On the Global Convergence Rates of Softmax Policy Gradient Methods

Jincheng Mei [1 2 *] Chenjun Xiao [1] Csaba Szepesvári [3 1] Dale Schuurmans [2 1]

Abstract

We make three contributions toward better understanding policy gradient methods in the tabular setting. First, we show that with the true gradient, policy gradient with a softmax parametrization converges at a $O(1/t)$ rate, with constants depending on the problem and initialization. This result significantly expands the recent asymptotic convergence results. The analysis relies on two findings: that the softmax policy gradient satisfies a Łojasiewicz inequality, and the minimum probability of an optimal action during optimization can be bounded in terms of its initial value. Second, we analyze entropy regularized policy gradient and show that it enjoys a significantly faster linear convergence rate $O(e^{-t})$ toward softmax optimal policy. This result resolves an open question in the recent literature. Finally, combining the above two results and additional new $\Omega(1/t)$ lower bound results, we explain how entropy regularization improves policy optimization, even with the true gradient, from the perspective of convergence rate. The separation of rates is further explained using the notion of non-uniform Łojasiewicz degree. These results provide a theoretical understanding of the impact of entropy and corroborate existing empirical studies.

1. Introduction

The *policy gradient* is one of the most foundational concepts in Reinforcement Learning (RL), lying at the core of policy-search and actor-critic methods. This paper is concerned with the analysis of the convergence rate of *policy gradient methods* (Sutton et al., 2000). As an approach to RL, the appeal of policy gradient methods is that they are conceptually straightforward and under some regularity conditions they guarantee monotonic improvement of the value. A secondary appeal is that policy gradient methods were shown to achieve effective empirical performance (e.g., Schulman et al., 2015; 2017).

Despite the prevalence and importance of policy optimization in RL, the theoretical understanding of policy gradient method has, until recently, been severely limited. A key barrier to understanding is the inherent non-convexity of the value landscape with respect to standard policy parametrizations. As a result, little has been known about the global convergence behavior of policy gradient method. Recently, important new progress in understanding the convergence behavior of policy gradient has been achieved. As in this paper we will restrict ourselves to the tabular setting, we analyze the part of the literature that also deals with this setting. While the tabular setting is clearly limiting, this is the setting where so far the cleanest results have been achieved and understanding this setting is a necessary first step towards the bigger problem of understanding RL algorithms. Returning to the discussion of recent work, Bhandari & Russo (2019) showed that, without parametrization, projected gradient ascent on the simplex does not suffer from spurious local optima. In concurrent work, Agarwal et al. (2019) showed that *(i)* without parametrization, projected gradient ascent converges at rate $O(1/\sqrt{t})$ to a global optimum; and *(ii)* with softmax parametrization, policy gradient converges asymptotically. Agarwal et al. also analyze other variants of policy gradient, and show that policy gradient with relative entropy regularization converges at rate $O(1/\sqrt{t})$, natural policy gradient (mirror descent) converges at rate $O(1/t)$, and given a "compatible" function approximation (thus, going beyond the tabular case) natural policy gradient converges at rate $O(1/\sqrt{t})$. Shani et al. (2020) obtains the slower rate $O(1/\sqrt{t})$ for mirror descent. They also proposed a variant that adds entropy regularization and prove a rate of $O(1/t)$ for this modified problem.

Despite these advances, many open questions remain in understanding the behavior of policy gradient methods, even in the tabular setting and even when the true gradient is available in the updates. In this paper, we provide answers to the following three questions left open by previous work in this area: *(i)* What is the convergence rate of policy gradient methods with softmax parametrization? The best previous result, due to Agarwal et al. (2019), established asymptotic

[*] Work done as an intern at Google Brain. [1] University of Alberta [2] Google Research, Brain Team [3] DeepMind. Correspondence to: Jincheng Mei <jmei2@ualberta.ca>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

convergence but gave no rates. *(ii)* What is the convergence rate of entropy regularized softmax policy gradient? Figuring out the answer to this question was explicitly stated as an open problem by Agarwal et al. (2019). *(iii)* Empirical results suggest that entropy helps optimization (Ahmed et al., 2019). Can this empirical observation be turned into a rigorous theoretical result?[1]

First, we prove that with the true gradient, policy gradient methods with a softmax parametrization converge to the optimal policy at a $O(1/t)$ rate, with constants depending on the problem and initialization. This result significantly strengthens the recent asymptotic convergence results of Agarwal et al. (2019). Our analysis relies on two novel findings: *(i)* that softmax policy gradient satisfies what we call a non-uniform Łojasiewicz-type inequality with the constant in the inequality depending on the optimal action probability under the current policy; *(ii)* the minimum probability of an optimal action during optimization can be bounded in terms of its initial value. Combining these two findings, with a few other properties we describe, it can be shown that softmax policy gradient method achieves a $O(1/t)$ convergence rate.

Second, we analyze entropy regularized policy gradient and show that it enjoys a linear convergence rate of $O(e^{-t})$ toward the softmax optimal policy, which is significantly faster than that of the unregularized version. This result resolves an open question in Agarwal et al. (2019), where the authors analyzed a more aggressive relative entropy regularization rather than the more common entropy regularization. A novel insight is that entropy regularized gradient updates behave similarly to the contraction operator in value learning, with a contraction factor that depends on the current policy.

Third, we provide a theoretical understanding of entropy regularization in policy gradient methods. *(i)* We prove a new lower bound of $\Omega(1/t)$ for softmax policy gradient, implying that the upper bound of $O(1/t)$ that we established, apart from constant factors, is unimprovable. This result also provides a theoretical explanation of the optimization advantage of entropy regularization: even with access to the true gradient, entropy helps policy gradient *converge faster than any achievable rate of softmax policy gradient method without regularization*. *(ii)* We study the concept of non-uniform Łojasiewicz degree and show that, without regularization, the Łojasiewicz degree of expected reward cannot be positive, which allows $O(1/t)$ rates to be established. We then show that with entropy regularization, the Łojasiewicz degree of maximum entropy reward becomes $1/2$, which is sufficient to obtain linear $O(e^{-t})$ rates. This change of the relationship between gradient norm and sub-optimality reveals a deeper reason for the improvement in convergence rates. The theoretical study we provide corroborates existing empirical studies on the impact of entropy in policy optimization (Ahmed et al., 2019).

The remainder of the paper is organized as follows. After introducing notation and defining the setting in Section 2, we present the three main contributions in Sections 3 to 5 as aforementioned. Section 6 gives our conclusions.

2. Notations and Settings

For a finite set $\mathcal{X}$, we use $\Delta(\mathcal{X})$ to denote the set of probability distributions over $\mathcal{X}$. A finite Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$ is determined by a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, transition function $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and discount factor $\gamma \in [0, 1)$. Given a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, the value of state s under π is defined as

$$V^\pi(s) := \underset{\substack{s_0 = s, a_t \sim \pi(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}}{\mathbb{E}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]. \qquad (1)$$

We also let $V^\pi(\rho) := \mathbb{E}_{s \sim \rho} [V^\pi(s)]$, where $\rho \in \Delta(\mathcal{S})$ is an initial state distribution. The state-action value of π at $(s, a) \in \mathcal{S} \times \mathcal{A}$ is defined as

$$Q^\pi(s, a) := r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) V^\pi(s'). \qquad (2)$$

We let $A^\pi(s, a) := Q^\pi(s, a) - V^\pi(s)$ be the so-called advantage function of π. The (discounted) state distribution of π is defined as

$$d_{s_0}^\pi(s) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s | s_0, \pi, \mathcal{P}), \qquad (3)$$

and we let $d_\rho^\pi(s) := \mathbb{E}_{s_0 \sim \rho} \left[d_{s_0}^\pi(s) \right]$. Given ρ, there exists an optimal policy π^* such that

$$V^{\pi^*}(\rho) = \max_{\pi : \mathcal{S} \to \Delta(\mathcal{A})} V^\pi(\rho). \qquad (4)$$

We denote $V^*(\rho) := V^{\pi^*}(\rho)$ for conciseness. Since $\mathcal{S} \times \mathcal{A}$ is finite, for convenience, without loss of generality, we assume that the one step reward lies in the $[0, 1]$ interval:

Assumption 1 (Bounded reward)**.** $r(s, a) \in [0, 1]$, $\forall (s, a)$.

The softmax transform of a vector exponentiates the components of the vector and normalizes it so that the result lies in the simplex. This can be used to transform vectors assigned to state-action pairs into policies:

Softmax transform. Given the function $\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the softmax transform of θ is defined as $\pi_\theta(\cdot|s) :=$

[1] While Shani et al. (2020) suggest that entropy regularization speeds up mirror descent to achieve the rate of $O(1/t)$, in light of the corresponding result of Agarwal et al. (2019) who established the same rate for the unregularized version of mirror descent, their conclusion needs further support (e.g., lower bounds).

$\mathrm{softmax}(\theta(s,\cdot))$, where for all $a \in \mathcal{A}$,

$$\pi_\theta(a|s) = \frac{\exp\{\theta(s,a)\}}{\sum_{a'} \exp\{\theta(s,a')\}}. \tag{5}$$

Due to its origin in logistic regression, we call the values $\theta(s,a)$ the *logit* values and the function θ itself a logit function. We also extend this notation to the case when there are no states: For $\theta : [K] \to \mathbb{R}$, we define $\pi_\theta := \mathrm{softmax}(\theta)$ using $\pi_\theta(a) = \exp\{\theta(a)\} / \sum_{a'} \exp\{\theta(a')\}$ $(a \in [K])$.

H matrix. Given any distribution π over $[K]$, let $H(\pi) := \mathrm{diag}(\pi) - \pi\pi^\top \in \mathbb{R}^{K\times K}$, where $\mathrm{diag}(x) \in \mathbb{R}^{K\times K}$ is the diagonal matrix that has $x \in \mathbb{R}^K$ at its diagonal. The H matrix will play a central role in our analysis because $H(\pi_\theta)$ is the Jacobian of the $\theta \mapsto \pi_\theta := \mathrm{softmax}(\theta)$ map that maps $\mathbb{R}^{[K]}$ to the $(K-1)$-simplex:

$$\left(\frac{d\pi_\theta}{d\theta}\right)^\top = H(\pi_\theta). \tag{6}$$

Here, we are using the standard convention that derivatives give row-vectors. Finally, we recall the definition of smoothness from convex analysis:

Smoothness. A function $f : \Theta \to \mathbb{R}$ with $\Theta \subset \mathbb{R}^d$ is β-smooth (w.r.t. ℓ_2 norm, $\beta > 0$) if for all $\theta, \theta' \in \Theta$,

$$\left| f(\theta') - f(\theta) - \left\langle \frac{df(\theta)}{d\theta}, \theta' - \theta \right\rangle \right| \le \frac{\beta}{2} \cdot \|\theta' - \theta\|_2^2. \tag{7}$$

3. Policy Gradient

Policy gradient is a special policy search method. In policy search, one considers a family of policies parametrized by finite-dimensional parameter vectors, reducing the search for a good policy to searching in the space of parameters. This search is usually accomplished by making incremental changes (additive updates) to the parameters. Representative policy-based RL methods include REINFORCE (Williams, 1992), natural policy gradient (Kakade, 2002), deterministic policy gradient (Silver et al., 2014), and trust region policy optimization (Schulman et al., 2015). In policy gradient methods, the parameters are updated by following the gradient of the map that maps policy parameters to values. Under mild conditions, the gradient can be reexpressed in a convenient form in terms of the policy's action-value function and the gradients of the policy parametrization:

Theorem 1 (Policy gradient theorem (Sutton et al., 2000))**.** *Fix a map $\theta \mapsto \pi_\theta(a|s)$ that for any (s,a) is differentiable and fix an initial distribution $\mu \in \Delta(\mathcal{S})$. Then,*

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d_\mu^{\pi_\theta}} \left[\sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \cdot Q^{\pi_\theta}(s,a) \right].$$

3.1. Vanilla Softmax Policy Gradient

We focus on the policy gradient method that uses the softmax parametrization. Since we consider the tabular case, the policy is then parametrized using the logit $\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ function and $\pi_\theta(\cdot|s) = \mathrm{softmax}(\theta(s,\cdot))$. The vanilla form of policy gradient for this case is shown in Algorithm 1.

Algorithm 1 Policy Gradient Method

```
Input: Learning rate $\eta > 0$.
Initialize logit $\theta_1(s,a)$ for all $(s,a)$.
for $t = 1$ to $T$ do
    $\theta_{t+1} \leftarrow \theta_t + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}$.
end for
```

With some calculation, Theorem 1 can be used to show that the gradient takes the following special form in this case:

Lemma 1. *Softmax policy gradient w.r.t. θ is*

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot A^{\pi_\theta}(s,a). \tag{8}$$

Due to space constraints, the proof of this, as well as of all the remaining results are given in the appendix. While this lemma was known (Agarwal et al., 2019), we included a proof for the sake of completeness.

Recently, Agarwal et al. (2019) showed that softmax policy gradient asymptotically converges to π^*, i.e., $V^{\pi_{\theta_t}}(\rho) \to V^*(\rho)$ as $t \to \infty$ provided that $\mu(s) > 0$ holds for all states $s \in \mathcal{S}$. We strengthen this result to show that the rate of convergence (in terms of value sub-optimality) is $O(1/t)$. The next section is devoted to this result. For better accessibility, we start with the result for the bandit case which presents an opportunity to explaining the main ideas underlying our result in a clean fashion.

3.2. Convergence Rates

3.2.1. The Instructive Case of Bandits

As promised, in this section we consider "bandit case": In particular, assume that the MDP has a single state and the discount factor γ is zero: $\gamma = 0$. In this case, Eq. (1) reduces to maximizing the expected reward,

$$\max_{\theta:\mathcal{A}\to\mathbb{R}} \mathop{\mathbb{E}}_{a\sim\pi_\theta} [r(a)]. \tag{9}$$

With $\pi_\theta = \mathrm{softmax}(\theta)$, even in this simple setting, the objective is non-concave in θ, as shown by a simple example:

Proposition 1. *On some problems, $\theta \mapsto \mathbb{E}_{a\sim\pi_\theta}[r(a)]$ is a non-concave function over $\mathbb{R}^K$.*

As $\gamma = 0$ and there is a single state, Lemma 1 simplifies to

$$\frac{d\pi_\theta^\top r}{d\theta(a)} = \pi_\theta(a) \cdot (r(a) - \pi_\theta^\top r). \tag{10}$$

Putting things together, we see that in this case the update in Algorithm 1 takes the following form:

Update 1 (Softmax policy gradient, expected reward)**.** $\theta_{t+1}(a) \leftarrow \theta_t(a) + \eta \cdot \pi_{\theta_t}(a) \cdot (r(a) - \pi_{\theta_t}^\top r)$, $\forall a \in [K]$.

As is well known, if a function is smooth, then a small gradient update will be guaranteed to improve the objective value. As it turns out, for the softmax parametrization, the expected reward objective is β-smooth with $\beta \le 5/2$:

Lemma 2 (Smoothness)**.** *$\forall r \in [0,1]^K$, $\theta \mapsto \pi_\theta^\top r$ is $5/2$-smooth.*

Smoothness alone (as is also well known) is not sufficient to guarantee that gradient updates converge to a global optimum. For non-concave objectives, the next best thing to guarantee convergence to global maxima is to establish that the gradient of the objective at any parameter dominates the sub-optimality of the parameter. Inequalities of this form are known as a Łojasiewicz inequality (Łojasiewicz, 1963). The reason gradient dominance helps is because it prevents the gradient vanishing before reaching a maximum. The objective function of our problem also satisfies such an inequality, although of a weaker, "non-uniform" form. For the following result, for simplicity, we assume that the optimal action is unique. This assumption can be lifted with a little extra work, which is discussed at the end of this section.

Lemma 3 (Non-uniform Łojasiewicz)**.** *Assume r has one unique maximizing action a^*. Let $\pi^* = \arg\max_{\pi \in \Delta} \pi^\top r$. Then,*

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \ge \pi_\theta(a^*) \cdot (\pi^* - \pi_\theta)^\top r \,. \tag{11}$$

The weakness of this inequality is that the right-hand side scales with $\pi_\theta(a^*)$ – hence we call it non-uniform. As a result, Lemma 3 is not very useful if $\pi_{\theta_t}(a^*)$, the optimal action's probability, becomes very small during the updates.

Nevertheless, the inequality still suffices to get an following intermediate result. The proof of this result combines smoothness and the Łojasiewicz inequality we derived.

Lemma 4 (Pseudo-rate)**.** *Let $c_t = \min_{1 \le s \le t} \pi_{\theta_s}(a^*)$. Using Update 1 with $\eta = 2/5$, for all $t \ge 1$,*

$$(\pi^* - \pi_{\theta_t})^\top r \le 5/(t \cdot c_t^2), \qquad \textit{and}$$
$$\sum_{t=1}^{T} (\pi^* - \pi_{\theta_t})^\top r \le \min\left\{ \sqrt{5T}/c_T,\ (5 \log T)/c_T^2 + 1 \right\}.$$

In the remainder of this section we assume that $\eta = 2/5$.

Remark 1. *The value of $\pi_{\theta_t}(a^*)$, while it is nonzero (and so is c_t) can be small (e.g., because of the choice of θ_1). Consequently, its minimum c_t can be quite small and the upper bound in Lemma 4 can be large, or even vacuous. The dependence of the previous result on $\pi_{\theta_t}(a^*)$ comes from Lemma 3. As it turns out, it is not possible to eliminate or improve the dependence on $\pi_\theta(a^*)$ in Lemma 3. To see this consider $r = (5,4,4)^\top$, $\pi_\theta = (2\epsilon, 1/2 - 2\epsilon, 1/2)$ where $\epsilon > 0$ is small number. By algebra, $(\pi^* - \pi_\theta)^\top r = 1 - 2\epsilon > 1/2$, $\frac{d\pi_\theta^\top r}{d\theta} = (2\epsilon - 4\epsilon^2, -\epsilon + 4\epsilon^2, -\epsilon)^\top$, $\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 = \epsilon \cdot \sqrt{6 - 24\epsilon + 32\epsilon^2} \le 3\epsilon$. Hence, for any constant $C > 0$,*

$$C \cdot (\pi^* - \pi_\theta)^\top r > C/2 > 3\epsilon \ge \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 , \tag{12}$$

which means for any Łojasiewicz-type inequality, C necessarily depends on ϵ and hence on $\pi_\theta(a^*) = 2\epsilon$.

The necessary dependence on $\pi_{\theta_t}(a^*)$ makes it clear that Lemma 4 is insufficient to conclude a $O(1/t)$ rate. since c_t may vanish faster than $O(1/t)$ as t increases. Our next result eliminates this possibility. In particular, the result follows from the asymptotic convergence result of Agarwal et al. (2019) which states that $\pi_{\theta_t}(a^*) \to 1$ as $t \to \infty$. From this and because $\pi_\theta(a) > 0$ for any $\theta \in \mathbb{R}^K$ and action a, we conclude that $\pi_{\theta_t}(a^*)$ remains bounded away from zero during the course of the updates:

Lemma 5. *We have $\inf_{t \ge 1} \pi_{\theta_t}(a^*) > 0$.*

With some extra work, one can also show that eventually θ_t enters a region where $\pi_{\theta_t}(a^*)$ can only increase:

Proposition 2. *For any initialization there exist $t_0 \ge 1$ such that for any $t \ge t_0$, $t \mapsto \pi_{\theta_t}(a^*)$ is increasing. In particular, when π_{θ_1} is the uniform distribution, $t_0 = 1$.*

With Lemmas 4 and 5, we can now obtain an $O(1/t)$ convergence rate for softmax policy gradient method[2]:

Theorem 2 (Arbitrary initialization)**.** *Using Update 1 with $\eta = 2/5$, for all $t \ge 1$,*

$$(\pi^* - \pi_{\theta_t})^\top r \le 1/(c^2 \cdot t), \tag{13}$$

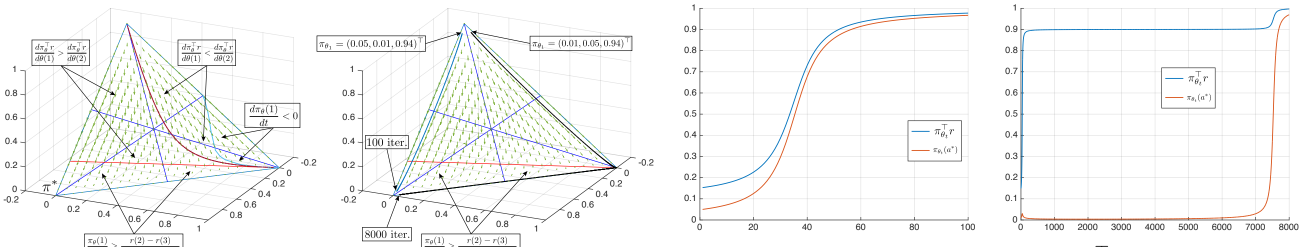
where $c = \inf_{t \ge 1} \pi_{\theta_t}(a^*) > 0$ is a constant that depends on r and θ_1, but it does not depend on the time t.

Proposition 2 suggests that one should set θ_1 so that π_{θ_1} is uniform. Using this initialization, we can show that $\inf_{t \ge 1} \pi_{\theta_t}(a^*) \ge 1/K$, strengthening Theorem 2:

Theorem 3 (Uniform initialization)**.** *Using Update 1 with $\eta = 2/5$ and θ_1 such that $\pi_{\theta_1}(a) = 1/K$, $\forall a$, for all $t \ge 1$,*

$$(\pi^* - \pi_{\theta_t})^\top r \le 5K^2/t, \qquad \textit{and}$$
$$\sum_{t=1}^{T} (\pi^* - \pi_{\theta_t})^\top r \le \min\left\{ K\sqrt{5T},\ 5K^2 \log T + 1 \right\}.$$

[2] For a continuous version of Update 1, Walton (2020) proves a $O(1/t)$ rate, using a Lyapunov function argument.



(a) Softmax gradient flow. (b) Good & bad initializations. (c) $\pi_{\theta_t}^\top r$ and $\pi_{\theta_t}(a^*)$ for good initialization. (d) $\pi_{\theta_t}^\top r$ and $\pi_{\theta_t}(a^*)$ for bad initialization.

Figure 1. Visualization of proof idea for Lemma 5.

Remark 2. *In Section 5, we prove a lower bound $\Omega(1/t)$ for the same update rule, showing that the upper bound $O(1/t)$ of Theorem 2, apart from constant factors, is unimprovable.*

In general it is difficult to characterize how the constant C in Theorem 2 depends on the problem and initialization. For the simple 3-armed case, this dependence is relatively clear:

Lemma 6. *Let $r(1) > r(2) > r(3)$. Then, $a^* = 1$ and $\inf_{t\geq 1} \pi_{\theta_t}(a^*) = \min_{1\leq t\leq t_0} \pi_{\theta_t}(1)$, where*

$$t_0 = \min\left\{ t \geq 1 \,:\, \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} \geq \frac{r(2)-r(3)}{2\cdot(r(1)-r(2))} \right\}. \quad (14)$$

Note that the smaller $r(1) - r(2)$ and $\pi_{\theta_1}(1)$ are, the larger t_0 is, which potentially means C in Theorem 2 can be larger.

Visualization. Let $r = (1.0, 0.9, 0.1)^\top$. In Fig. 1(a), the region below the red line corresponds to $\mathcal{R} = \{\pi_\theta : \pi_\theta(1)/\pi_\theta(3) \geq (r(2)-r(3))/(2\cdot(r(1)-r(2)))\}$. Any globally convergent iteration will enter $\mathcal{R}$ within finite time (the closure of $\mathcal{R}$ contains π^*) and never leaves $\mathcal{R}$ (this is the main idea in Lemma 5). Subfigure (b) shows the behavior of the gradient updates with "good" ($\pi_{\theta_1} = (0.05, 0.01, 0.94)^\top$) and "bad" ($\pi_{\theta_1} = (0.01, 0.05, 0.94)^\top$) initial policies. While these are close to each other, the iterates behave quite differently (in both cases $\eta = 2/5$). From the good initialization, the iterates converge quickly: after 100 iterations the distance to the optimal policy is already quite small. At the same time, starting from a "bad" initial value, the iterates are first attracted toward a sub-optimal action. It takes more than 7000 iterations for the algorithm to escape this sub-optimal corner! In subfigure (c), we see that $\pi_{\theta_t}(a^*)$ increases for the good initialization, while in subfigure (d), for the bad initialization, we see that it initially decreases. These experiments confirm that the dependence of the error bound in Theorem 2 on the initial values cannot be removed.

Non-unique optimal actions. When the optimal action is non-unique, the arguments need to be slightly modified. Instead of using a single $\pi_\theta(a^*)$, we need to consider $\sum_{a^*\in\mathcal{A}^*} \pi_\theta(a^*)$, i.e., the sum of probabilities of all optimal actions. Details are given in the appendix.

3.2.2. General MDPs

For general MDPs, the optimization problem takes the form

$$\max_{\theta:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} V^{\pi_\theta}(\rho) = \max_{\theta:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \mathop{\mathbb{E}}_{s\sim\rho} \sum_a \pi_\theta(a|s)\cdot Q^{\pi_\theta}(s,a).$$

Here, as before, $\pi_\theta(\cdot|s) = \text{softmax}(\theta(s,\cdot))$, $s \in \mathcal{S}$. Following Agarwal et al. (2019), the values here are defined with respect to an initial state distribution ρ which may not be the same as the initial state distribution μ used in the gradient updates (cf. Algorithm 1), allowing for greater flexibility in our analysis. While the initial state distributions do not play any role in the bandit case, here, in the multi-state case, they have a strong influence. In particular, for the rest of this section, we will assume that the initial state distribution μ used in the gradient updates is bounded away from zero:

Assumption 2 (Sufficient exploration)**.** *The initial state distribution satisfies* $\min_s \mu(s) > 0$.

Assumption 2 was also adapted by Agarwal et al. (2019), which ensures "sufficient exploration" in the sense that the occupancy measure d_μ^π of any policy π when started from μ will be guaranteed to be positive over the whole state space. Agarwal et al. (2019) asked whether this assumption is necessary for convergence to global optimality.

Proposition 3. *There exists an MDP and μ with* $\min_s \mu(s) = 0$ *such that there exists $\theta^* : \mathcal{S}\times\mathcal{A} \to [0,\infty]$ such that θ^* is the stationary point of $\theta \mapsto V^{\pi_\theta}(\mu)$ while π_{θ^*} is not an optimal policy. Furthermore, this stationary point is an attractor, hence, starting gradient ascent in a small enough vicinity of θ^* will make it converge to θ^*.*

The MDP of this proposition is S bandit problems: Each state $s \in \mathcal{S}$ under each action deterministically gives itself as the next state. The reward is selected so that in each s there is a unique optimal action. If μ leaves out state s (i.e., $\mu(s) = 0$), clearly, the gradient of $\theta \mapsto V^{\pi_\theta}(\mu)$ w.r.t. θ is zero regardless of the choice of θ. Hence, any θ such that $\theta(s,a) = +\infty$ for a optimal in state s with $\mu(s) > 0$ and $\theta(s,a)$ finite otherwise will satisfy the properties of the proposition. It remains open whether the sufficient exploration condition is necessary for unichain MDPs.

According to Assumption 1, $r(s,a) \in [0,1]$, $Q(s,a) \in [0, 1/(1-\gamma)]$, and hence the objective function is still smooth, as was also shown by Agarwal et al. (2019):

Lemma 7 (Smoothness)**.** *$V^{\pi_\theta}(\rho)$ is $8/(1-\gamma)^3$-smooth.*

As mentioned in Section 3.2.1, smoothness and (uniform) Łojasiewicz inequality are sufficient to prove a convergence rate. As noted by Agarwal et al. (2019), the main difficulty is to establish a (uniform) Łojasiewicz inequality for softmax parametrization. As it turns out, the results from the bandit case carry over to multi-state MDPs.

For stating this and the remaining results, we fix a *deterministic* optimal policy π^* and denote by $a^*(s)$ the action that π^* selects in state s. With this, the promised result on the non-uniform Łojasiewicz inequality is as follows:

Lemma 8 (Non-uniform Łojasiewicz)**.** *We have,*

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \ge \frac{\min_s \pi_\theta(a^*(s)|s)}{\sqrt{S} \cdot \left\| d_\rho^{\pi^*} / d_\mu^{\pi_\theta} \right\|_\infty} \cdot \left[V^*(\rho) - V^{\pi_\theta}(\rho) \right] .$$

By Assumption 2, $d_\mu^{\pi_\theta}$ is also bounded away from zero on the whole state space and thus the multiplier of the sub-optimality in the above inequality is positive.

Generalizing Lemma 5, we show that $\min_s \pi_{\theta_t}(a^*(s)|s)$ is uniformly bounded away from zero:

Lemma 9. *Let Assumption 2 hold. Using Algorithm 1, we have, $c := \inf_{s \in \mathcal{S}, t \ge 1} \pi_{\theta_t}(a^*(s)|s) > 0$.*

Using Lemmas 7 to 9, we prove that softmax policy gradient converges to an optimal policy at a $O(1/t)$ rate in MDPs, just like what we have seen in the bandit case:

Theorem 4. *Let Assumption 2 hold and let $\{\theta_t\}_{t \ge 1}$ be generated using Algorithm 1 with $\eta = (1-\gamma)^3/8$, c the positive constant from Lemma 9. Then, for all $t \ge 1$,*

$$V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \le \frac{16 S}{c^2 (1-\gamma)^6 t} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^2 \cdot \left\| \frac{1}{\mu} \right\|_\infty .$$

As far as we know, this is the first convergence-rate result for softmax policy gradient for MDPs.

Remark 3. *Theorem 4 implies that the iteration complexity of Algorithm 1 to achieve $O(\epsilon)$ sub-optimality is $O\Big(\frac{S}{c^2(1-\gamma)^6\epsilon} \cdot \Big\|\frac{d_\mu^{\pi^*}}{\mu}\Big\|_\infty^2 \cdot \Big\|\frac{1}{\mu}\Big\|_\infty\Big)$, which, as a function of ϵ, is better than the results of Agarwal et al. (2019) for* (i) *projected gradient ascent on the simplex ($O\Big(\frac{SA}{(1-\gamma)^6\epsilon^2} \cdot \Big\|\frac{d_\rho^{\pi^*}}{\mu}\Big\|_\infty^2\Big)$) or for* (ii) *softmax policy gradient with relative-entropy regularization ($O\Big(\frac{S^2A^2}{(1-\gamma)^6\epsilon^2} \cdot \Big\|\frac{d_\rho^{\pi^*}}{\mu}\Big\|_\infty^2\Big)$). The improved dependence on ϵ (or t) in our result follows from Lemmas 8 and 9 and a different proof technique utilized to prove Theorem 4, while we pay a price because our bound depends on c, which adds an extra dependence on the MDP as well as on the initialization of the algorithm.*

4. Entropy Regularized Policy Gradient

Agarwal et al. (2019) considered relative-entropy regularization in policy gradient to get an $O(1/\sqrt{t})$ convergence rate. As they note, relative-entropy is more "agressive" in penalizing small probabilities than the more "common" entropy regularizer (cf. Remark 5.5 in their paper) and it remains unclear whether this latter regularizer leads to an algorithm with the same rate. In this section, we answer this positively and in fact prove a much better rate. In particular, we show that entropy regularized policy gradient with the softmax parametrization enjoys a linear rate of $O(e^{-t})$. In retrospect, perhaps this is unsurprising as entropy regularization bears a strong similarity to introducing a strongly convex regularizer in convex optimization, where this change is known to significantly improve the rate of convergence of first-order methods (e.g., Nesterov, 2018, Chapter 2).

4.1. Maximum Entropy RL

In entropy regularized RL, or sometimes called maximum entropy RL, near-deterministic policies are penalized (Williams & Peng, 1991; Mnih et al., 2016; Nachum et al., 2017; Haarnoja et al., 2018; Mei et al., 2019), which is achieved by modifying the value of a policy π to

$$\tilde{V}^\pi(\rho) := V^\pi(\rho) + \tau \cdot \mathbb{H}(\rho, \pi) , \tag{15}$$

where $\mathbb{H}(\rho, \pi)$ is the "discounted entropy", defined as

$$\mathbb{H}(\rho, \pi) := \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, a_t \sim \pi(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\sum_{t=0}^\infty -\gamma^t \log \pi(a_t|s_t) \right] . \tag{16}$$

and $\tau \ge 0$, the "temperature", determines the strength of the penalty.[3] Clearly, the value of any policy can be obtained by adding an entropy penalty to the rewards (as proposed originally by Williams & Peng (1991)). Hence, similarly to Lemma 1, one can obtain the following expression for the gradient of the entropy regularized objective under the softmax policy parametrization:

Lemma 10. *It holds that for all (s, a),*

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot \tilde{A}^{\pi_\theta}(s,a), \tag{17}$$

[3] To better align with naming conventions in information-theory, discounted entropy should be rather called the discounted action-entropy rate as entropy itself in the literature on Markov chain information theory would normally refer to the entropy of the stationary distribution of the chain, while entropy rate refers to what is being used here.

where $\tilde{A}^{\pi_\theta}(s,a)$ is the "soft" advantage function defined as

$$\tilde{A}^{\pi_\theta}(s,a) \coloneqq \tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) - \tilde{V}^{\pi_\theta}(s), \quad (18)$$

$$\tilde{Q}^{\pi_\theta}(s,a) \coloneqq r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \tilde{V}^{\pi_\theta}(s'). \quad (19)$$

4.2. Convergence Rates

As in the non-regularized case, to gain insight, we first consider MDPs with a single state and $\gamma = 0$.

4.2.1. Bandit Case

In the one-state case with $\gamma = 0$, Eq. (15) reduces to maximizing the entropy-regularized reward,

$$\max_{\theta : \mathcal{A} \to \mathbb{R}} \mathop{\mathbb{E}}_{a \sim \pi_\theta} [r(a) - \tau \log \pi_\theta(a)]. \quad (20)$$

Again, Eq. (20) is a non-concave function of θ. In this case, regularized policy gradient reduces to

$$\frac{d\{\pi_\theta^\top (r - \tau \log \pi_\theta)\}}{d\theta} = H(\pi_\theta)(r - \tau \log \pi_\theta), \quad (21)$$

where $H(\pi_\theta)$ is the same as in Eq. (6). Using the above gradient in Algorithm 1 we have the following update rule:

Update 2 (Softmax policy gradient, maximum entropy reward)**.** $\theta_{t+1} \leftarrow \theta_t + \eta \cdot H(\pi_{\theta_t})(r - \tau \log \pi_{\theta_t})$.

Due to the presence of regularization, the optimal solution will be biased with the bias disappearing as $\tau \to 0$:

Softmax optimal policy. $\pi_\tau^* \coloneqq \mathrm{softmax}(r/\tau)$ is the optimal solution of Eq. (20).

Remark 4. *At this stage, we could use arguments similar to those of Section 3 to show the $O(1/t)$ convergence of π_{θ_t} to π_τ^*. However, we can use an alternative idea to show that entropy-regularized policy gradient converges significantly faster. The issue of bias will be discussed later.*

Our alternative idea is to show that Update 2 defines a contraction but with a contraction coefficient that depends on the parameter that the update is applied to:

Lemma 11 (Non-uniform contraction)**.** *Using Update 2 with $\tau\eta \le 1$, $\forall t > 0$,*

$$\|\zeta_{t+1}\|_2 \le \left(1 - \tau\eta \cdot \min_a \pi_{\theta_t}(a)\right) \cdot \|\zeta_t\|_2, \quad (22)$$

where $\zeta_t \coloneqq \tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}$.

This lemma immediately implies the following bound:

Lemma 12. *Using Update 2 with $\tau\eta \le 1$, $\forall t > 0$,*

$$\|\zeta_t\|_2 \le \frac{2(\tau\|\theta_1\|_\infty + 1)\sqrt{K}}{\exp\left\{\tau\eta \sum_{s=1}^{t-1} [\min_a \pi_{\theta_s}(a)]\right\}}. \quad (23)$$

Similarly to Lemma 5, we can show that the minimum action probability can be lower bounded by its initial value.

Lemma 13. *There exists $c = c(\tau, K, \|\theta_1\|_\infty) > 0$, such that for all $t \ge 1$, $\min_a \pi_{\theta_t}(a) \ge c$. Thus, $\sum_{s=1}^{t-1} [\min_a \pi_{\theta_s}(a)] \ge c \cdot (t-1)$.*

A closed-form expression for c is given in the appendix. Note that when $\tau = 0$ (no regularization), the result would no longer hold true. The key here is that $\min_a \pi_{\theta_t}(a) \to \min_a \pi_\tau^*(a) > 0$ as $t \to \infty$ and the latter inequality holds thanks to $\tau > 0$. From Lemmas 12 and 13, it follows that entropy regularized softmax policy gradient enjoys a linear convergence rate:

Theorem 5. *Using Update 2 with $\eta \le 1/\tau$, for all $t \ge 1$,*

$$\tilde{\delta}_t \le \frac{2(\tau\|\theta_1\|_\infty + 1)^2 K/\tau}{\exp\{2\tau\eta \cdot c \cdot (t-1)\}}, \quad (24)$$

where $\tilde{\delta}_t \coloneqq \pi_\tau^{*\top}(r - \tau \log \pi_\tau^*) - \pi_{\theta_t}^\top (r - \tau \log \pi_{\theta_t})$ and $c > 0$ is from Lemma 13.

4.2.2. General MDPs

For general MDPs, the problem is to maximize $\tilde{V}^{\pi_\theta}(\rho)$ in Eq. (15). The softmax optimal policy π_τ^* is known to satisfy the following consistency conditions (Nachum et al., 2017):

$$\pi_\tau^*(a|s) = \exp\left\{(\tilde{Q}^{\pi_\tau^*}(s,a) - \tilde{V}^{\pi_\tau^*}(s))/\tau\right\}, \quad (25)$$

$$\tilde{V}^{\pi_\tau^*}(s) = \tau \log \sum_a \exp\left\{\tilde{Q}^{\pi_\tau^*}(s,a)/\tau\right\}. \quad (26)$$

Using a somewhat lengthy calculation, we show that the discounted entropy in Eq. (16) is smooth:

Lemma 14 (Smoothness)**.** *$\mathbb{H}(\rho, \pi_\theta)$ is $(4 + 8\log A)/(1-\gamma)^3$-smooth, where $A \coloneqq |\mathcal{A}|$ is the total number of actions.*

Our next key result shows that the augmented value function $\tilde{V}^{\pi_\theta}(\rho)$ satisfies a "better type" of Łojasiewicz inequality:

Lemma 15 (Non-uniform Łojasiewicz)**.** *Suppose $\mu(s) > 0$ for all state $s \in \mathcal{S}$. Then,*

$$\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \ge C(\theta) \cdot \left[\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho)\right]^{\frac{1}{2}}, \quad (27)$$

where

$$C(\theta) \coloneqq \frac{\sqrt{2\tau}}{\sqrt{S}} \cdot \min_s \sqrt{\mu(s)} \cdot \min_{s,a} \pi_\theta(a|s) \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}}.$$

The main difference to the previous versions of the non-uniform Łojasiewicz inequality is that the sub-optimality gap appears under the square root. For small sub-optimality gaps this means that the gradient must be larger – a stronger "signal". Next, we show that action probabilities are still uniformly bounded away from zero:

Lemma 16. *Using Algorithm 1 with the entropy regularized objective, we have $c := \inf_{t \geq 1} \min_{s,a} \pi_{\theta_t}(a|s) > 0$.*

With Lemmas 14 to 16, we show a $O(e^{-t})$ rate for entropy regularized policy gradient in general MDPs:

Theorem 6. *Suppose $\mu(s) > 0$ for all state s. Using Algorithm 1 with the entropy regularized objective and softmax parametrization and $\eta = (1-\gamma)^3/(8 + \tau(4 + 8 \log A))$, there exists a constant $C > 0$ such that for all $t \geq 1$,*

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

The value of the constant C in this theorem appears in the proof of the result in the appendix in a closed form.

4.2.3. Controlling the Bias

As noted in Remark 4, π^*_τ is biased, i.e., $\pi^*_\tau \neq \pi^*$ for fixed $\tau > 0$. We discuss two possible approaches to deal with the bias, but much remains to be done to properly address the bias. For simplicity, we consider the bandit case.

A two-stage approach. Note that for any fixed $\tau > 0$, $\pi^*_\tau(a^*) \geq \pi^*_\tau(a)$ for all $a \neq a^*$. Therefore, using policy gradient with $\pi_{\theta_1} = \pi^*_\tau$, we have $\pi_{\theta_t}(a^*) \geq c_t \geq 1/K$. This suggests a two-stage method: first, to ensure $\pi_{\theta_t}(a^*) \geq \max_a \pi_{\theta_t}(a)$, use entropy-regularized policy gradient some iterations and then turn off regularization.

Theorem 7. *Denote $\Delta = r(a^*) - \max_{a \neq a^*} r(a) > 0$. Using Update 2 for $t_1 \in O(e^{1/\tau} \cdot \log(\frac{\tau+1}{\Delta}))$ iterations and then Update 1 for $t_2 \geq 1$ iterations, we have,*

$$(\pi^* - \pi_{\theta_t})^\top r \leq 5/(C^2 \cdot t_2), \tag{28}$$

where $t = t_1 + t_2$, and $C \in [1/K, 1)$.

This approach removes the nasty dependence on the choice of the initial parameters. While this dependence is also removed if we initialize with the uniform policy, uniform initialization is insufficient if only noisy estimates of the gradients are available. However, we leave the study of this case for future work. An obvious problem with this approach is that Δ is unknown. This can be helped by exiting the first phase when we detect "convergence" e.g. by detecting that the relative change of the policy is small.

Decreasing the penalty. Another simple idea is to decrease the strength of regularization, e.g., set $\tau_t \in O(1/\log t)$. Consider the following update, which is a slight variation of the previous one:

Update 3. $\theta_{t+1} \leftarrow \frac{\tau_t}{\tau_{t+1}} \cdot (\theta_t + \eta_t \cdot H(\pi_{\theta_t})(r - \tau_t \log \pi_{\theta_t}))$.

The rationale for the scaling factor is that it allows one to prove a variant of Lemma 11. While this is promising, the proof cannot be finished as before. The difficulty is that $\pi_{\theta_t} \to \pi^*$ (which is what we want to achieve) implies that $\min_a \pi_{\theta_t}(a) \to 0$, which prevents the use of our previous proof technique. We show the following partial results.

Theorem 8. *Using Update 3 with $\tau_t = \frac{\alpha \cdot \Delta}{\log t}$ for $t \geq 2$, where $\alpha > 0$, and $\eta_t = 1/\tau_t$, we have, for all $t \geq 1$,*

$$(\pi^* - \pi_{\theta_t})^\top r \leq \frac{K}{t^{1/\alpha}} + \frac{C \cdot \log t}{\exp\{\sum_{s=1}^{t-1} [\min_a \pi_{\theta_s}(a)]\}},$$

where $C := \frac{2(\tau_1 \|\theta_1\|_\infty + 1)\sqrt{K}}{\alpha \cdot \Delta}$.

The final rates then depend on how fast $\min_a \pi_{\theta_t}(a)$ diminishes as function of t. We conjecture that the rate in some cases degenerates to $O(\frac{\log t}{t^{1/\alpha}})$, which is strictly faster than $O(1/t)$ in non-regularized case when $\alpha \in (0, 1)$ and is observed in simulations in the appendix. We leave it as an open problem to study decaying entropy in general MDPs.

5. Does Entropy Regularization Really Help?

The previous section indicated that entropy regularization may speed up convergence. In addition, ample empirical evidence suggest that this may be the case (e.g., Williams & Peng, 1991; Mnih et al., 2016; Nachum et al., 2017; Haarnoja et al., 2018; Mei et al., 2019). In this section, we aim to provide new insights into why entropy may help policy optimization, taking an optimization perspective.

We start by establishing a lower bound that shows that the $O(1/t)$ rate we established earlier for softmax policy gradient without entropy regularization cannot be improved. Next, we introduce the notion of Łojasiewicz degree, which we show to increase in the presence of entropy regularization. We then connect a higher degree to faster convergence rates. Note that our proposal to view entropy regularization as an optimization aid is somewhat conflicting with the more common explanation that entropy regularization helps by encouraging exploration. While it is definitely true that entropy regularization encourages exploration, the form of exploration it encourages is not sensitive to epistemic uncertainty and as such it fails to provide a satisfactory solution to the exploration problem (e.g., O'Donoghue et al., 2020).

5.1. Lower Bounds

The purpose of this section is to establish that the $O(1/t)$ rates established earlier for unpenalized policy gradient is tight. To get lower bounds, we need to show that progress in every iteration cannot be too large. This holds when we can reverse the inequality in the Łojasiewicz inequality. To this regard, in bandit problems we have the following result:

Lemma 17 (Reversed Łojasiewicz)**.** *Take any $r \in [0, 1]^K$.*

Denote $\Delta = r(a^*) - \max_{a \neq a^*} r(a) > 0$. *Then,*

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \le (\sqrt{2}/\Delta) \cdot (\pi^* - \pi_\theta)^\top r. \tag{29}$$

Using this result gives the desired lower bound:

Theorem 9 (Lower bound)**.** *Take any* $r \in [0,1]^K$. *For large enough* $t \ge 1$, *using Update 1 with learning rate* $\eta_t \in (0,1]$,

$$(\pi^* - \pi_{\theta_t})^\top r \ge \Delta^2/(6 \cdot t). \tag{30}$$

Note that Theorem 9 is a special case of general MDPs. Next, we strengthen this result and show that the $\Omega(1/t)$ lower bound also holds for *any* MDP:

Theorem 10 (Lower bound)**.** *Take any MDP. For large enough* $t \ge 1$, *using Algorithm 1 with* $\eta_t \in (0,1]$,

$$V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \ge \frac{(1-\gamma)^5 \cdot (\Delta^*)^2}{12 \cdot t}, \tag{31}$$

where $\Delta^* := \min_{s \in \mathcal{S}, a \neq a^*(s)} \{Q^*(s, a^*(s)) - Q^*(s,a)\} > 0$ *is the optimal value gap of the MDP.*

Remark 5. *Our convergence rates in Section 3 match the lower bounds up to constant. However, the constant gap is large, e.g.,* K^2 *in Theorem 3, and* Δ^2 *in Theorem 9. The gap is because the reversed Łojasiewicz inequality of Lemma 17 uses* Δ, *which is unavoidable when* π_θ *is close to* π^*. *We leave it as an open problem to close this gap.*

With the lower bounds established, we confirm that entropy regularization helps policy optimization by speeding up convergence, though the question remains as to the mechanism by which the improved convergence rate manifests itself.

5.2. Non-uniform Łojasiewicz Degree

To gain further insight into how entropy regularization helps, we introduce the non-uniform Łojasiewicz degree:

Definition 1 (Non-uniform Łojasiewicz degree)**.** *A function* $f : \mathcal{X} \to \mathbb{R}$ *has Łojasiewicz degree* $\xi \in [0,1]$ *if*[4]

$$\|\nabla_x f(x)\|_2 \ge C(x) \cdot |f(x) - f(x^*)|^{1-\xi}, \tag{32}$$

$\forall x \in \mathcal{X}$, *where* $C(x) > 0$ *holds for all* $x \in \mathcal{X}$.

The uniform degree, where $C(x)$ is a positive constant, has previously been connected to convergence speed in the optimization literature. Bárta (2017) studied this effect for first-, while Nesterov & Polyak (2006); Zhou et al. (2018) studied this for second-order methods. As noted beforehand, a larger degree (smaller exponent of the sub-optimality) is expected to improve the convergence speed of algorithms that rely on gradient information. Intuitively, we expect this to continue to hold for the non-uniform Łojasiewicz degree as well. With this, we now study what Łojasiewicz degrees can one obtain with and without entropy regularization.

Our first result shows that the Łojasiewicz degree of the expected reward objective (in bandits) cannot be positive:

Proposition 4. *Let* $r \in [0,1]^K$ *be arbitrary and consider* $\theta \mapsto \mathbb{E}_{a \sim \pi_\theta}[r(a)]$. *The non-uniform Łojasiewicz degree of this map with constant* $C(\theta) = \pi_\theta(a^*)$ *is zero.*

Note that according to Remark 1, it is necessary that $C(\theta)$ depends on $\pi_\theta(a^*)$. The difference between Proposition 4 and the reversed Łojasiewicz inequality of Lemma 17 is subtle. Lemma 17 is a condition that implies impossibility to get rates faster than $O(1/t)$, while Proposition 4 says it is not sufficient to get rates faster than $O(1/t)$ *using the same technique as in Lemma 4*. However, this does not preclude that other techniques could give faster rates.

Next, we show that the Łojasiewicz degree of the entropy-regularized expected reward objective is at least $1/2$:

Proposition 5. *Fix* $\tau > 0$. *With* $C(\theta) = \sqrt{2\tau} \cdot \min_a \pi_\theta(a)$, *the Łojasiewicz degree of* $\theta \mapsto \mathbb{E}_{a \sim \pi_\theta}[r(a) - \tau \log \pi_\theta(a)]$ *is at least* $1/2$.

6. Conclusions and Future Work

We set out to study the convergence speed of softmax policy gradient methods with and without entropy regularization in the tabular setting. Here, the error is measured in terms of the sub-optimality of the policy obtained after some number of updates. Our main findings is that without entropy regularization, the rate is $\Theta(1/t)$, which is faster than rates previously obtained. Our analysis also uncovered an unpleasant dependence on the initial parameter values. With entropy regularization, the rate becomes linear, where now the constant in the exponent is influenced by the initial choice of parameters. Thus, our analysis shows that entropy regularization substantially changes the rate at which gradient methods converge. Our main technical innovation is the introduction of a non-uniform variant of the Łojasiewicz inequality. Our work leaves open a number of interesting questions: While we have some lower bounds, there remains some gaps to be filled between the lower and upper bounds. Other interesting directions are extending the results for alternative (e.g., restricted) policy parametrizations or studying policy gradient when the gradient must be estimated from data. One also expects that non-uniform Łojasiewicz inequalities and the Łojasiewicz degree could also be put to good use in other areas of non-convex optimization.

[4] In literature (Łojasiewicz, 1963), C cannot depend on x. Based on the examples we have seen, we relax this requirement.

Acknowledgements

Jincheng Mei would like to thank Bo Dai and Lihong Li for helpful discussions and for providing feedback on a draft of this manuscript. Jincheng Mei would like to thank Ruitong Huang for enlightening early discussions. Csaba Szepesvári gratefully acknowledges funding from the Canada CIFAR AI Chairs Program, Amii and NSERC.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in Markov decision processes, 2019.

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

Bárta, T. Rate of convergence to equilibrium and Łojasiewicz-type estimates. *Journal of Dynamics and Differential Equations*, 29(4):1553–1568, 2017.

Bhandari, J. and Russo, D. Global optimality guarantees for policy gradient methods, 2019.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870, 2018.

Kakade, S. M. A natural policy gradient. In *Advances in neural information processing systems*, pp. 1531–1538, 2002.

Łojasiewicz, S. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

Mei, J., Xiao, C., Huang, R., Schuurmans, D., and Müller, M. On principled entropy exploration in policy optimization. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 3130–3136. AAAI Press, 2019.

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

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 2775–2785, 2017.

Nesterov, Y. *Lectures on convex optimization*, volume 137. Springer, 2018.

Nesterov, Y. and Polyak, B. T. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

O'Donoghue, B., Osband, I., and Ionescu, C. Making sense of reinforcement learning and probabilistic inference. *arXiv preprint arXiv:2001.00805*, 2020.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shani, L., Efroni, Y., and Mannor, S. Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs. In *AAAI*, 2020.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *International Conference on Machine Learning*, pp. 387–395, 2014.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pp. 1057–1063, 2000.

Walton, N. A short note on soft-max and policy gradients in bandits problems. *arXiv preprint arXiv:2007.10297*, 2020.

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

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

Zhou, Y., Wang, Z., and Liang, Y. Convergence of cubic regularization for nonconvex optimization under KL property. In *Advances in Neural Information Processing Systems*, pp. 3760–3769, 2018.