note that with the exact PG , $\inf_{t\geq 1}\min_{s,a}\pi_{\theta_t}(a|s)$ is only dependent on the initialization θ_1 and step-size η (apart from problem dependent constants). Hence hereafter we denote $c_{\theta_1,\eta} = \inf_{t\geq 1}\min_{s,a}\pi_{\theta_t}(a|s)$.

With Lemmas 5 and 6, it is shown in Theorem 6 of [10] that the convergence rate for the entropy regularized PG is $O\left(e^{-Ct}\right)$, where the value of C depends on $\inf_{t\geq 1}\min_{s,a}\pi_{\theta_t}(a|s) > 0$ and $\{\theta_t\}_{t=1}^\infty$ is generated by the exact PG method. With a bad initialization θ_1, $\min_{s,a}\pi_{\theta_1}(a|s)$ could be very small and result in a slow convergence rate. When studying the stochastic PG, this issue of bad initialization will create more severe challenges on the convergence and finding a good initial policy becomes critical. However, even with a good initial policy, it remains unknown whether the stochastic policy gradient with the entropy regularization will guarantee the convergence and how to characterize the region for the good initial policy.

V. 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. We denote $D(\theta_t) = V_\lambda^{\theta^*}(\rho) - V_\lambda^{\theta_t}(\rho)$ as the sub-optimality gap between $V_\lambda^{\theta^*}(\rho)$ and $V_\lambda^{\theta_t}(\rho)$.

Lemma 7: Given a tolerance level $\delta > 0$, let π_λ^* be the optimal policy of $V_\lambda^\theta(\rho)$. Assume further that Algorithm 3 is run for T iterates with a step-size sequence of the form $\eta_t = 1/(t+t_0)$ and a batch-size sequence $B \geq \frac{1}{\eta_t}$ for all $t = 1, 2, \ldots, T$. If $t_0 \geq \sqrt{\frac{3\sigma^2}{2\delta\epsilon_0}}$, and π_{θ_1} is initialized in a neighborhood $\mathcal{U}_1$ such that

$$\mathcal{U}_1 = \left\{\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|} : D(\pi) \leq \epsilon_0\right\}, \tag{5}$$

where $\epsilon_0 = \min\left\{\left(\frac{\lambda \min_s \rho(s)}{6\ln 2}\right)^2\left(\alpha\exp(\frac{-\bar{r}}{(1-\gamma)\lambda})\right)^4, 1\right\}$ and the constant $\alpha \in (0,1)$, then the event

$$\Omega_{\alpha,1}^T = \left\{\min_{s,a}\pi_{\theta_t}(a|s) \geq (1-\alpha)\min_{s,a}\pi_\lambda^*(a|s), \forall t = 1, 2, \ldots, T\right\} \tag{6}$$

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

A. Helpful lemmas

Since the optimal policy of (1) is unique [11], there must exist a continuum of optimal solutions

$$\Theta^* \coloneqq \{\theta^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} : \frac{\exp(\theta_{s,a}^*)}{\sum_{a'}\exp(\theta_{s,a'}^*)} = \pi_\lambda^*(a \mid s), \forall s \in \mathcal{S}, a \in \mathcal{A}\}.$$

In addition, we use π_{θ^*} and π_λ^* interchangeably to denote the optimal policy of the entropy-regularized RL.

Lemma 8: Suppose that $f(x)$ is L-smooth. Given $0 < \eta_t \leq \frac{1}{2L}$ for all $t \geq 1$, let $\{x_t\}_{t=1}^T$ be generated by a general update of the form $x_{t+1} = x_t + \eta_t u_t$ and let $e_t = u_t - \nabla f(x_t)$. We have

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

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

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

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

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

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

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

This completes the proof. □

To prove Lemma 7, we first characterize the maximum amount by which $D(\theta_t)$ can grow at each step.

Lemma 9: Suppose that $\{\theta_t\}$ is generated by Algorithm 3 with $0 < \eta_t \le \frac{(1-\gamma)^3}{16\bar{r}+\lambda(8+16\log|\mathcal{A}|)}$ for all $t \ge 1$. We have

$$D(\theta_{t+1}) \le \left(1 - \frac{\eta_t C(\theta_t)}{4}\right) D(\theta_t) - \frac{\eta_t}{2}\xi_t + \frac{\eta_t}{4} \|e_t\|_2^2 , \quad (7)$$

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

Proof. Since $\nabla V_\lambda^\theta(\rho)$ is L-smooth in light of Lemmas 7 and 14 in [10], it follows from Lemma 8 that

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

for every $\eta_t \le \frac{1}{2L}$, where the last inequality is due to Lemma 5. □

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. The difficulty of controlling the errors in $D(\theta_t)$ lies in the fact that $C(\theta_t)$ may be close to 0. 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 $D(\theta^*)$. Similar as the techniques used in [30], [31], [20], [32], [33], we now encode the error terms in (7) as $M_n = \sum_{t=1}^n \eta_t \xi_t$ and $S_n = \sum_{t=1}^n \frac{\eta_t}{4} \|e_t\|_2^2$. Since $\mathbb{E}[\xi_n] = 0$, we have $\mathbb{E}[M_n] = M_{n-1}$. Therefore, M_n is a zero-mean martingale; likewise, $\mathbb{E}[S_n] \ge S_{n-1}$, and therefore S_n is a submartingale. We begin by introducing the "cumulative mean square error" $R_n = M_n^2 + S_n$. By construction, we have

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

Hence, $\mathbb{E}[R_n] = R_{n-1} + 2M_{n-1}\eta_n\mathbb{E}[\xi_n] + \eta_n^2\mathbb{E}[\xi_n^2] + \frac{1}{4}\eta_n\mathbb{E}\left[\|e_n\|^2\right] \ge R_{n-1}$, i.e., R_n is a submartingale. With a fair degree of hindsight, we define $\mathcal{U}$ as:

$$\mathcal{U} = \left\{\pi \in \Delta(\mathcal{A})^{|\mathcal{S}|} : D(\pi) \le 2\epsilon_0 + \sqrt{\epsilon_0}\right\} . \quad (8)$$

To condition it further, we also define the events

$$\begin{aligned} \Omega_n &\equiv \Omega_n(\epsilon_0) = \{\pi_{\theta_t} \in \mathcal{U} \text{ for all } t = 1, 2, \ldots, n\} \\ E_n &\equiv E_n(\epsilon_0) = \{R_t \le \epsilon_0 \text{ for all } t = 1, 2, \ldots, n\} \end{aligned}$$

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

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

Lemma 10: Let π_λ^* be the optimal policy. Then, for all $n \in \{1, 2, \ldots\}$, the following statements hold:

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

$$\begin{aligned} \tilde{E}_n &\equiv E_{n-1} \backslash E_n = E_{n-1} \cap \{R_n > \epsilon_0\} \\ &= \{R_t \le \epsilon_0 \text{ for all } t = 1, 2, \ldots, n-1 \text{ and } R_n > \epsilon_0\} \end{aligned}$$

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

$$\mathbb{E}\left[\tilde{R}_n\right] \le \mathbb{E}\left[\tilde{R}_{n-1}\right] + G^2\sigma^2\eta_n^2 + \frac{\eta_n\sigma^2}{4B} - \epsilon_0\mathbb{P}\left(\tilde{E}_{n-1}\right) . \quad (9)$$

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

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

- For the base case $n = 1$, we have $\Omega_1 = \{\pi_{\theta_1} \in \mathcal{U}\} \supseteq \{\pi_{\theta_1} \in \mathcal{U}_1\} = \Omega$ because π_{θ_1} is initialized in $\mathcal{U}_1 \subseteq \mathcal{U}$. Since $E_0 = \Omega$, our claim follows.
- For the inductive step, assume that $E_{n-1} \subseteq \Omega_n$ for some $n \ge 1$. To show that $E_n \subseteq \Omega_{n+1}$, we fix a realization in E_n such that $R_t \le \varepsilon$ for all $t = 1, 2, \ldots, n$. Since $E_n \subseteq E_{n-1}$, the inductive hypothesis posits that Ω_n also occurs, i.e., $\pi_{\theta_t} \in \mathcal{U}$ for all $t = 1, 2, \ldots, n$; hence, it suffices to show that $\pi_{\theta_{n+1}} \in \mathcal{U}$. To that end, given that $\pi_{\theta_t} \in \mathcal{U}$ for all $t = 1, 2, \ldots n$, the distance estimate (7) readily gives $D(\theta_{t+1}) \le D(\theta_t) + \eta_t\xi_t + \frac{\eta_t}{4}\|e_t\|_2^2$ for all $t = 1, 2, \ldots, n$. Therefore, after telescoping, we obtain

$$\begin{aligned} D(\theta_{n+1}) \le & D(\theta_1) + M_n + S_n \le D(\theta_1) + \sqrt{R_n} + R_n \\ \le & \varepsilon + \sqrt{\varepsilon} + \varepsilon \\ = & 2\varepsilon + \sqrt{\varepsilon} \end{aligned}$$

by the inductive hypothesis. This completes the induction.

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

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

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

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

and therefore

$$\mathbb{E}\left[(R_n - R_{n-1})\,\mathbb{1}_{E_{n-1}}\right] = \tag{10}$$

$$2\eta_n\mathbb{E}\left[M_{n-1}\xi_n\mathbb{1}_{E_{n-1}}\right] + \eta_n^2\mathbb{E}\left[\xi_n^2\mathbb{1}_{E_{n-1}}\right] + \frac{1}{4}\eta_n\mathbb{E}\left[\|e_n\|^2\mathbb{1}_{E_{n-1}}\right].$$

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

- For the term in (10), by the unbiasedness of the gradient estimator shown in Lemma 3, we have: $\mathbb{E}\left[M_{n-1}\xi_n\mathbb{1}_{E_{n-1}}\right] = \mathbb{E}\left[M_{n-1}\mathbb{1}_{E_{n-1}}\mathbb{E}\left[\xi_n \mid \mathcal{F}_n\right]\right] = 0$.
- The second term in (10) is where the conditioning on E_{n-1} plays the most important role. It holds that:

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

  where the first inequality is due to the Cauchy-Schwarz inequality, the second inequality follows from $E_{n-1} \subseteq \Omega_n$ and the last inequality results from Lemmas 2 and 4.
- Finally, for the third term in (10), we have:

$$\frac{\eta_n}{4}\mathbb{E}\left[\|e_n\|_2^2\mathbb{1}_{E_{n-1}}\right] \le \frac{\eta_n\sigma^2}{4B}. \tag{11}$$

Thus, putting together all of the above, we obtain $\mathbb{E}\left[(R_n - R_{n-1})\,\mathbb{1}_{E_{n-1}}\right] \le G^2\sigma^2\eta_n^2 + \frac{\eta_n\sigma^2}{4B}$. Since $R_{n-1} > \varepsilon$ if $\tilde{E}_{n-1}$ occurs, we obtain $\mathbb{E}\left[R_{n-1}\mathbb{1}_{\tilde{E}_{n-1}}\right] \ge \varepsilon\mathbb{E}\left[\mathbb{1}_{\tilde{E}_{n-1}}\right] = \varepsilon\mathbb{P}\left(\tilde{E}_{n-1}\right)$. This completes the proof of Statement 3. □

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

Lemma 11: Consider an arbitrary tolerance level $\delta > 0$. If Algorithm 3 is run with a step-size schedule of the form $\eta_t = 1/(t + t_0)$ where $t_0 \ge \sqrt{\frac{3\sigma^2}{2\delta\epsilon_0}}$ and a batch size schedule $B_t \ge \frac{1}{\eta_t}$, we have $\mathbb{P}(E_n) \ge 1 - \delta/6,$ for all $n = 1, 2, \ldots$

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

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

which is derived by using the fact that $R_n \ge 0$ (so $\mathbb{1}_{\{R_n > \varepsilon\}} \le R_n/\varepsilon$). Now, by summing up (9), we conclude that $\mathbb{E}\left[\tilde{R}_n\right] \le \mathbb{E}\left[\tilde{R}_0\right] + \frac{\sigma^2}{4B}\sum_{t=1}^n \eta_t - \varepsilon\sum_{t=1}^n \mathbb{P}\left(\tilde{E}_{t-1}\right)$. Hence, combining the above results, we obtain the estimate

$$\sum_{t=1}^n \mathbb{P}\left(\tilde{E}_k\right) \le \frac{\sigma^2}{4B\epsilon_0}\sum_{t=1}^n \eta_t \le \frac{\sigma^2}{4\epsilon_0}\sum_{t=1}^n \eta_t^2 \le \frac{\sigma^2\Gamma}{4\epsilon_0},$$

where $\Gamma = \sum_{t=1}^\infty \eta_t^2 = \sum_{t=1}^\infty (t + t_0)^{-2}$, and we have used the relations that $\tilde{R}_0 = 0$ and $\tilde{E}_0 = \varnothing$ (by convention). By choosing $t_0 \ge \sqrt{\frac{3\sigma^2}{2\delta\epsilon_0}}$, we ensure that $\frac{\sigma^2\Gamma}{4\epsilon_0} < \delta/6$; moreover, since the events $\tilde{E}_t$ are disjoint for all $t = 1, 2, \ldots$, we obtain $\mathbb{P}\left(\bigcup_{t=1}^n \tilde{E}_t\right) = \sum_{t=1}^n \mathbb{P}\left(\tilde{E}_t\right) \le \delta/6$. Hence, $\mathbb{P}(E_n) = \mathbb{P}\left(\bigcap_{t=1}^n \tilde{E}_t^{\mathrm{c}}\right) \ge 1 - \delta/6$ as claimed. □

Furthermore, we can show that the entropy-regularized value function $V_\lambda^\theta(\rho)$ is locally quadratic around the optimal policy π_{θ^*}.

Lemma 12: For every policy π_θ, we have

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

Proof. It follows from the soft sub-optimality difference lemma (Lemma 26 in [34]) that

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

where the first inequality is due to Theorem 11.6 in [35] stating that

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

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

B. Proof of Lemma 7

Since the sequence Ω_n is decreasing and $\Omega_n \supseteq E_{n-1}$ (by the second part of Lemma 10), Lemma 11 yields that $\mathbb{P}(\Omega_T) \ge \inf_n \mathbb{P}(\Omega_n) \ge \inf_n \mathbb{P}(E_{n-1}) \ge 1 - \delta/6$ provided that t_0 is chosen large enough.

Now, it remains to show that $\Omega_T \subseteq \Omega_{\alpha,1}^T$. We fix a realization in Ω_T such that $D(\theta_t) \le 2\epsilon_0 + \sqrt{\epsilon_0}$ for all

$t = 1, 2, \ldots, T$. By Lemma 12, we have

$$\begin{aligned}
&|\pi_{\theta_t}(a \mid s) - \pi_{\theta^*}(a \mid s)| \\
\leq &\sqrt{\frac{2D(\theta_t)\ln 2}{\lambda \min_s \rho(s)}} \leq \sqrt{\frac{2(2\epsilon_0 + \sqrt{\epsilon_0})\ln 2}{\lambda \min_s \rho(s)}} \\
\leq &\sqrt{\frac{6\sqrt{\epsilon_0}\ln 2}{\lambda \min_s \rho(s)}} \leq \alpha \exp(\frac{-\bar{r}}{(1-\gamma)\lambda}) \leq \alpha \min_{s,a} \pi_{\theta^*}(a \mid s),
\end{aligned}$$

where the second inequality is due to the condition that the event Ω_T occurs, the third inequality is due to $\epsilon_0 \leq \sqrt{\epsilon_0}$ when $\epsilon_0 \leq 1$, the forth inequality is due to the definition of ϵ_0, and the last inequality is due to Theorem 1 in [36] where it holds that $\log \pi^*_\lambda(a \mid s) = \frac{1}{\lambda}\left(Q^{\pi^*_\lambda}(s,a) - V^{\pi^*_\lambda}(s)\right) \geq \frac{-\bar{r}}{(1-\gamma)\lambda}$, $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$.

Now, it can be easily verified that $\pi_{\theta_t}(a \mid s) \geq \pi_{\theta^*}(a \mid s) - \alpha \min_{s,a} \pi_{\theta^*}(a \mid s)$. For every $t \in \{1, 2, \ldots, T\}$, let $\bar{s}, \bar{a} = \operatorname{argmin}_{s,a} \pi_{\theta_t}(a \mid s)$. One can write

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

where the last inequality is due to $\pi(a|s) \geq \min_{s,a} \pi(a|s)$ for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Thus, we obtain $\mathbb{P}\left(\Omega^T_{\alpha,1}\right) \geq \mathbb{P}\left(\Omega_T\right) \geq 1 - \delta/6$. This completes the proof.

VI. Last iterate convergence

From Lemma 7, we know that, with a good initialization, the policies will remain in the interior of the probability simplex with high probability. We are now ready to prove the "last iterate" convergence and the sample complexity of the stochastic PG for entropy-regularized RL when a good initilization is given.

Theorem 1: Consider an arbitrary tolerance level $\delta > 0$ and a small enough tolerance level $\epsilon > 0$. For initial point θ_1 satisfying the condition (5), if θ_{T+1} is generated by Algorithm 3 with

$$T \geq \frac{t_0 \epsilon_0}{6\delta\epsilon} - t_0, \;\; B \geq \frac{\sigma^2 \ln(T + t_0)}{6 C_\alpha \delta \epsilon}, \;\; \eta_t = \frac{1}{t + t_0},$$

where

$$\epsilon_0 = \min\left\{\left(\frac{\lambda \min_s \rho(s)}{6\ln 2}\right)^2 \left(\alpha \exp(\frac{-\bar{r}}{(1-\gamma)\lambda})\right)^4, 1\right\}, \quad (12)$$

$$t_0 \geq \sqrt{\frac{3\sigma^2}{2\delta\epsilon_0}}, \quad (13)$$

$$C_\alpha = \frac{2\lambda}{|\mathcal{S}|} \min_s \rho(s) (1-\alpha)^2 \min_{s,a} \pi_{\theta^*}(a|s)^2 \left\| \frac{d_\rho^{\pi^*_\lambda}}{\rho} \right\|_\infty^{-1} > 0, \quad (14)$$

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

Proof. When $D(\theta_1) \leq \epsilon_0$, it follows from Lemma 9 that

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

for all $t \geq T_1$, where $\xi_t = \langle e_t, \nabla V^{\theta_t}_\lambda(\rho)\rangle$. When $D(\theta_1) \leq \epsilon_0$, from Lemma 7 we know that the event $\Omega^t_{\alpha,1}$ defined in (6) occurs and $C(\theta_t) \geq C_\alpha$, with high probability, where C_α is defined in (14). By taking the expectation, we have

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

where the first equality is because $\Omega^t_{\alpha,1}$ is deterministic conditioning on $\mathcal{F}_t$, the second equality is due to the unbiasedness of ξ_t conditioning on $\mathcal{F}_t$, and the first inequality is due to (11). Therefore, $\mathbb{E}[D(\theta_{t+1})\mathbb{1}_{\Omega^t_{\alpha,1}}] \leq \left(1 - \frac{\eta_t C_\alpha}{4}\right)\mathbb{E}\left[D(\theta_t)\mathbb{1}_{\Omega^t_{\alpha,1}}\right] + \frac{\eta_t \sigma^2}{4B}$. Arguing inductively yields that

$$\begin{aligned}
&\mathbb{E}[D(\theta_{T+1})\mathbb{1}_{\Omega^T_{\alpha,1}}] \\
\leq &\prod_{i=1}^T \left(1 - \frac{\eta_i C_\alpha}{4}\right) D(\theta_1) + \sum_{i=1}^T \left(1 - \frac{\eta_i C_\alpha}{4}\right)^i \frac{\eta_i \sigma^2}{4B} \\
\leq &\prod_{i=1}^T \left(1 - \frac{\eta_i C_\alpha}{4}\right) D(\theta_1) + \sum_{i=1}^T \frac{\eta_i \sigma^2}{4B}.
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[D(\theta_{T+1})\mathbb{1}_{\Omega^T_{\alpha,1}}] \leq &\prod_{i=1}^T \left(\frac{i + t_0 - 1}{i + t_0}\right) D(\theta_T) + \frac{\sigma^2}{C_\alpha B} \sum_{i=1}^T \frac{1}{i + t_0} \\
\leq &\frac{t_0}{T + t_0} D(\theta_1) + \frac{\sigma^2 \ln(T + t_0)}{B C_\alpha}.
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}(D(\theta_{T+1}) \geq \epsilon) \\
=&\mathbb{P}(D(\theta_{T+1}) \geq \epsilon, \Omega^T_{\alpha,1}) + \mathbb{P}(D(\theta_{T+1}) \geq \epsilon, \left(\Omega^T_{\alpha,1}\right)^c) \\
=&\mathbb{P}(D(\theta_{T+1}) \geq \epsilon \mid \Omega^T_{\alpha,1}) \mathbb{P}(\Omega^T_{\alpha,1}) \\
&+ \mathbb{P}(D(\theta_{T+1}) \geq \epsilon \mid \left(\Omega^T_{\alpha,1}\right)^c) \mathbb{P}(\left(\Omega^T_{\alpha,1}\right)^c) \\
\leq &\frac{\mathbb{E}[D(\theta_{T+1}) \mid \Omega^T_{\alpha,1}]}{\epsilon} \mathbb{P}(\Omega^T_{\alpha,1}) \\
&+ \mathbb{P}(D(\theta_{T+1}) \geq \epsilon \mathbb{1}^c_{\Omega^T_{\alpha,1}}) \mathbb{P}(\left(\Omega^T_{\alpha,1}\right)^c) \\
\leq &\frac{\mathbb{E}[D(\theta_{T+1})\mathbb{1}_{\Omega^T_{\alpha,1}}]}{\epsilon} + \delta/6 \\
\leq &\frac{t_0}{(T + t_0)\epsilon} D(\theta_1) + \frac{\sigma^2 \ln(T + t_0)}{B C_\alpha \epsilon} + \delta/6,
\end{aligned}$$

where the second inequality follows from Lemma 7. To guarantee $\mathbb{P}(D(\theta_{T+1}) \geq \epsilon) \leq \frac{\delta}{2}$, it suffices to have $T = \frac{t_0 D(\theta_1)}{6\delta\epsilon} - t_0, B = \frac{\sigma^2 \ln(T+t_0)}{6 C_\alpha \delta \epsilon}$. This completes the proof. □

VII. Conclusion

In this work, we studied the convergence and the sample complexity of stochastic PG methods for the entropy-regularized RL with the soft-max parameterization when a good initial policy is given. We proposed a new unbiased PG estimator for the entropy-regularized RL and proved that it

has a bounded variance even though it could be unbounded. In addition, this work provided the first "last iterate" convergence result for stochastic PG methods for the entropy-regularized RL and obtained the sample complexity of $\widetilde{\mathcal{O}}(\frac{1}{\epsilon^2})$, 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. The future direction includes proving the global convergence of the stochastic PG methods for the entropy-regularized RL with an arbitrary initial policy.

ACKNOWLEDGMENT

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

REFERENCES

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

[2] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.

[3] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[4] Brendan O'Donoghue, Remi Munos, Koray Kavukcuoglu, and Volodymyr Mnih. Combining policy gradient and q-learning. *arXiv preprint arXiv:1611.01626*, 2016.

[5] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pages 1352–1361. PMLR, 2017.

[6] Hongyu Zang, Xin Li, Li Zhang, Peiyao Zhao, and Mingzhong Wang. Teac: Intergrating trust region and max entropy actor critic for continuous control. *https://openreview.net/references/pdf?id=bzTQQZQ6ix*, 2020.

[7] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22(98):1–76, 2021.

[8] Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, pages 1–48, 2022.

[9] Lin Xiao. On the convergence rates of policy gradient methods. *arXiv preprint arXiv:2201.07443*, 2022.

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

[11] Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 2021.

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

[13] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

[14] Junyu Zhang, Chengzhuo Ni, Csaba Szepesvari, Mengdi Wang, et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240, 2021.

[15] Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4572–4583. Curran Associates, Inc., 2020.

[16] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. In *Conference on Learning Theory*, pages 3107–3110. PMLR, 2021.

[17] Donghao Ying, Mengzi Guo, Yuhao Ding, Javad Lavaei, et al. Policy-based primal-dual methods for convex constrained markov decision processes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[18] Donghao Ying, Yuhao Ding, and Javad Lavaei. A dual approach to constrained markov decision processes with entropy regularization. In *International Conference on Artificial Intelligence and Statistics*, pages 1887–1909. PMLR, 2022.

[19] Vanshaj Khattar, Yuhao Ding, Bilgehan Sel, Javad Lavaei, and Ming Jin. A cmdp-within-online framework for meta-safe reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[20] Jincheng Mei, Yue Gao, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. *International Conference on Machine Learning*, 2021.

[21] Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33, 2020.

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

[23] Wesley Chung, Valentin Thomas, Marlos C Machado, and Nicolas Le Roux. Beyond variance reduction: Understanding the true impact of baselines on policy optimization. In *International Conference on Machine Learning*, pages 1999–2009. PMLR, 2021.

[24] Jincheng Mei, Bo Dai, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. Understanding the effect of stochasticity in policy optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

[25] John Schulman, Pieter Abbeel, and Xi Chen. Equivalence between policy gradients and soft q-learning. *CoRR*, abs/1704.06440, 2017.

[26] Benjamin Eysenbach and Sergey Levine. If MaxEnt RL is the answer, what is the question? *arXiv preprint arXiv:1910.01913*, 2019.

[27] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[28] Semih Cayci, Niao He, and R Srikant. Linear convergence of entropy-regularized natural policy gradient with linear function approximation. In *International conference on machine learning*. PMLR, 2021.

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

[30] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[31] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. *Advances in Neural Information Processing Systems*, 33:16223–16234, 2020.

[32] Panayotis Mertikopoulos and Zhengyuan Zhou. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1):465–507, 2019.

[33] Panayotis Mertikopoulos, Nadav Hallak, Ali Kavis, and Volkan Cevher. On the almost sure convergence of stochastic gradient descent in non-convex problems. *Advances in Neural Information Processing Systems*, 33:1117–1128, 2020.

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

[35] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[36] Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. *Advances in neural information processing systems*, 30, 2017.