On the Global Convergence Rates of Softmax Policy Gradient Methods

Jincheng Mei[♣][♠][*] **Chenjun Xiao**[♣] **Csaba Szepesvári**[♡][♣] **Dale Schuurmans**[♠][♣]

[♣]University of Alberta [♡]DeepMind, Edmonton [♠]Google Research, Brain Team

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. The policy gradient theorem (Sutton et al., 2000), in particular, establishes a general foundation for policy search methods, by showing that an unbiased estimate of the gradient of a policy's expected return with respect to its parameters can still be recovered from an approximate value function (provided the approximation is a best fit). As an approach to RL, policy gradient ascent is particularly appealing due to its simplicity and directness: it targets the quantity of interest, it is inherently sound given appropriate step size control, and it can be readily combined with network function approximation to achieve effective empirical performance (e.g., Schulman et al., 2015; 2017).

Despite the prevalence and importance of policy optimization methods in RL, the theoretical understanding of policy gradient ascent has, until recently, been severely limited. A key barrier to understanding is the inherent non-convexity of the expected return landscape with respect to standard policy parametrizations. As a result, little has been known about the global convergence behavior of policy gradient ascent. Recently, important new progress in understanding the convergence behavior of policy gradient has been achieved in the tabular setting. Although tabular RL is an extremely simplified scenario, it has often provided a necessary first step to understanding deeper questions about RL algorithms. In particular, Bhandari & Russo (2019) have shown that, without parametrization, projected gradient ascent on the simplex does not suffer from spurious local optima. Subsequently, Agarwal et al. (2019) contributed further progresses by showing that (1) without parametrization, projected gradient ascent converges at rate $O(1/\sqrt{t})$ to a global optimum; and (2) with softmax parametrization, policy gradient converges asymptotically. Agarwal et al. (2019) also analyze other variants of policy gradient, and show that policy gradient with relative entropy converges at rate $O(1/\sqrt{t})$, natural policy gradient (mirror descent) converges at rate $O(1/t)$, and given a "compatible" function approximation natural policy gradient converges at rate $O(1/\sqrt{t})$.

Despite these recent advances, many open questions remain in understanding the behavior of policy gradient methods, even in the tabular setting and even with the true gradient. In this paper, we consider the following three open questions raised by the current work in this area. (1) The convergence rate of policy gradient methods with softmax parametrization was previously unknown. The best previous result, due to Agarwal et al. (2019), established asymptotic convergence without any characterization of rate. (2) The convergence rate of entropy regularized softmax policy gra-

Work done as an intern at Google Research, Brain Team. Correspondence to: Jincheng Mei <jmei2@ualberta.ca>.

dient methods was also unknown, and explicitly stated as an open problem in Agarwal et al. (2019). (3) It was not previously understood, theoretically, why entropy helps policy optimization. There have been recent empirical studies that provide suggestive observations (Ahmed et al., 2019), but a theoretical explanation has been missing.

In this paper, we provide answers to these three stated open questions.

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: (1) that the softmax policy gradient satisfies a Łojasiewicz-type inequality but with dependence on the optimal action probability under the current policy; (2) 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 ascent 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 softmax optimal policy, which is significantly faster than vanilla softmax policy gradient. This result resolves an open question in Agarwal et al. (2019, Remark 5.5), where the authors analyzed a more aggressive relative entropy regularization rather than the more common entropy regularization. A novel insight is that the entropy regularized gradient update behaves 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. (1) We prove a new lower bound of $\Omega(1/t)$ for softmax policy gradient. This means the upper bound of $O(1/t)$ for softmax policy gradient we establish is optimal up to constant factors. 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 ascent without regularization*. (2) We study the concept of non-uniform Łojasiewicz degree and show that, without regularization, the Łojasiewicz degree of expected reward cannot be positive, which only allows $O(1/t)$ rates to be established. We then show that after adding 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 Łojasiewicz degree and 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 $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, state value of π is defined as

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

We also let $V^\pi(\rho) \coloneqq \mathbb{E}_{s\sim\rho}\left[V^\pi(s)\right]$, 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

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

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

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

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

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

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

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

The following softmax transform can extract a corresponding probability distribution from any given vector.

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

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

Due to its origin in logistic regression, we call the values of θ the *logit* values and the function 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 $\pi_\theta := \text{softmax}(\theta)$ using $\pi_\theta(a) = \exp\{\theta(a)\} / \sum_{a'} \exp\{\theta(a')\}$ $(a \in [K])$.

H matrix. Given any distribution π over $[K]$, let $H(\pi) := \text{diag}(\pi) - \pi\pi^\top \in \mathbb{R}^{K \times K}$, where $\text{diag}(\pi) \in \mathbb{R}^{K \times K}$ is the diagonal matrix that has π at its diagonal. The H matrix will play a central role in our analysis, e.g., in Section 4, because for $\pi_\theta := \text{softmax}(\theta)$ and $\theta \in \mathbb{R}^{[K]}$, $H(\pi_\theta)$ is the Jacobian of the $\theta \mapsto \pi_\theta$ map:

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

Finally, we recall the definition of smoothness from convex analysis:

Smoothness. A function $f : \Theta \to \mathbb{R}$ 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-based RL methods usually represent policy as parametric functions, and employ different update rules to do policy improvement in parameter spaces. 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). Policy-based RL methods require gradient information of parameters. The policy gradient theorem expresses the gradient in a convenient form, which we will need:

Theorem 1 (Policy gradient theorem (Sutton et al., 2000))**.** *Suppose $\theta \mapsto \pi_\theta(a|s)$ is differentiable w.r.t. θ, $\forall (s, a)$,*

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

where $\mu \in \Delta(\mathcal{S})$ is an initial state distribution.

3.1. Vanilla Softmax Policy Gradient

In this paper 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 $\pi_\theta(\cdot|s) := \text{softmax}(\theta(s, \cdot))$. The vanilla form of policy gradient for this case is shown in Algorithm 1.

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

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
```

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 the space, 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 state 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 (when the MDP has a single state and $\gamma = 0$) so that we can better focus on explaining the main ideas underlying our result.

3.2. Convergence Rate

3.2.1. One State: An Illustrative Case

We illustrate main idea using MDPs with one state, K actions and discount $\gamma = 0$ (i.e., a bandit problem). 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 softmax parametrization $\pi_\theta := \text{softmax}(\theta)$, even in this simple setting, the objective is non-concave in θ, as can be shown by means of a simple example:

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

The one-state case allows some simplifications which are worth describing. In particular, here Lemma 1 simplifies to

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

Applying the above gradient in Algorithm 1 for any π_{θ_t}, we have the following update rule:

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. By some calculations, we show that the expected reward objective in Eq. (9) is β-smooth with $\beta \le 5/2$:

Lemma 2 (Smoothness)**.** *$\forall r \in [0,1]^K$, $\pi_\theta^\top r$ 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-convex objectives, the next best thing to guarantee convergence to global optima is to establish that the gradient of the objective at any parameter of interest dominates the sub-optimality of the parameter. Inequalities of this form are known as a Łojasiewicz inequality (Łojasiewicz, 1963). 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. We will comment on how to lift this assumption later.

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

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

where $a^* \coloneqq \arg\max_{a \in [K]} r(a)$ is the optimal action.

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

Nevertheless, this already suffices to get an intermediate result, which we state next. The proof of this result combines smoothness and the Łojasiewicz inequality we derived.

Lemma 4 (Pseudo-rate)**.** *Using Update 1 with $\eta = 2/5$,*

$$(\pi^* - \pi_{\theta_t})^\top r \le 5/(t \cdot c_t^2), \tag{12}$$

for all $t > 0$, where $c_t \coloneqq \min_{1 \le s \le t} \pi_{\theta_s}(a^*) > 0$, also

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

Remark 1. *$\pi_{\theta_t}(a^*)$ can be small at initialization and during optimization. Consequently, its minimum c_t can be quite small, and the upper bound in Lemma 4 can be large, or even vacuous. The dependence on $\pi_{\theta_t}(a^*)$ is from Lemma 3. We show that it is impossible to eliminate or improve dependence on $\pi_\theta(a^*)$ in this result. Consider $r = (5,4,4)^\top$, $\pi_\theta = (2\epsilon, 1/2 - 2\epsilon, 1/2)$ where $\epsilon > 0$ is small number. By calculation, $(\pi^* - \pi_\theta)^\top r = 1 - 2\epsilon > 1/2$, $\frac{d\pi_\theta^\top r}{d\theta} = (2\epsilon - 4\epsilon^2, -\epsilon + 4\epsilon^2, -\epsilon)^\top$, $\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 = \epsilon \cdot \sqrt{6 - 24\epsilon + 32\epsilon^2} \le 3\epsilon$. Hence, for any constant $C > 0$,*

$$C \cdot (\pi^* - \pi_\theta)^\top r > C/2 > 3\epsilon \ge \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2, \tag{13}$$

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 not sufficient to claim a true $O(1/t)$ rate just by Lemma 4, since it is still unclear whether c_t can be a function of t. Our next result eliminates this possibility. In particular, this follows by recalling the asymptotic convergence result of Agarwal et al. (2019) that 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 immediately 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 $t_0 > 0$ 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:

Theorem 2 (Arbitrary initialization)**.** *Using Update 1 with $\eta = 2/5$, for $t > 0$,*

$$(\pi^* - \pi_{\theta_t})^\top r \le C/t, \tag{14}$$

where $1/C = \left[\inf_{t \ge 1} \pi_{\theta_t}(a^*)\right]^2 > 0$ 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 the uniform distribution, with which we can strengthen the previous result by showing that $\inf_{t \ge 1} \pi_{\theta_t}(a^*) \ge 1/K$, leading to the promised strengthening of the asymptotic convergence results of Agarwal et al. (2019):

Theorem 3 (Uniform initialization)**.** *Using Update 1 with $\eta = 2/5$ and $\pi_{\theta_1}(a) = 1/K$, $\forall a$, for all $t > 0$,*

$$(\pi^* - \pi_{\theta_t})^\top r \le 5K^2/t,$$

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

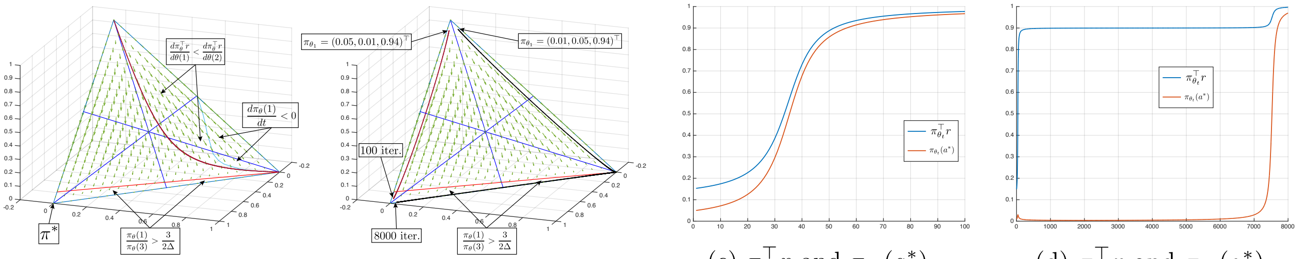
Remark 2. *In Section 5, we prove a lower bound $\Omega(1/t)$ for the same update rule, showing that the upper bound $O(1/t)$ of Theorem 2, apart from constant factors, is unimprovable.*

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

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

$$t_0 \coloneqq \min_{t \ge 1} \left\{ \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} \ge \frac{3}{2\Delta} \right\}. \tag{15}$$

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



(a) Softmax gradient flow.

(b) Good & bad initializations.

(c) $\pi_{\theta_t}^\top r$ and $\pi_{\theta_t}(a^*)$ of good initialization.

(d) $\pi_{\theta_t}^\top r$ and $\pi_{\theta_t}(a^*)$ of bad initialization.

Figure 1. Visualization of proof idea for Lemma 5.

Visualization. Let $r = (1.0, 0.9, 0.1)^\top$. In Fig. 1(a), the region $\{\pi_\theta : \pi_\theta(1)/\pi_\theta(3) \ge 3/(2\Delta)\}$ is the one that is below the red line. Any globally convergent iteration will enter this region within finite time (it contains π^* in its closure) and never goes out (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 toward the optimal policy: 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! This is a typical behavior of non-convex optimization, and it verifies our theoretical findings. 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.

Non-unique optimal actions When the optimal action is not unique, the earlier statements remain valid. However, 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.

3.2.2. General MDPs

For general MDPs, the optimization problem takes the form

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

According to Assumption 1, $r(s, a) \in [0, 1]$, $Q(s, a) \in [0, 1/(1-\gamma)]$, and hence the smoothness property still holds, as also shown by Agarwal et al. (2019).

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

As shown before, smoothness and the Łojasiewicz inequality are sufficient to prove a convergence rate. As noted by Agarwal et al. (2019), the main difficulty is to establish a Łojasiewicz inequality for softmax parametrization. In Lemma 3, we showed that a non-uniform Łojasiewicz inequality holds as in the one-state cases. Fortunately, Lemma 3 can be generalized to general MDPs, under the exploration assumption considered by Agarwal et al. (2019):

Lemma 8 (Non-uniform Łojasiewicz)**.** *Suppose $\mu(s) > 0$ for all state s. Then,*

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

where $a^*(s) \coloneqq \arg\max_a \pi^*(a|s)$, $s \in \mathcal{S}$.

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

Lemma 9. $\inf_{s \in \mathcal{S}, t \ge 1} \pi_{\theta_t}(a^*(s)|s) > 0$.

Using Lemmas 7 to 9, we prove that softmax policy gradient converges to optimal policy with rate $O(1/t)$ in MDPs:

Theorem 4. *Suppose $\mu(s) > 0$ for all state s. Using Algorithm 1 with $\eta = (1-\gamma)^3/8$ and $\pi_{\theta_1}(a^*(s)|s) \in \Omega(1)$ for every $s \in \mathcal{S}$, with some constant $C > 0$, for all $t > 0$,*

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

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 $O\Big(\frac{SC}{(1-\gamma)^6 \epsilon} \cdot \Big\| \frac{d_\mu^{\pi^*}}{\mu} \Big\|_\infty^2 \cdot \Big\| \frac{1}{\mu} \Big\|_\infty \Big)$, 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 ($O\Big(\frac{S^2 A^2}{(1-\gamma)^6 \epsilon^2} \cdot \Big\| \frac{d_\rho^{\pi^*}}{\mu} \Big\|_\infty^2 \Big)$). Our better dependence on ϵ (or t) results from Lemmas 8 and 9 and a different proof technique utilized in Theorem 4.*

4. Entropy Regularized Policy Gradient

It Agarwal et al. (2019, Remark 5.5), an "aggressive" relative-entropy regularization is considered to get finite

$O(1/\sqrt{t})$ rate and Agarwal et al. (2019) pose as an open problem whether the more common entropy regularization also enjoys a similar rate. In this section, we resolve this open question. Surprisingly, we show that entropy regularization improves policy gradient significantly, achieving linear $O(e^{-t})$ convergence rate toward softmax optimal policy. In retrospect, perhaps this is not surprising as adding strongly convex regularizers is a well known technique in convex optimization to improve convergence of first-order methods (Nesterov, 2018, Chapter 2).

4.1. Maximum Entropy RL

Entropy regularization has been widely used in RL objectives (Mnih et al., 2016; Nachum et al., 2017; Haarnoja et al., 2018; Mei et al., 2019). The idea here is to regularize the values as follows:

$$\tilde{V}^{\pi}(\rho) \coloneqq V^{\pi}(\rho) + \tau \cdot \mathbb{H}(\rho, \pi). \tag{16}$$

Here, $\tau \ge 0$, the "temperature", determines the strength of regularization, and $\mathbb{H}(\rho, \pi)$ is the "discounted entropy", as defined by Nachum et al. (2017):

$$\mathbb{H}(\rho, \pi) \coloneqq \underset{\substack{s_0 \sim \rho, a_t \sim \pi(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}}{\mathbb{E}} \left[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t|s_t) \right]. \tag{17}$$

Similar with Lemma 1, one can obtain the following expression for the gradient of the entropy regularized with the softmax policy parametrization:

Lemma 10. *It holds that*

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s, a)} = \frac{1}{1-\gamma} \cdot d_{\mu}^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot \tilde{A}^{\pi_\theta}(s, a), \tag{18}$$

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

$$\tilde{A}^{\pi_\theta}(s, a) \coloneqq \tilde{Q}^{\pi_\theta}(s, a) - \tau \log \pi_\theta(a|s) - \tilde{V}^{\pi_\theta}(s), \tag{19}$$

$$\tilde{Q}^{\pi_\theta}(s, a) \coloneqq r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) \tilde{V}^{\pi_\theta}(s'). \tag{20}$$

4.2. Convergence Rate

As in the non-regularized case, we first consider the one-state (bandit) case to gain some insight.

4.2.1. One-State Case

In the one-state case, Eq. (16) reduces to maximizing the entropy-regularized reward,

$$\max_{\theta: \mathcal{A} \to \mathbb{R}} \underset{a \sim \pi_\theta}{\mathbb{E}} [r(a) - \tau \log \pi_\theta(a)]. \tag{21}$$

Again, Eq. (21) is a non-concave function of θ. In this case, regularized policy gradient reduces to

$$\frac{d\{\pi_\theta^\top (r - \tau \log \pi_\theta)\}}{d\theta} = H(\pi_\theta)(r - \tau \log \pi_\theta), \tag{22}$$

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 disapparing as $\tau \to 0$:

Softmax optimal policy. $\pi_\tau^* \coloneqq \mathrm{softmax}(r/\tau)$ is the optimal solution of Eq. (21).

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 analyzes the following update rule.

Update 3. $\tilde{\theta}_{t+1} \leftarrow \tilde{\theta}_t + \eta \cdot H(\pi_{\tilde{\theta}_t})(r - \tau \log \pi_{\tilde{\theta}_t}) - \frac{\tilde{\theta}_t^\top \mathbf{1}}{K} \cdot \mathbf{1}$.

Updates 2 and 3 are equivalent in the sense that they provide the same softmax policy sequence.

Lemma 11. *If $\theta_1 = \tilde{\theta}_1 + c \cdot \mathbf{1}$ for some constant $c \in \mathbb{R}$, then $\pi_{\theta_t} = \pi_{\tilde{\theta}_t}$, $\forall t \ge 1$.*

As it turns out, in the case of Update 3, the update behaves like a contraction operator in value learning, but with a contraction factor that depends on the current policy.

Lemma 12 (Non-uniform contraction)**.** *Using Update 3 with $\tau\eta \le 1$, $\forall t > 0$,*

$$\|\zeta_{t+1}\|_2 \le \left(1 - \tau\eta \cdot \min_a \pi_{\tilde{\theta}_t}(a)\right) \cdot \|\zeta_t\|_2, \tag{23}$$

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

This lemma immediately implies the following bound:

Lemma 13. *Using Update 3 with $\tau\eta \le 1$, $\forall t > 0$,*

$$\|\zeta_t\|_2 \le \frac{2(\tau C + 1)\sqrt{K}}{\exp\left\{\tau\eta \sum_{s=1}^{t-1} [\min_a \pi_{\tilde{\theta}_s}(a)]\right\}}, \tag{24}$$

where we assume $\|\tilde{\theta}_1\|_\infty \le C$ for some constant $C > 0$.

Similar with Lemma 5, we show that minimum action probability can be lower bounded by its initial value.

Lemma 14. $\min_a \pi_{\tilde{\theta}_t}(a) / \min_a \pi_{\tilde{\theta}_1}(a) \in \Omega(1)$*, for $t > 0$. Thus, $\sum_{s=1}^{t-1} [\min_a \pi_{\tilde{\theta}_s}(a)] \in \Omega(t)$.*

Essentially, what happens is $\min_a \pi_{\tilde{\theta}_t}(a) \to \min_a \pi_\tau^*(a) > 0$, where the latter inequality holds thanks to $r \in [0, 1]^K$ and $\tau > 0$. With Lemmas 11, 13 and 14, we prove that entropy regularized softmax policy gradient enjoys a linear convergence rate:

Theorem 5. *Using Update 2 with $\eta \le 1/\tau$, for all $t > 0$,*

$$\tilde{\delta}_t \le \frac{2(\tau C + 1)^2 K/\tau}{\exp\left\{2\tau\eta \cdot \Omega(1) \cdot t\right\}}, \tag{25}$$

where $\tilde{\delta}_t := {\pi_\tau^*}^\top (r - \tau \log \pi_\tau^*) - {\pi_{\theta_t}}^\top (r - \tau \log \pi_{\theta_t})$.

4.2.2. General MDPs

For general MDPs, the problem is to maximize $\tilde{V}^{\pi_\theta}(\rho)$ in Eq. (16). The softmax optimal policy π_τ^* is known to satisfy the following consistency conditions (Nachum et al., 2017):

$$\pi_\tau^*(a|s) = \exp\left\{(\tilde{Q}^{\pi_\tau^*}(s,a) - \tilde{V}^{\pi_\tau^*}(s))/\tau\right\}, \tag{26}$$

$$\tilde{V}^{\pi_\tau^*}(s) = \tau \log \sum_a \exp\left\{\tilde{Q}^{\pi_\tau^*}(s,a)/\tau\right\}. \tag{27}$$

Using a somewhat lengthy calculation, we show that the discounted entropy in Eq. (17) is smooth:

Lemma 15 (Smoothness)**.** *$\mathbb{H}(\rho, \pi_\theta)$ is $(4 + 8 \log A)/(1-\gamma)^3$-smooth, where $A := |\mathcal{A}|$ 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 16 (Non-uniform Łojasiewicz)**.** *Suppose $\mu(s) > 0$ for all state $s \in \mathcal{S}$. Then,*

$$\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \ge C(\theta) \cdot \left[\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho)\right]^{\frac{1}{2}}, \tag{28}$$

where

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

The entropy regularization helps prove the following result:

Lemma 17. *Using Algorithm 1 with soft policy gradient Eq. (18), we have $\inf_{t \ge 1} \min_{s,a} \pi_{\theta_t}(a|s) > 0$.*

With Lemmas 15 to 17, 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 entropy regularized softmax policy gradient Eq. (18), $\eta = (1-\gamma)^3/(8 + \tau(4 + 8\log A))$ and $\pi_{\theta_1}(a|s) \in \Omega(1)$, $\forall (s,a)$,*

$$\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) \le \frac{\left\|1/\mu\right\|_\infty}{\exp\left\{C_\tau \cdot \Omega(1) \cdot t\right\}} \cdot \frac{1 + \tau \log A}{(1-\gamma)^2},$$

for all $t > 0$, where C_τ, $\Omega(1) > 0$ are independent with t.

4.2.3. Biased Softmax Optimal Policy

As noted in Remark 4, π_τ^* is biased, i.e., $\pi_\tau^* \ne \pi^*$ for fixed $\tau > 0$. We discuss two usual ways to deal with this issue.

Two-stage. Note $\pi_\tau^*(a^*) \ge \pi_\tau^*(a)$, $\forall a$, for any $\tau > 0$. Therefore, using policy gradient with $\pi_{\theta_1} = \pi_\tau^*$, we have $\pi_{\theta_t}(a^*) \ge c_t \ge 1/K$. This suggests a two-stage method: first, use entropy-regularized policy gradient for constant $O(\log(\tau/\Delta))$ iterations (to make $\pi_{\theta_t}(a^*) \ge \pi_{\theta_t}(a)$, $\forall a$); second, use vanilla policy gradient. The convergence rate is

$$(\pi^* - \pi_{\theta_t})^\top r \le 5/(C^2 \cdot t_2), \tag{29}$$

where $t_1 + t_2 = t$, $t_1 \in O(\log(\tau/\Delta))$, t_2 is time of second stage, and $C \in [1/K, 1)$. The initialization dependent c_t no longer exists, which is better than $5/(c_t^2 \cdot t)$ as in Lemma 4.

Decayed entropy. Another usual way is to decay the regularization, e.g., $\tau_t := 1/t$. Consider the following update.

Update 4. $\theta_{t+1} \leftarrow \frac{\tau_t}{\tau_{t+1}} \cdot (\theta_t + \eta \cdot H(\pi_{\theta_t})(r - \tau_t \log \pi_{\theta_t}))$.

There is also an alternative update that Lemma 12 holds but with $\zeta_t := \tau_t \tilde{\theta}_t - r - \frac{(\tau_t \tilde{\theta}_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}$. However, as $\tau_t \to 0$, $\min_a \pi_{\tilde{\theta}_t}(a) \to 0$, which means $\min_a \pi_{\tilde{\theta}_t}(a) \notin \Omega(1)$ as $\pi_{\tilde{\theta}_t} \to \pi^*$. We conjecture the rate degenerates to $O(1/t)$.

5. A Theoretical Understanding of Entropy Regularization in Policy Gradient Methods

Ahmed et al. (2019) perform an empirical study to explain the impact of entropy in policy optimization by introducing a loss perturbation method to facilitate visualization of optimization landscape. They qualitatively show that in *some* tasks, policies with higher entropy have smoother landscapes, and they possibly enjoy better optimization properties. However, this leaves open the question of whether the conclusions of this empirical study hold in general.

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 $O(1/t)$ we established earlier for policy gradient and softmax parametrization when entropy regularization is not used cannot be improved. Next, we introduce the notion of Łojasiewicz degree, which we show to increase in the presence of entropy regularization, which, we connect 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 fails to provide a satisfactory solution to the exploration problem as explained by O'Donoghue et al. (2020).

5.1. Lower Bounds

Sections 3 and 4 show that sofmax policy gradient converges with a rate $O(1/t)$, while entropy regularized softmax policy gradient has a faster $O(e^{-t})$ rate. To claim that regularization makes policy gradient converge faster, we need a lower bound for softmax policy gradient.

Intuitively, smoothness and the Łojasiewicz inequality of Lemma 3 together guarantee enough progress in each iteration toward a global optimum. To get lower bounds, we need to show that progress in every iteration cannot be too large. As it turns out, the expected reward satisfies a reverse Łojasiewicz inequality, with a problem-dependent constant:

Lemma 18 (Reverse Łojasiewicz). *Denote $\Delta := r(a^*) - \max_{a \neq a^*} r(a) > 0$ as the reward gap of r. Then,*

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \le \frac{\sqrt{2}}{\Delta} \cdot (\pi^* - \pi_\theta)^\top r. \tag{30}$$

Using smoothness, one can then upper bound the progress of one step gradient update by the squared norm of the gradient. Then, by Lemma 18, this progress is further upper bounded by the square of the current sub-optimality. From this an elementary calculation gives the desired lower bound:

Theorem 7 (Lower bound). *For large enough $t > 0$, using Update 1 with learning rate $\eta \in (0, 1]$,*

$$(\pi^* - \pi_{\theta_t})^\top r \ge \frac{\Delta^2}{6\eta \cdot t}. \tag{31}$$

Note that Theorem 7 is a special case of general MDPs. Therefore, the $\Omega(1/t)$ lower bound also holds for MDPs:

Theorem 8 (Lower bound). *For large enough $t > 0$, using softmax policy gradient Algorithm 1 with $\eta \in (0, 1]$,*

$$V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \ge \frac{(1-\gamma)^5 \cdot (\Delta^*)^2}{12\eta \cdot t}, \tag{32}$$

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 7, which is from Lemma 18 and happens when π_θ is in vicinity of π^*. Since any globally convergent iteration must be in vicinity of the optimal policy π^* after large enough time, this large constant gap seems unavoidable. We leave the improvement of constant difference as an open problem.*

With the lower bound established, we can confirm that entropy regularization helps policy optimization by speeding up convergence.

5.2. Non-uniform Łojasiewicz Degree

The discrepancy between convergence rates can explain the advantage of entropy regularization, but the difference itself is a result of a deeper reason. We investigate this point through lens of Łojasiewicz degree, which is a key property that is related to rates of non-convex optimization.

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

$$\left\| \nabla_x f(x) \right\|_2 \ge C(x) \cdot \left| f(x) - f(x^*) \right|^{1-\xi}, \tag{33}$$

$\forall x \in \mathcal{X}$, where $C(x) > 0$ can be independent with x.

Łojasiewicz degree is closely related to convergence rates of first- (Bárta, 2017) and second-order methods (Nesterov & Polyak, 2006; Zhou et al., 2018) in non-convex optimization. Large Łojasiewicz degree corresponds to faster convergence rate for the same optimization method.

First, we show that the Łojasiewicz degree of the expected reward objective cannot be positive:

Proposition 3. *The Łojasiewicz degree of $\mathbb{E}_{a \sim \pi_\theta}[r(a)]$ cannot be larger than 0 with $C(\theta) = \pi_\theta(a^*)$.*

Note that according to Remark 1, it is necessary that $C(\theta)$ depends on $\pi_\theta(a^*)$. The difference between Proposition 3 and the reverse Łojasiewicz inequality of Lemma 18 is subtle. Lemma 18 is a condition that implies impossibility to get rates faster than $O(1/t)$, while Proposition 3 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 becomes $1/2$:

Proposition 4. *With $C(\theta) = \sqrt{2\tau} \cdot \min_a \pi_\theta(a)$, the Łojasiewicz degree of $\mathbb{E}_{a \sim \pi_\theta}[r(a) - \tau \log \pi_\theta(a)]$ is $1/2$.*

We postulate that the increase of Łojasiewicz degree is the key to the faster convergence rate of policy gradient when used on the entropy-regularized objective.

6. Conclusions and Future Work

We show matching bounds $O(1/t)$ and $\Omega(1/t)$ for the tabular setting of softmax policy gradient methods, which is a faster rate than those obtained for closely related policy gradient methods in previous work. Important directions for future work include the generalization of our results to the case when the gradient needs to be estimated and/or a function approximator is used to represent policies. Moreover, it may also be interesting to find new uses for non-uniform Łojasiewicz inequalities in non-convex optimization and for the notion of Łojasiewicz degree.

[1] Note that 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 discussion and for providing feedback on a draft of this manuscript. Jincheng Mei would like to thank Ruitong Huang for enlightening early discussions.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. *arXiv preprint arXiv:1908.00261*, 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. *arXiv preprint arXiv:1906.01786*, 2019.

Golub, G. H. Some modified matrix eigenvalue problems. *SIAM Review*, 15(2):318–334, 1973.

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. and Langford, J. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pp. 267–274, 2002.

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.

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.

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

Xiao, C., Huang, R., Mei, J., Schuurmans, D., and Müller, M. Maximum entropy monte-carlo planning. In *Advances in Neural Information Processing Systems*, pp. 9516–9524, 2019.

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.

The appendix is organized as follows.

- Appendix A: proofs for the technical results in the main paper.
  - Appendix A.1: proofs for the results of softmax policy gradient in Section 3.
  - Appendix A.2: proofs for the results of entropy regularized softmax policy gradient in Section 4.
  - Appendix A.3: proofs for the results of theoretical understanding of entropy in Section 5.
- Appendix B: supporting lemmas which are not presented in the main paper.
- Appendix C: remarks on sub-optimality guarantees for other entropy-based RL methods, which are not presented in the main paper.
- Appendix D: simulation results to verify the convergence rates.

A. Proofs

A.1. Proofs for Section 3

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

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot A^{\pi_\theta}(s,a). \tag{34}$$

Proof. See Agarwal et al. (2019, Lemma C.1). Our proof is for completeness. According to Theorem 1,

$$\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s'\sim d_\mu^{\pi_\theta}} \left[\sum_a \frac{\partial \pi_\theta(a|s')}{\partial \theta} \cdot Q^{\pi_\theta}(s',a) \right]. \tag{35}$$

For $s' \neq s$, $\frac{\partial \pi_\theta(a|s')}{\partial \theta(s,\cdot)} = \mathbf{0}$ since $\pi_\theta(a|s')$ does not depend on $\theta(s,\cdot)$. Therefore,

$$\begin{aligned}
\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} &= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \left[\sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)} \cdot Q^{\pi_\theta}(s,a) \right] && (36)\\
&= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \left(\frac{d \pi(\cdot|s)}{d \theta(s,\cdot)} \right)^\top Q^{\pi_\theta}(s,\cdot) && (37)\\
&= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(\cdot|s)) Q^{\pi_\theta}(s,\cdot). \qquad \text{(Eq. (6))} && (38)
\end{aligned}$$

According to $H(\pi_\theta(\cdot|s)) = \text{diag}(\pi_\theta(\cdot|s)) - \pi_\theta(\cdot|s)\pi_\theta(\cdot|s)^\top$, for each component a, we have

$$\begin{aligned}
\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)} &= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot \left[Q^{\pi_\theta}(s,a) - \sum_a \pi_\theta(a|s) \cdot Q^{\pi_\theta}(s,a) \right] && (39)\\
&= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot \left[Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s) \right] \qquad \left(V^{\pi_\theta}(s) = \sum_a \pi_\theta(a|s) \cdot Q^{\pi_\theta}(s,a) \right) && (40)\\
&= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot A^{\pi_\theta}(s,a).
\end{aligned}$$

□

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

Proof. Consider the following example: $r = (1, 9/10, 1/10)^\top$, $\theta_1 = (0,0,0)^\top$, $\pi_{\theta_1} = \text{softmax}(\theta_1) = (1/3, 1/3, 1/3)^\top$, $\theta_2 = (\ln 9, \ln 16, \ln 25)^\top$, and $\pi_{\theta_2} = \text{softmax}(\theta_2) = (9/50, 16/50, 25/50)^\top$. We have,

$$\frac{1}{2} \cdot \left(\pi_{\theta_1}^\top r + \pi_{\theta_2}^\top r \right) = \frac{1}{2} \cdot \left(\frac{2}{3} + \frac{259}{500} \right) = \frac{1777}{3000} = \frac{14216}{24000}. \tag{41}$$

On the other hand, $\bar{\theta} \coloneqq \frac{1}{2} \cdot (\theta_1 + \theta_2) = (\ln 3, \ln 4, \ln 5)^\top$, and $\pi_{\bar{\theta}} = \mathrm{softmax}(\bar{\theta}) = (3/12, 4/12, 5/12)^\top$.

$$\pi_{\bar{\theta}}^\top r = \frac{71}{120} = \frac{14200}{24000}. \tag{42}$$

Since $\frac{1}{2} \cdot \left(\pi_{\theta_1}^\top r + \pi_{\theta_2}^\top r \right) > \pi_{\bar{\theta}}^\top r$, $\mathbb{E}_{a \sim \pi_\theta(\cdot)} \left[r(a) \right]$ is a non-concave function of θ. □

Lemma 2 (Smoothness)**.** Let $\pi_\theta \coloneqq \mathrm{softmax}(\theta)$ and $\pi_{\theta'} \coloneqq \mathrm{softmax}(\theta')$. $\forall r \in [0,1]^K$,

$$\left| (\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d \pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle \right| \le \frac{5}{4} \cdot \| \theta' - \theta \|_2^2. \tag{43}$$

Proof. Denote the second derivative w.r.t. θ (i.e., Hessian) as

$$\begin{aligned} S(r, \theta) &\coloneqq \frac{d}{d\theta} \left\{ \frac{d \pi_\theta^\top r}{d\theta} \right\} && (44) \\ &= \frac{d}{d\theta} \left\{ \left(\frac{d \pi_\theta}{d\theta} \right)^\top \left(\frac{d \pi_\theta^\top r}{d \pi_\theta} \right) \right\} && (45) \\ &= \frac{d}{d\theta} \left\{ H(\pi_\theta) r \right\} \qquad \text{(Eq. (6))} && (46) \\ &= \frac{d}{d\theta} \left\{ (\mathrm{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top) r \right\}. && (47) \end{aligned}$$

Note that $S(r, \theta) \in \mathbb{R}^{K \times K}$, and $\forall i, j \in [K]$, the value of $S(r, \theta)$ is,

$$\begin{aligned} S_{i,j} &= \frac{d \{ \pi_\theta(i) \cdot (r(i) - \pi_\theta^\top r) \}}{d \theta(j)} && (48) \\ &= \frac{d \pi_\theta(i)}{d\theta(j)} \cdot (r(i) - \pi_\theta^\top r) + \pi_\theta(i) \cdot \frac{d \{ r(i) - \pi_\theta^\top r \}}{d \theta(j)} && (49) \\ &= (\delta_{ij} \pi_\theta(j) - \pi_\theta(i) \pi_\theta(j)) \cdot (r(i) - \pi_\theta^\top r) - \pi_\theta(i) \cdot (\pi_\theta(j) r(j) - \pi_\theta(j) \pi_\theta^\top r) && (50) \\ &= \delta_{ij} \pi_\theta(j) \cdot (r(i) - \pi_\theta^\top r) - \pi_\theta(i) \pi_\theta(j) \cdot (r(i) - \pi_\theta^\top r) - \pi_\theta(i) \pi_\theta(j) \cdot (r(j) - \pi_\theta^\top r), && (51) \end{aligned}$$

where

$$\delta_{ij} \coloneqq \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases} \tag{52}$$

We show that the spectral radius of $S(r, \theta)$ is smaller than or equal to $5/2$. For any $y \in \mathbb{R}^K$,

$$\begin{aligned} \left| y^\top S(r, \theta) y \right| &= \left| \sum_{i=1}^{K} \sum_{j=1}^{K} S_{i,j} y(i) y(j) \right| && (53) \\ &= \left| \sum_i \pi_\theta(i) (r(i) - \pi_\theta^\top r) y(i)^2 - 2 \sum_i \pi_\theta(i) (r(i) - \pi_\theta^\top r) y(i) \sum_j \pi_\theta(j) y(j) \right| && (54) \\ &= \left| (H(\pi_\theta) r)^\top (y \odot y) - 2 \cdot (H(\pi_\theta) r)^\top y \cdot \left(\pi_\theta^\top y \right) \right| && (55) \\ &\le \| H(\pi_\theta) r \|_\infty \cdot \| y \odot y \|_1 + 2 \cdot \| H(\pi_\theta) r \|_1 \cdot \| y \|_\infty \cdot \| \pi_\theta \|_1 \cdot \| y \|_\infty, && (56) \end{aligned}$$

where $\odot$ is Hadamard (component-wise) product, and the last inequality is by triangle inequality and Hölder's inequality.

Note that $\|y \odot y\|_1 = \|y\|_2^2$, $\|\pi_\theta\|_1 = 1$, and $\|y\|_\infty \le \|y\|_2$. Denote $H_{i,:}(\pi_\theta)$ as the i-th row vector of $H(\pi_\theta)$, $\forall i$,

$$\begin{aligned}
\left\| H_{i,:}(\pi_\theta) \right\|_1 &= \pi_\theta(i) - \pi_\theta(i)^2 + \pi_\theta(i) \cdot \sum_{j \ne i} \pi_\theta(j) && (57)\\
&= \pi_\theta(i) - \pi_\theta(i)^2 + \pi_\theta(i) \cdot (1 - \pi_\theta(i)) && (58)\\
&= 2 \cdot \pi_\theta(i) \cdot (1 - \pi_\theta(i)) && (59)\\
&\le 1/2. \qquad (x \cdot (1-x) \le 1/4 \text{ for } x \in [0,1]) && (60)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\left\| H(\pi_\theta) r \right\|_1 &= \sum_i \pi_\theta(i) \cdot \left| r(i) - \pi_\theta^\top r \right| && (61)\\
&\le \max_i \left| r(i) - \pi_\theta^\top r \right| && (62)\\
&\le 1. \qquad \left(r \in [0,1]^K \right) && (63)
\end{aligned}$$

Therefore we have,

$$\begin{aligned}
\left| y^\top S(r,\theta) y \right| &\le \left\| H(\pi_\theta) r \right\|_\infty \cdot \|y\|_2^2 + 2 \cdot \left\| H(\pi_\theta) r \right\|_1 \cdot \|y\|_2^2 && (64)\\
&= \max_i \left| \left(H_{i,:}(\pi_\theta) \right)^\top r \right| \cdot \|y\|_2^2 + 2 \cdot \left\| H(\pi_\theta) r \right\|_1 \cdot \|y\|_2^2 && (65)\\
&\le \max_i \left\| H_{i,:}(\pi_\theta) \right\|_1 \cdot \|r\|_\infty \cdot \|y\|_2^2 + 2 \cdot 1 \cdot \|y\|_2^2 && (66)\\
&\le (1/2 + 2) \cdot \|y\|_2^2 = 5/2 \cdot \|y\|_2^2 . && (67)
\end{aligned}$$

Denote $\theta_\xi = \theta + \xi(\theta' - \theta)$, where $\xi \in [0,1]$. According to Taylor's theorem, we have,

$$\begin{aligned}
\left| (\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d\pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle \right| &= \frac{1}{2} \cdot \left| (\theta' - \theta)^\top S(r, \theta_\xi) \, (\theta' - \theta) \right| && (68)\\
&\le \frac{5}{4} \cdot \|\theta' - \theta\|_2^2. && \square
\end{aligned}$$

Lemma 3 (Non-uniform Łojasiewicz)**.** Assume r has one unique optimal action. Let $\pi^* \coloneqq \arg\max_{\pi \in \Delta} \pi^\top r$. $\forall \pi_\theta \coloneqq \mathrm{softmax}(\theta)$,

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \ge \pi_\theta(a^*) \cdot (\pi^* - \pi_\theta)^\top r, \tag{69}$$

where $a^* \coloneqq \arg\max_{a \in [K]} r(a)$ is the optimal action. Also, for non-unique optimal action cases,

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \ge \frac{1}{\sqrt{|\mathcal{A}^*|}} \cdot \left[\sum_{a^* \in \mathcal{A}^*} \pi_\theta(a^*) \right] \cdot (\pi^* - \pi_\theta)^\top r, \tag{70}$$

where $\mathcal{A}^* \coloneqq \{ a^* : r(a^*) = \max_a r(a) \}$ is the optimal action set.

Proof. The claim follows from calculating the ℓ_2 norm of gradient,

$$\begin{aligned}
\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 &= \left(\sum_{a=1}^K \left[\pi_\theta(a) \cdot (r(a) - \pi_\theta^\top r) \right]^2 \right)^{\frac{1}{2}} && (71)\\
&\ge \left(\left[\pi_\theta(a^*) \cdot (r(a^*) - \pi_\theta^\top r) \right]^2 \right)^{\frac{1}{2}} && (72)\\
&= \pi_\theta(a^*) \cdot (\pi^* - \pi_\theta)^\top r. && (73)
\end{aligned}$$

In case of non-unique optimal actions, define the optimal action set

$$\mathcal{A}^* \coloneqq \left\{ a^* : r(a^*) = \max_a r(a) \right\}. \tag{74}$$

The argument holds with $\pi_\theta(a^*)$ replaced with $\sum_{a^* \in \mathcal{A}^*} \pi_\theta(a^*)$.

$$\begin{aligned}
\left\| \frac{d \pi_\theta^\top r}{d \theta} \right\|_2 &\ge \left(\sum_{a^* \in \mathcal{A}^*} \left[\pi_\theta(a^*) \cdot (r(a^*) - \pi_\theta^\top r) \right]^2 \right)^{\frac{1}{2}} \qquad &(75) \\
&\ge \frac{1}{\sqrt{|\mathcal{A}^*|}} \sum_{a^* \in \mathcal{A}^*} \pi_\theta(a^*) \cdot (r(a^*) - \pi_\theta^\top r) \qquad \text{(Cauchy-Schwarz)} \qquad &(76) \\
&= \frac{1}{\sqrt{|\mathcal{A}^*|}} \cdot \left[\sum_{a^* \in \mathcal{A}^*} \pi_\theta(a^*) \right] \cdot (\pi^* - \pi_\theta)^\top r. \qquad &\square
\end{aligned}$$

Lemma 4 (Pseudo-rate)**.** Let $\pi_{\theta_t} \coloneqq \mathrm{softmax}(\theta_t)$. Using Update 1 with $\eta = 2/5$,

$$(\pi^* - \pi_{\theta_t})^\top r \le \frac{5}{c_t^2} \cdot \frac{1}{t}, \tag{77}$$

for all $t > 0$, where $c_t \coloneqq \min_{1 \le s \le t} \pi_{\theta_s}(a^*) > 0$. And

$$\sum_{t=1}^{T} (\pi^* - \pi_{\theta_t})^\top r \le \min \left\{ \frac{\sqrt{5T}}{c_T}, \ \frac{5 \log T}{c_T^2} + 1 \right\}. \tag{78}$$

Proof. According to Lemma 2,

$$\left| (\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d \pi_{\theta_t}^\top r}{d \theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| \le \frac{5}{4} \cdot \| \theta_{t+1} - \theta_t \|_2^2, \tag{79}$$

which implies

$$\begin{aligned}
\pi_{\theta_t}^\top r - \pi_{\theta_{t+1}}^\top r &\le - \left\langle \frac{d \pi_{\theta_t}^\top r}{d \theta_t}, \theta_{t+1} - \theta_t \right\rangle + \frac{5}{4} \cdot \| \theta_{t+1} - \theta_t \|_2^2 \qquad &(80) \\
&= - \eta \cdot \left\| \frac{d \pi_{\theta_t}^\top r}{d \theta_t} \right\|_2^2 + \frac{5}{4} \cdot \eta^2 \cdot \left\| \frac{d \pi_{\theta_t}^\top r}{d \theta_t} \right\|_2^2 \qquad \left(\theta_{t+1} = \theta_t + \eta \cdot \frac{d \pi_{\theta_t}^\top r}{d \theta_t} \right) \qquad &(81) \\
&= - \frac{1}{5} \cdot \left\| \frac{d \pi_{\theta_t}^\top r}{d \theta_t} \right\|_2^2 \qquad (\eta = 2/5) \qquad &(82) \\
&\le - \frac{1}{5} \cdot \left[\pi_{\theta_t}(a^*) \cdot (\pi^* - \pi_{\theta_t})^\top r \right]^2 \qquad \text{(Lemma 3)} \qquad &(83) \\
&\le - \frac{c_t^2}{5} \cdot \left[(\pi^* - \pi_{\theta_t})^\top r \right]^2, \qquad \text{(by the definition of } c_t) \qquad &(84)
\end{aligned}$$

which is equivalent with

$$(\pi^* - \pi_{\theta_{t+1}})^\top r - (\pi^* - \pi_{\theta_t})^\top r \le - \frac{c_t^2}{5} \cdot \left[(\pi^* - \pi_{\theta_t})^\top r \right]^2. \tag{85}$$

Denote $\delta_t \coloneqq (\pi^* - \pi_{\theta_t})^\top r$. We prove $\delta_t \le \frac{5}{c_t^2} \cdot \frac{1}{t}$ by induction on t. For $t = 2$, since $c_2 \in (0, 1)$,

$$\delta_2 \le 1 \le \frac{5}{c_2^2} \cdot \frac{1}{2}. \tag{86}$$

Suppose $\delta_t \le \frac{5}{c_t^2} \cdot \frac{1}{t}$, $t \ge 2$. Consider $f_t : \mathbb{R} \to \mathbb{R}$, $f_t(x) \coloneqq x - \frac{c_t^2}{5} \cdot x^2$. f_t is monotonically increasing in $\left[0, \frac{5}{2 \cdot c_t^2}\right]$.

$$\begin{aligned}
\delta_{t+1} &\le \delta_t - \frac{c_t^2}{5} \cdot \delta_t^2 \qquad \text{(Eq. (85))} && (87)\\
&\le \frac{5}{c_t^2} \cdot \frac{1}{t} - \frac{c_t^2}{5} \cdot \left(\frac{5}{c_t^2} \cdot \frac{1}{t} \right)^2 \qquad \left(\delta_t \le \frac{5}{c_t^2} \cdot \frac{1}{t} \le \frac{5}{2 \cdot c_t^2}, \ t \ge 2 \right) && (88)\\
&= \frac{5}{c_t^2} \cdot \left(\frac{1}{t} - \frac{1}{t^2} \right) && (89)\\
&\le \frac{5}{c_t^2} \cdot \frac{1}{t+1} && (90)\\
&\le \frac{5}{c_{t+1}^2} \cdot \frac{1}{t+1}, \qquad (c_t \ge c_{t+1} > 0) && (91)
\end{aligned}$$

which completes proof for $\delta_t \le \frac{5}{c_t^2} \cdot \frac{1}{t}$. Summing up $\delta_t \le \frac{5}{c_t^2} \cdot \frac{1}{t} \le \frac{5}{c_T^2} \cdot \frac{1}{t}$, we have

$$\sum_{t=1}^{T} (\pi^* - \pi_{\theta_t})^\top r \le \frac{5 \log T}{c_T^2} + 1. \tag{92}$$

On the other hand, rearranging Eq. (85) and summing up $\delta_t^2 \le \frac{5}{c_t^2} \cdot (\delta_t - \delta_{t+1}) \le \frac{5}{c_T^2} \cdot (\delta_t - \delta_{t+1})$ from $t = 1$ to T,

$$\begin{aligned}
\sum_{t=1}^{T} \delta_t^2 &\le \frac{5}{c_T^2} \sum_{t=1}^{T} (\delta_t - \delta_{t+1}) && (93)\\
&= \frac{5}{c_T^2} \cdot (\delta_1 - \delta_{T+1}) && (94)\\
&\le \frac{5}{c_T^2}. \qquad (\text{since } \delta_{T+1} \ge 0, \ \delta_1 \le 1) && (95)
\end{aligned}$$

Finally, we have,

$$\sum_{t=1}^{T} (\pi^* - \pi_{\theta_t})^\top r = \sum_{t=1}^{T} \delta_t \le \sqrt{T} \cdot \sqrt{\sum_{t=1}^{T} \delta_t^2} \le \frac{\sqrt{5T}}{c_T}. \qquad \square$$

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

Proof. In particular, we prove $\inf_{t \ge 1} \pi_{\theta_t}(a^*) = \min_{1 \le t \le t_0} \pi_{\theta_t}(a^*)$, where $t_0 \coloneqq \min\{t : \pi_{\theta_t}(a^*) \ge \frac{c}{c+1}\}$, and $c \coloneqq \frac{K}{2\Delta} \cdot \left(1 - \frac{\Delta}{K}\right)$, and $\Delta \coloneqq r(a^*) - \max_{a \ne a^*} r(a) > 0$ is the reward gap of r. Note that t_0 depends only on θ_1 and c, and c depends only on the problem. Define the following regions,

$$\begin{aligned}
\mathcal{R}_1 &\coloneqq \left\{ \theta : \frac{d\pi_\theta^\top r}{d\theta(a^*)} \ge \frac{d\pi_\theta^\top r}{d\theta(a)}, \ \forall a \ne a^* \right\}, && (96)\\
\mathcal{N}_c &\coloneqq \left\{ \theta : \pi_\theta(a^*) \ge \frac{c}{c+1} \right\}, \text{ where } c \coloneqq \frac{K}{2\Delta} \cdot \left(1 - \frac{\Delta}{K} \right). && (97)
\end{aligned}$$

The main proof idea consists of the following three parts.

- First, we show that $\mathcal{R}_1$ is a "nice" region, in the sense that, following gradient update, (i) if $\theta_t \in \mathcal{R}_1$, then $\theta_{t+1} \in \mathcal{R}_1$; (ii) $\pi_{\theta_{t+1}}(a^*) \ge \pi_{\theta_t}(a^*)$.
- Second, we show that $\mathcal{N}_c \subset \mathcal{R}_1$.
- Third, there exists a finite time $t_0 > 0$, such that $\theta_{t_0} \in \mathcal{N}_c$, and thus $\theta_{t_0} \in \mathcal{R}_1$, which implies $\inf_{t \ge 1} \pi_{\theta_t}(a^*) = \min_{1 \le t \le t_0} \pi_{\theta_t}(a^*)$.

First part. (i) if $\theta_t \in \mathcal{R}_1$, then $\theta_{t+1} \in \mathcal{R}_1$. Suppose $\frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \ge \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)}$ for action a. There are two cases.

(a) If $\pi_{\theta_t}(a^*) \ge \pi_{\theta_t}(a)$, then $\theta_t(a^*) \ge \theta_t(a)$. After one step gradient update,

$$\theta_{t+1}(a^*) \leftarrow \theta_t(a^*) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \tag{98}$$

$$\ge \theta_t(a) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)} \tag{99}$$

$$:= \theta_{t+1}(a), \tag{100}$$

which implies $\pi_{\theta_{t+1}}(a^*) \ge \pi_{\theta_{t+1}}(a)$. Since $r(a^*) - \pi_{\theta_{t+1}}^\top r > 0$ and $r(a^*) > r(a)$,

$$\pi_{\theta_{t+1}}(a^*) \cdot \left[r(a^*) - \pi_{\theta_{t+1}}^\top r \right] \ge \pi_{\theta_{t+1}}(a) \cdot \left[r(a) - \pi_{\theta_{t+1}}^\top r \right], \tag{101}$$

which is equivalent with $\frac{d\pi_{\theta_{t+1}}^\top r}{d\theta_{t+1}(a^*)} \ge \frac{d\pi_{\theta_{t+1}}^\top r}{d\theta_{t+1}(a)}$, i.e., $\theta_{t+1} \in \mathcal{R}_1$.

(b) If $\pi_{\theta_t}(a^*) < \pi_{\theta_t}(a)$, then by $\frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \ge \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)}$,

$$\pi_{\theta_t}(a^*) \cdot \left[r(a^*) - \pi_{\theta_t}^\top r \right] \ge \pi_{\theta_t}(a) \cdot \left[r(a) - \pi_{\theta_t}^\top r \right] \tag{102}$$

$$= \pi_{\theta_t}(a) \cdot \left[r(a^*) - \pi_{\theta_t}^\top r \right] - \pi_{\theta_t}(a) \cdot \left[r(a^*) - r(a) \right], \tag{103}$$

which after rearranging is equivalent with

$$r(a^*) - r(a) \ge \left(1 - \frac{\pi_{\theta_t}(a^*)}{\pi_{\theta_t}(a)} \right) \cdot \left[r(a^*) - \pi_{\theta_t}^\top r \right] \tag{104}$$

$$= \left(1 - \exp\left\{ \theta_t(a^*) - \theta_t(a) \right\} \right) \cdot \left[r(a^*) - \pi_{\theta_t}^\top r \right]. \tag{105}$$

After one step gradient update, according to smoothness argument as in Eq. (80), $\pi_{\theta_{t+1}}^\top r \ge \pi_{\theta_t}^\top r$, i.e.,

$$0 < r(a^*) - \pi_{\theta_{t+1}}^\top r \le r(a^*) - \pi_{\theta_t}^\top r. \tag{106}$$

On the other hand,

$$\theta_{t+1}(a^*) - \theta_{t+1}(a) = \theta_t(a^*) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} - \theta_t(a) - \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)} \tag{107}$$

$$\ge \theta_t(a^*) - \theta_t(a), \tag{108}$$

which implies

$$1 - \exp\left\{ \theta_{t+1}(a^*) - \theta_{t+1}(a) \right\} \le 1 - \exp\left\{ \theta_t(a^*) - \theta_t(a) \right\}. \tag{109}$$

And since $1 - \exp\left\{ \theta_t(a^*) - \theta_t(a) \right\} = 1 - \frac{\pi_{\theta_t}(a^*)}{\pi_{\theta_t}(a)} > 0$ (in this case $\pi_{\theta_t}(a^*) < \pi_{\theta_t}(a)$),

$$\left(1 - \exp\left\{ \theta_{t+1}(a^*) - \theta_{t+1}(a) \right\} \right) \cdot \left[r(a^*) - \pi_{\theta_{t+1}}^\top r \right] \le \left(1 - \exp\left\{ \theta_t(a^*) - \theta_t(a) \right\} \right) \cdot \left[r(a^*) - \pi_{\theta_t}^\top r \right] \tag{110}$$

$$\le r(a^*) - r(a), \tag{111}$$

which is equivalent with

$$\left(1 - \frac{\pi_{\theta_{t+1}}(a^*)}{\pi_{\theta_{t+1}}(a)} \right) \cdot \left[r(a^*) - \pi_{\theta_{t+1}}^\top r \right] \le r(a^*) - r(a). \tag{112}$$

Rearranging the above inequality,

$$\pi_{\theta_{t+1}}(a^*) \cdot \left[r(a^*) - \pi_{\theta_{t+1}}^\top r \right] \ge \pi_{\theta_{t+1}}(a) \cdot \left[r(a) - \pi_{\theta_{t+1}}^\top r \right], \tag{113}$$

which means $\frac{d\pi_{\theta_{t+1}}^\top r}{d\theta_{t+1}(a^*)} \ge \frac{d\pi_{\theta_{t+1}}^\top r}{d\theta_{t+1}(a)}$, i.e, $\theta_{t+1} \in \mathcal{R}_1$. Now we have (i) if $\theta_t \in \mathcal{R}_1$, then $\theta_{t+1} \in \mathcal{R}_1$.

Next we prove (ii) $\pi_{\theta_{t+1}}(a^*) > \pi_{\theta_t}(a^*)$. If $\theta_t \in \mathcal{R}_1$, then $\frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \ge \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)}$, $\forall a \ne a^*$. After one step gradient update,

$$\pi_{\theta_{t+1}}(a^*) = \frac{\exp\left\{\theta_{t+1}(a^*)\right\}}{\sum_a \exp\left\{\theta_{t+1}(a)\right\}} \tag{114}$$

$$= \frac{\exp\left\{\theta_t(a^*) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)}\right\}}{\sum_a \exp\left\{\theta_t(a) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)}\right\}} \tag{115}$$

$$\ge \frac{\exp\left\{\theta_t(a^*) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)}\right\}}{\sum_a \exp\left\{\theta_t(a) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)}\right\}} \qquad \left(\frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \ge \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)}\right) \tag{116}$$

$$= \frac{\exp\left\{\theta_t(a^*)\right\}}{\sum_a \exp\left\{\theta_t(a)\right\}} = \pi_{\theta_t}(a^*). \tag{117}$$

Second part. $\mathcal{N}_c \subset \mathcal{R}_1$. Suppose $\pi_\theta(a^*) \ge \frac{c}{c+1}$. There are two cases.

(a) If $\pi_\theta(a^*) \ge \max_{a \ne a^*}\{\pi_\theta(a)\}$, then $\theta \in \mathcal{R}_2 \subset \mathcal{R}_1$ in the proof for Proposition 2.

(b) If $\pi_\theta(a^*) < \max_{a \ne a^*}\{\pi_\theta(a)\}$, we show that $\frac{d\pi_\theta^\top r}{d\theta(a^*)} \ge \frac{d\pi_\theta^\top r}{d\theta(a)}$, $\forall a \ne a^*$. For any specific a, we re-label the actions for convenience, such that $a^* = 1$, $a = 2$. Then we have,

$$\frac{d\pi_\theta^\top r}{d\theta(a^*)} - \frac{d\pi_\theta^\top r}{d\theta(a)} = \frac{d\pi_\theta^\top r}{d\theta(1)} - \frac{d\pi_\theta^\top r}{d\theta(2)} = \pi_\theta(1) \cdot \left[r(1) - \pi_\theta^\top r\right] - \pi_\theta(2) \cdot \left[r(2) - \pi_\theta^\top r\right] \tag{118}$$

$$= 2\pi_\theta(1) \cdot \left[r(1) - \pi_\theta^\top r\right] + \sum_{i=3}^K \pi_\theta(i) \cdot \left[r(i) - \pi_\theta^\top r\right] \tag{119}$$

$$= \left(2\pi_\theta(1) + \sum_{i=3}^K \pi_\theta(i)\right) \cdot \left[r(1) - \pi_\theta^\top r\right] - \sum_{i=3}^K \pi_\theta(i) \cdot \left[r(1) - r(i)\right] \tag{120}$$

$$\ge \left(2\pi_\theta(1) + \sum_{i=3}^K \pi_\theta(i)\right) \cdot \left[r(1) - \pi_\theta^\top r\right] - \sum_{i=3}^K \pi_\theta(i) \tag{121}$$

$$\ge \left(2\pi_\theta(1) + \sum_{i=3}^K \pi_\theta(i)\right) \cdot \frac{\Delta}{K} - \sum_{i=3}^K \pi_\theta(i), \tag{122}$$

where the second equation is according to

$$\pi_\theta(2) \cdot \left[r(2) - \pi_\theta^\top r\right] + \sum_{i \ne 2} \pi_\theta(i) \cdot \left[r(i) - \pi_\theta^\top r\right] = \pi_\theta^\top r - \pi_\theta^\top r = 0, \tag{123}$$

and the first inequality is by $0 < r(1) - r(i) \le 1$, and the second inequality is because of

$$r(1) - \pi_\theta^\top r = \left[1 - \pi_\theta(1)\right] \cdot r(1) - \sum_{i=2}^K \pi_\theta(i) \cdot r(i) = \sum_{i=2}^K \pi_\theta(i) \cdot \left[r(1) - r(i)\right] \tag{124}$$

$$\ge \sum_{i=2}^K \pi_\theta(i) \cdot \Delta \ge \max_{a \ne a^*}\{\pi_\theta(a)\} \cdot \Delta \tag{125}$$

$$\ge \frac{\Delta}{K}. \qquad \left(\pi_\theta(a^*) < \max_{a \ne a^*}\{\pi_\theta(a)\},\ \max_{a \ne a^*}\{\pi_\theta(a)\} = \max_a\{\pi_\theta(a)\} \ge \frac{1}{K}\right) \tag{126}$$

Note that $\sum_{i=3}^{K} \pi_\theta(i) = 1 - \pi_\theta(1) - \pi_\theta(2)$, we have

$$\begin{aligned} \frac{d\pi_\theta^\top r}{d\theta(a^*)} - \frac{d\pi_\theta^\top r}{d\theta(a)} &\ge \pi_\theta(1) \cdot \frac{2\Delta}{K} - [1 - \pi_\theta(1) - \pi_\theta(2)] \cdot \left(1 - \frac{\Delta}{K}\right) && (127) \\ &= \frac{2\Delta}{K} \cdot \left[\pi_\theta(1) \cdot \left(1 + \frac{K}{2\Delta} \cdot \left(1 - \frac{\Delta}{K}\right)\right) - [1 - \pi_\theta(2)] \cdot \frac{K}{2\Delta} \cdot \left(1 - \frac{\Delta}{K}\right) \right] && (128) \\ &\ge \frac{2\Delta}{K} \cdot \left[\frac{K}{2\Delta} \cdot \left(1 - \frac{\Delta}{K}\right) - [1 - \pi_\theta(2)] \cdot \frac{K}{2\Delta} \cdot \left(1 - \frac{\Delta}{K}\right) \right] \qquad \left(\pi_\theta(1) \ge \frac{c}{c+1}\right) && (129) \\ &= \pi_\theta(2) \cdot \left(1 - \frac{\Delta}{K}\right) \ge 0, && (130) \end{aligned}$$

which means $\theta \in \mathcal{R}_1$, and thus $\mathcal{N}_c \subset \mathcal{R}_1$.

Third part. According to the asymptotic convergence results of Agarwal et al. (2019, Theorem 5.1), $\pi_{\theta_t}(a^*) \to 1$ as $t \to \infty$. Hence, there exists $t_0 > 0$, such that $\pi_{\theta_{t_0}}(a^*) \ge \frac{c}{c+1}$, which means $\theta_{t_0} \in \mathcal{N}_c \subset \mathcal{R}_1$. According to the first part in our proof, i.e., once θ_t is in $\mathcal{R}_1$, following gradient update θ_{t+1} will be in $\mathcal{R}_1$, and $\pi_{\theta_t}(a^*)$ is increasing in $\mathcal{R}_1$, we have $\inf_t \pi_{\theta_t}(a^*) = \min_{1 \le t \le t_0} \pi_{\theta_t}(a^*)$. t_0 depends on initialization and c, which only depends on the problem. □

Proposition 2. For any initialization there exist $t_0 > 0$ 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, $t_0 = 1$.

Proof. $t_0 := \min\{t : \pi_{\theta_t}(a^*) \ge \frac{c}{c+1}\}$, where $c := \frac{K}{2\Delta} \cdot \left(1 - \frac{\Delta}{K}\right)$ in the proof for Lemma 5 satisfies for any $t \ge t_0$, $t \mapsto \pi_{\theta_t}(a^*)$ is increasing. Next we show that when π_{θ_1} is the uniform distribution, $t_0 = 1$. Recall the definition of $\mathcal{R}_1$ in the proof for Lemma 5, and define another region $\mathcal{R}_2$ as,

$$\mathcal{R}_1 := \left\{ \theta : \frac{d\pi_\theta^\top r}{d\theta(a^*)} \ge \frac{d\pi_\theta^\top r}{d\theta(a)}, \ \forall a \ne a^* \right\}, \tag{131}$$

$$\mathcal{R}_2 := \left\{ \theta : \pi_\theta(a^*) \ge \pi_\theta(a), \ \forall a \ne a^* \right\}. \tag{132}$$

It is obvious that $\theta_1 \in \mathcal{R}_2$ if π_{θ_1} is the uniform distribution. Next we show $\mathcal{R}_2 \subset \mathcal{R}_1$. Suppose $\pi_\theta(a^*) \ge \pi_\theta(a)$. Then,

$$\begin{aligned} \frac{d\pi_\theta^\top r}{d\theta(a^*)} &= \pi_\theta(a^*) \cdot \left[r(a^*) - \pi_\theta^\top r \right] && (133) \\ &> \pi_\theta(a) \cdot \left[r(a) - \pi_\theta^\top r \right] \qquad \left(r(a^*) - \pi_\theta^\top r > 0, \ r(a^*) > r(a) \right) && (134) \\ &:= \frac{d\pi_\theta^\top r}{d\theta(a)}. && (135) \end{aligned}$$

Therefore we have $\theta_1 \in \mathcal{R}_1$ and $t_0 = 1$. □

Theorem 2 (Arbitrary initialization)**.** Using Update 1 with $\eta = 2/5$, for $t > 0$,

$$(\pi^* - \pi_{\theta_t})^\top r \le C/t, \tag{136}$$

where $1/C = \left[\inf_{t \ge 1} \pi_{\theta_t}(a^*)\right]^2 > 0$ is a constant that depends on r and θ_1, but it does not depend on the time t.

Proof. According to Lemmas 4 and 5, the claim immediately holds, with $1/C = \left[\inf_{t \ge 1} \pi_{\theta_t}(a^*)\right]^2 \in \Omega(1)$. □

Theorem 3 (Uniform initialization)**.** Using Update 1 with $\eta = 2/5$ and $\pi_{\theta_1}(a) = 1/K, \forall a$, for all $t > 0$,

$$(\pi^* - \pi_{\theta_t})^\top r \le 5K^2/t, \tag{137}$$

$$\sum_{t=1}^{T} (\pi^* - \pi_{\theta_t})^\top r \le \min\left\{ K\sqrt{5T}, \ 5K^2 \log T + 1 \right\}. \tag{138}$$

Proof. Since initial policy is uniform policy, $\pi_{\theta_1}(a^*) \ge 1/K$. According to Proposition 2, for all $t \ge t_0 = 1$, $t \mapsto \pi_{\theta_t}(a^*)$ is increasing, we have $\pi_{\theta_t}(a^*) \ge 1/K$, $\forall t > 0$, and $c_t := \min_{1 \le s \le t} \pi_{\theta_s}(a^*) \ge 1/K$. According to Lemma 4,

$$(\pi^* - \pi_{\theta_t})^\top r \le \frac{5}{c_t^2} \cdot \frac{1}{t}, \tag{139}$$

we have $(\pi^* - \pi_{\theta_t})^\top r \le 5K^2/t, \forall t > 0$. Remaining results follow from Eq. (78) and $c_T \ge 1/K$. □

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

$$t_0 := \min_{t \ge 1} \left\{ \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} \ge \frac{3}{2\Delta} \right\}. \tag{140}$$

Proof. Recall the definition of $\mathcal{R}_1$ in the proof for Lemma 5,

$$\mathcal{R}_1 := \left\{ \theta : \frac{d\pi_\theta^\top r}{d\theta(a^*)} \ge \frac{d\pi_\theta^\top r}{d\theta(a)}, \ \forall a \ne a^* \right\}. \tag{141}$$

We prove if $\frac{\pi_\theta(1)}{\pi_\theta(3)} \ge \frac{3}{2\Delta}$, then $\theta \in \mathcal{R}_1$. Suppose $\frac{\pi_\theta(1)}{\pi_\theta(3)} \ge \frac{3}{2\Delta}$. There are two cases.

(a) If $\pi_\theta(1) \ge \max\{\pi_\theta(2), \pi_\theta(3)\}$, then $\theta \in \mathcal{R}_2 \subset \mathcal{R}_1$ in the proof for Proposition 2.

(b) If $\pi_\theta(1) < \max\{\pi_\theta(2), \pi_\theta(3)\}$, then

$$\begin{aligned}
\frac{d\pi_\theta^\top r}{d\theta(1)} - \frac{d\pi_\theta^\top r}{d\theta(2)} &= \pi_\theta(1) \cdot \left[r(1) - \pi_\theta^\top r \right] - \pi_\theta(2) \cdot \left[r(2) - \pi_\theta^\top r \right] && (142) \\
&= 2\pi_\theta(1) \cdot \left[r(1) - \pi_\theta^\top r \right] + \pi_\theta(3) \cdot \left[r(3) - \pi_\theta^\top r \right] && (143) \\
&= \pi_\theta(3) \cdot \left[\frac{2\pi_\theta(1)}{\pi_\theta(3)} \cdot \left[r(1) - \pi_\theta^\top r \right] - \left[r(3) - \pi_\theta^\top r \right] \right] && (144) \\
&\ge \pi_\theta(3) \cdot \left[\frac{2\pi_\theta(1)}{\pi_\theta(3)} \cdot \left[r(1) - \pi_\theta^\top r \right] - 1 \right] && (145) \\
&\ge \pi_\theta(3) \cdot \left[\frac{2\pi_\theta(1)}{\pi_\theta(3)} \cdot \frac{\Delta}{3} - 1 \right] && (146) \\
&\ge \pi_\theta(3) \cdot (1 - 1) = 0, && (147)
\end{aligned}$$

where the second equation is according to

$$\pi_\theta(1) \cdot \left[r(1) - \pi_\theta^\top r \right] + \pi_\theta(2) \cdot \left[r(2) - \pi_\theta^\top r \right] + \pi_\theta(3) \cdot \left[r(3) - \pi_\theta^\top r \right] = \pi_\theta^\top r - \pi_\theta^\top r = 0, \tag{148}$$

and the first inequality is by $0 < \pi_\theta^\top r - r(3) \le 1$, and the second inequality is because of

$$\begin{aligned}
r(1) - \pi_\theta^\top r &= [1 - \pi_\theta(1)] \cdot r(1) - [\pi_\theta(2) \cdot r(2) + \pi_\theta(3) \cdot r(3)] && (149) \\
&= \pi_\theta(2) \cdot [r(1) - r(2)] + \pi_\theta(3) \cdot [r(1) - r(3)] && (150) \\
&\ge [\pi_\theta(2) + \pi_\theta(3)] \cdot \Delta && (151) \\
&\ge \max\{\pi_\theta(2), \pi_\theta(3)\} \cdot \Delta && (152) \\
&\ge \frac{\Delta}{3}. \qquad \left(\pi_\theta(1) < \max\{\pi_\theta(2), \pi_\theta(3)\}, \ \max\{\pi_\theta(2), \pi_\theta(3)\} = \max_a \{\pi_\theta(a)\} \ge \frac{1}{3} \right) && (153)
\end{aligned}$$

Note that since $r(1) > \pi_\theta^\top r$, and $r(3) < \pi_\theta^\top r$, we have

$$\begin{aligned}
\frac{d\pi_\theta^\top r}{d\theta(1)} - \frac{d\pi_\theta^\top r}{d\theta(3)} &= \pi_\theta(1) \cdot \left[r(1) - \pi_\theta^\top r \right] - \pi_\theta(3) \cdot \left[r(3) - \pi_\theta^\top r \right] && (154) \\
&\ge 0 - 0 = 0. && (155)
\end{aligned}$$

Therefore we have $\frac{d\pi_\theta^\top r}{d\theta(1)} \ge \frac{d\pi_\theta^\top r}{d\theta(2)}$ and $\frac{d\pi_\theta^\top r}{d\theta(1)} \ge \frac{d\pi_\theta^\top r}{d\theta(3)}$, i.e., $\theta \in \mathcal{R}_1$. □

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

Proof. See Agarwal et al. (2019, Lemma E.4). Our proof is for completeness. Denote $\theta_\alpha = \theta + \alpha u$, where $\alpha \in \mathbb{R}$ and $u \in \mathbb{R}^{SA}$. For any $s \in \mathcal{S}$,

$$\sum_a \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}\Big|_{\alpha=0} \right| = \sum_a \left| \left\langle \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \theta_\alpha}\Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \tag{156}$$

$$= \sum_a \left| \left\langle \frac{\partial \pi_\theta(a|s)}{\partial \theta}, u \right\rangle \right|. \tag{157}$$

Since $\frac{\partial \pi_\theta(a|s)}{\partial \theta(s',\cdot)} = 0$, for $s' \neq s$,

$$\sum_a \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}\Big|_{\alpha=0} \right| = \sum_a \left| \left\langle \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)}, u(s,\cdot) \right\rangle \right| \tag{158}$$

$$= \sum_a \pi_\theta(a|s) \cdot \left| u(s,a) - \pi_\theta(\cdot|s)^\top u(s,\cdot) \right| \tag{159}$$

$$\leq \max_a |u(s,a)| + |\pi_\theta(\cdot|s)^\top u(s,\cdot)| \leq 2 \cdot \|u\|_2. \tag{160}$$

Similarly,

$$\sum_a \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\Big|_{\alpha=0} \right| = \sum_a \left| \left\langle \frac{\partial}{\partial \theta_\alpha} \left\{ \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \right\} \Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \tag{161}$$

$$= \sum_a \left| \left\langle \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \theta_\alpha^2}\Big|_{\alpha=0} \frac{\partial \theta_\alpha}{\partial \alpha}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \tag{162}$$

$$= \sum_a \left| \left\langle \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} u(s,\cdot), u(s,\cdot) \right\rangle \right|. \tag{163}$$

Denote $S(a,\theta) \coloneqq \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} \in \mathbb{R}^{A\times A}$. $\forall i,j \in [A]$, the value of $S(a,\theta)$ is,

$$S_{i,j} = \frac{\partial \{\delta_{ia}\pi_\theta(a|s) - \pi_\theta(a|s)\pi_\theta(i|s)\}}{\partial \theta(s,j)} \tag{164}$$

$$= \delta_{ia} \cdot [\delta_{ja}\pi_\theta(a|s) - \pi_\theta(a|s)\pi_\theta(j|s)] - \pi_\theta(a|s) \cdot [\delta_{ij}\pi_\theta(j|s) - \pi_\theta(i|s)\pi_\theta(j|s)] - \pi_\theta(i|s) \cdot [\delta_{ja}\pi_\theta(a|s) - \pi_\theta(a|s)\pi_\theta(j|s)], \tag{165}$$

where the δ notation is as defined in Eq. (52). Then we have,

$$\left| \left\langle \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} u(s,\cdot), u(s,\cdot) \right\rangle \right| = \left| \sum_{i=1}^{A} \sum_{j=1}^{A} S_{i,j} u(s,i) u(s,j) \right| \tag{166}$$

$$= \pi_\theta(a|s) \cdot \left| u(s,a)^2 - 2 \cdot u(s,a) \cdot \pi_\theta(\cdot|s)^\top u(s,\cdot) - \pi_\theta(\cdot|s)^\top \left(u(s,\cdot) \odot u(s,\cdot) \right) + 2 \cdot \left(\pi_\theta(\cdot|s)^\top u(s,\cdot) \right)^2 \right|. \tag{167}$$

Therefore we have,

$$\sum_a \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\Big|_{\alpha=0} \right| \leq \max_a \left\{ u(s,a)^2 + 2 \cdot \left| u(s,a) \cdot \pi_\theta(\cdot|s)^\top u(s,\cdot) \right| \right\} + \pi_\theta(\cdot|s)^\top \left(u(s,\cdot) \odot u(s,\cdot) \right) + 2 \cdot \left(\pi_\theta(\cdot|s)^\top u(s,\cdot) \right)^2 \tag{168}$$

$$\leq \|u(s,\cdot)\|_2^2 + 2 \cdot \|u(s,\cdot)\|_2^2 + \|u(s,\cdot)\|_2^2 + 2 \cdot \|u(s,\cdot)\|_2^2 \leq 6 \cdot \|u\|_2^2. \tag{169}$$

Define $P(\alpha) \in \mathbb{R}^{S\times S}$, where $\forall (s,s')$,

$$\left[P(\alpha) \right]_{(s,s')} \coloneqq \sum_a \pi_{\theta_\alpha}(a|s) \cdot \mathcal{P}(s'|s,a). \tag{170}$$

The derivative w.r.t. α is

$$\left[\frac{\partial P(\alpha)}{\partial \alpha} \Big|_{\alpha=0} \right]_{(s,s')} = \sum_a \left[\frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \Big|_{\alpha=0} \right] \cdot \mathcal{P}(s'|s,a). \tag{171}$$

For any vector $x \in \mathbb{R}^S$, we have

$$\left[\frac{\partial P(\alpha)}{\partial \alpha} \Big|_{\alpha=0} x \right]_{(s)} = \sum_{s'} \sum_a \left[\frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \Big|_{\alpha=0} \right] \cdot \mathcal{P}(s'|s,a) \cdot x(s'). \tag{172}$$

The ℓ_∞ norm is upper bounded as

$$\begin{align} \left\| \frac{\partial P(\alpha)}{\partial \alpha} \Big|_{\alpha=0} x \right\|_\infty &= \max_s \left| \sum_{s'} \sum_a \left[\frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \Big|_{\alpha=0} \right] \cdot \mathcal{P}(s'|s,a) \cdot x(s') \right| \tag{173} \\ &\le \max_s \sum_a \sum_{s'} \mathcal{P}(s'|s,a) \cdot \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \Big|_{\alpha=0} \right| \cdot \|x\|_\infty \tag{174} \\ &= \max_s \sum_a \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \Big|_{\alpha=0} \right| \cdot \|x\|_\infty \tag{175} \\ &\le 2 \cdot \|u\|_2 \cdot \|x\|_\infty. \qquad \text{(Eq. (158))} \tag{176} \end{align}$$

Similarly, taking second derivative w.r.t. α,

$$\left[\frac{\partial^2 P(\alpha)}{\partial \alpha^2} \Big|_{\alpha=0} \right]_{(s,s')} = \sum_a \left[\frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2} \Big|_{\alpha=0} \right] \cdot \mathcal{P}(s'|s,a). \tag{177}$$

The ℓ_∞ norm is upper bounded as

$$\begin{align} \left\| \frac{\partial^2 P(\alpha)}{\partial \alpha^2} \Big|_{\alpha=0} x \right\|_\infty &= \max_s \left| \sum_{s'} \sum_a \left[\frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2} \Big|_{\alpha=0} \right] \cdot \mathcal{P}(s'|s,a) \cdot x(s') \right| \tag{178} \\ &\le \max_s \sum_a \sum_{s'} \mathcal{P}(s'|s,a) \cdot \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2} \Big|_{\alpha=0} \right| \cdot \|x\|_\infty \tag{179} \\ &= \max_s \sum_a \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2} \Big|_{\alpha=0} \right| \cdot \|x\|_\infty \tag{180} \\ &\le 6 \cdot \|u\|_2^2 \cdot \|x\|_\infty. \qquad \text{(Eq. (168))} \tag{181} \end{align}$$

Next, consider the state value function of π_{θ_α},

$$V^{\pi_{\theta_\alpha}}(s) = \sum_a \pi_{\theta_\alpha}(a|s) \cdot r(s,a) + \gamma \sum_a \pi_{\theta_\alpha}(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \cdot V^{\pi_{\theta_\alpha}}(s'), \tag{182}$$

which implies,

$$V^{\pi_{\theta_\alpha}}(s) = e_s^\top M(\alpha) r_{\theta_\alpha}, \tag{183}$$

where

$$M(\alpha) \coloneqq \left(\mathbf{Id} - \gamma P(\alpha) \right)^{-1}, \tag{184}$$

and $r_{\theta_\alpha} \in \mathbb{R}^S$, $\forall s$,

$$r_{\theta_\alpha}(s) \coloneqq \sum_a \pi_{\theta_\alpha}(a|s) \cdot r(s,a). \tag{185}$$

Since $[P(\alpha)]_{(s,s')} \ge 0$, $\forall (s,s')$, and

$$M(\alpha) = (\mathbf{Id} - \gamma P(\alpha))^{-1} = \sum_{t=0}^{\infty} \gamma^t \, [P(\alpha)]^t, \tag{186}$$

we have $[M(\alpha)]_{(s,s')} \ge 0$, $\forall (s,s')$. Denote $[M(\alpha)]_{i,:}$ as the i-th row vector of $M(\alpha)$. We have

$$\mathbf{1} = \frac{1}{1-\gamma} \cdot (\mathbf{Id} - \gamma P(\alpha))\, \mathbf{1} \Longrightarrow M(\alpha)\mathbf{1} = \frac{1}{1-\gamma} \cdot \mathbf{1}, \tag{187}$$

which implies, $\forall i$,

$$\left\| [M(\alpha)]_{i,:} \right\|_1 = \sum_j [M(\alpha)]_{(i,j)} = \frac{1}{1-\gamma}. \tag{188}$$

Therefore, for any vector $x \in \mathbb{R}^S$,

$$\begin{aligned} \| M(\alpha) x \|_\infty &= \max_i \left| [M(\alpha)]_{i,:}^\top x \right| && (189) \\ &\le \max_i \left\| [M(\alpha)]_{i,:} \right\|_1 \cdot \|x\|_\infty && (190) \\ &= \frac{1}{1-\gamma} \cdot \|x\|_\infty. && (191) \end{aligned}$$

According to Assumption 1, $r(s,a) \in [0,1]$, $\forall (s,a)$. We have,

$$\| r_{\theta_\alpha} \|_\infty = \max_s |r_{\theta_\alpha}(s)| = \max_s \left| \sum_a \pi_{\theta_\alpha}(a|s) \cdot r(s,a) \right| \le 1. \tag{192}$$

Since $\frac{\partial \pi_\theta(a|s)}{\partial \theta(s',\cdot)} = 0$, for $s' \ne s$,

$$\begin{aligned} \left| \frac{\partial r_{\theta_\alpha}(s)}{\partial \alpha} \right| &= \left| \left(\frac{\partial r_{\theta_\alpha}(s)}{\partial \theta_\alpha} \right)^\top \frac{\partial \theta_\alpha}{\partial \alpha} \right| && (193) \\ &= \left| \left(\frac{\partial \{ \pi_{\theta_\alpha}(\cdot|s)^\top r(s,\cdot) \}}{\partial \theta_\alpha(s,\cdot)} \right)^\top u(s,\cdot) \right| && (194) \\ &= \left| \left(H\left(\pi_{\theta_\alpha}(\cdot|s)\right) r(s,\cdot) \right)^\top u(s,\cdot) \right| && (195) \\ &\le \left\| H\left(\pi_{\theta_\alpha}(\cdot|s)\right) r(s,\cdot) \right\|_1 \cdot \| u(s,\cdot) \|_\infty. && (196) \end{aligned}$$

Similar with Eq. (61), the ℓ_1 norm is upper bounded as

$$\begin{aligned} \left\| H\left(\pi_{\theta_\alpha}(\cdot|s)\right) r(s,\cdot) \right\|_1 &= \sum_a \pi_{\theta_\alpha}(a|s) \cdot \left| r(s,a) - \pi_{\theta_\alpha}(\cdot|s)^\top r(s,\cdot) \right| && (197) \\ &\le \max_a \left| r(s,a) - \pi_{\theta_\alpha}(\cdot|s)^\top r(s,\cdot) \right| && (198) \\ &\le 1. \qquad (r(s,a) \in [0,1]) && (199) \end{aligned}$$

Therefore we have,

$$\begin{aligned} \left\| \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \right\|_\infty &= \max_s \left| \frac{\partial r_{\theta_\alpha}(s)}{\partial \alpha} \right| && (200) \\ &\le \max_s \left\| H\left(\pi_{\theta_\alpha}(\cdot|s)\right) r(s,\cdot) \right\|_1 \cdot \| u(s,\cdot) \|_\infty && (201) \\ &\le \|u\|_2. && (202) \end{aligned}$$

Similarly,

$$\begin{aligned}
\left\| \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2} \right\|_\infty &= \max_s \left| \frac{\partial^2 r_{\theta_\alpha}(s)}{\partial \alpha^2} \right| && (203) \\
&= \max_s \left| \left(\frac{\partial}{\partial \theta_\alpha} \left\{ \frac{\partial r_{\theta_\alpha}(s)}{\partial \alpha} \right\} \right)^\top \frac{\partial \theta_\alpha}{\partial \alpha} \right| && (204) \\
&= \max_s \left| \left(\frac{\partial^2 r_{\theta_\alpha}(s)}{\partial \theta_\alpha^2} \frac{\partial \theta_\alpha}{\partial \alpha} \right)^\top \frac{\partial \theta_\alpha}{\partial \alpha} \right| && (205) \\
&= \max_s \left| u(s, \cdot)^\top \frac{\partial^2 \{ \pi_{\theta_\alpha}(\cdot|s)^\top r(s, \cdot) \}}{\partial \theta_\alpha(s, \cdot)^2} u(s, \cdot) \right| && (206) \\
&\le 5/2 \cdot \| u(s, \cdot) \|_2^2 \le 3 \cdot \| u \|_2^2. \qquad \text{(Eq. (64))} && (207)
\end{aligned}$$

Taking derivative w.r.t. α in Eq. (183),

$$\frac{\partial V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha} = \gamma \cdot e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} + e_s^\top M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha}. \tag{208}$$

Taking second derivative w.r.t. α,

$$\begin{aligned}
\frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2} &= 2\gamma^2 \cdot e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} + \gamma \cdot e_s^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha} && (209) \\
&\quad + 2\gamma \cdot e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} + e_s^\top M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}. && (210)
\end{aligned}$$

For the last term,

$$\begin{aligned}
\left| e_s^\top M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0} \right| &\le \| e_s \|_1 \cdot \left\| M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0} \right\|_\infty && (211) \\
&\le \frac{1}{1-\gamma} \cdot \left\| \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0} \right\|_\infty \qquad \text{(Eq. (189))} && (212) \\
&\le \frac{3}{1-\gamma} \cdot \| u \|_2^2. \qquad \text{(Eq. (203))} && (213)
\end{aligned}$$

For the second last term,

$$\begin{aligned}
\left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right| &\le \left\| M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty && (214) \\
&\le \frac{1}{1-\gamma} \cdot \left\| \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty \qquad \text{(Eq. (189))} && (215) \\
&\le \frac{2 \cdot \| u \|_2}{1-\gamma} \cdot \left\| M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty \qquad \text{(Eq. (173))} && (216) \\
&\le \frac{2 \cdot \| u \|_2}{(1-\gamma)^2} \cdot \left\| \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty \qquad \text{(Eq. (189))} && (217) \\
&\le \frac{2 \cdot \| u \|_2}{(1-\gamma)^2} \cdot \| u \|_2 = \frac{2}{(1-\gamma)^2} \cdot \| u \|_2^2. \qquad \text{(Eq. (200))} && (218)
\end{aligned}$$

For the second term,

$$\left| e_s^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha} \Big|_{\alpha=0} \right| \le \left\| M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \tag{219}$$

$$\le \frac{1}{1-\gamma} \cdot \left\| \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \qquad \text{(Eq. (189))} \tag{220}$$

$$\le \frac{6 \cdot \|u\|_2^2}{1-\gamma} \cdot \left\| M(\alpha) r_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \qquad \text{(Eq. (178))} \tag{221}$$

$$\le \frac{6 \cdot \|u\|_2^2}{(1-\gamma)^2} \cdot \left\| r_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \qquad \text{(Eq. (189))} \tag{222}$$

$$\le \frac{6}{(1-\gamma)^2} \cdot \|u\|_2^2. \qquad \text{(Eq. (192))} \tag{223}$$

For the first term, according to Eqs. (173), (189) and (192),

$$\left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} \Big|_{\alpha=0} \right| \le \left\| M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \tag{224}$$

$$\le \frac{1}{1-\gamma} \cdot 2 \cdot \|u\|_2 \cdot \frac{1}{1-\gamma} \cdot 2 \cdot \|u\|_2 \cdot \frac{1}{1-\gamma} \cdot 1 \tag{225}$$

$$= \frac{4}{(1-\gamma)^3} \cdot \|u\|_2^2. \tag{226}$$

Combining Eqs. (211), (214), (219) and (224) with Eq. (209),

$$\left| \frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2} \Big|_{\alpha=0} \right| \le 2\gamma^2 \cdot \left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} \Big|_{\alpha=0} \right| + \gamma \cdot \left| e_s^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha} \Big|_{\alpha=0} \right| \tag{227}$$

$$+ 2\gamma \cdot \left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right| + \left| e_s^\top M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0} \right| \tag{228}$$

$$\le \left(2\gamma^2 \cdot \frac{4}{(1-\gamma)^3} + \gamma \cdot \frac{6}{(1-\gamma)^2} + 2\gamma \cdot \frac{2}{(1-\gamma)^2} + \frac{3}{1-\gamma} \right) \cdot \|u\|_2^2 \tag{229}$$

$$\le \frac{8}{(1-\gamma)^3} \cdot \|u\|_2^2, \tag{230}$$

which implies for all $y \in \mathbb{R}^{SA}$ and θ,

$$\left| y^\top \frac{\partial^2 V^{\pi_\theta}(s)}{\partial \theta^2} y \right| = \left| \left(\frac{y}{\|y\|_2} \right)^\top \frac{\partial^2 V^{\pi_\theta}(s)}{\partial \theta^2} \left(\frac{y}{\|y\|_2} \right) \right| \cdot \|y\|_2^2 \tag{231}$$

$$\le \max_{\|u\|_2 = 1} \left| \left\langle \frac{\partial^2 V^{\pi_\theta}(s)}{\partial \theta^2} u, u \right\rangle \right| \cdot \|y\|_2^2 \tag{232}$$

$$= \max_{\|u\|_2 = 1} \left| \left\langle \frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \theta_\alpha^2} \Big|_{\alpha=0} \frac{\partial \theta_\alpha}{\partial \alpha}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \cdot \|y\|_2^2 \tag{233}$$

$$= \max_{\|u\|_2 = 1} \left| \left\langle \frac{\partial}{\partial \theta_\alpha} \left\{ \frac{\partial V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha} \right\} \Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \cdot \|y\|_2^2 \tag{234}$$

$$= \max_{\|u\|_2 = 1} \left| \frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2} \Big|_{\alpha=0} \right| \cdot \|y\|_2^2 \tag{235}$$

$$\le \frac{8}{(1-\gamma)^3} \cdot \|y\|_2^2. \qquad \text{(Eq. (227))} \tag{236}$$

Denote $\theta_\xi = \theta + \xi(\theta' - \theta)$, where $\xi \in [0, 1]$. According to Taylor's theorem, $\forall s, \forall \theta,\ \theta'$,

$$\left| V^{\pi_{\theta'}}(s) - V^{\pi_\theta}(s) - \left\langle \frac{\partial V^{\pi_\theta}(s)}{\partial \theta}, \theta' - \theta \right\rangle \right| = \frac{1}{2} \cdot \left| (\theta' - \theta)^\top \frac{\partial^2 V^{\pi_{\theta_\xi}}(s)}{\partial \theta_\xi^2} (\theta' - \theta) \right| \tag{237}$$

$$\leq \frac{4}{(1-\gamma)^3} \cdot \|\theta' - \theta\|_2^2. \qquad \text{(Eq. (231))} \tag{238}$$

Since $V^{\pi_\theta}(s)$ is $8/(1-\gamma)^3$-smooth, for any state s, $V^{\pi_\theta}(\rho) \coloneqq \mathbb{E}_{s\sim\rho}\left[V^{\pi_\theta}(s)\right]$ is also $8/(1-\gamma)^3$-smooth. □

Lemma 8 (Non-uniform Łojasiewicz)**.** Suppose $\mu(s) > 0$ for all state s. $\pi_\theta(\cdot|s) \coloneqq \mathrm{softmax}(\theta(s, \cdot)), \forall s$.

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \geq \frac{1}{\sqrt{S}} \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot \min_s \pi_\theta(a^*(s)|s) \cdot \left[V^*(\rho) - V^{\pi_\theta}(\rho) \right], \tag{239}$$

where $a^*(s) \coloneqq \arg\max_a \pi^*(a|s), \forall s$. Also

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \geq \frac{1}{\sqrt{SA}} \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot \left[\min_s \sum_{\bar{a}(s) \in \bar{\mathcal{A}}^{\pi_\theta}(s)} \pi_\theta(\bar{a}(s)|s) \right] \cdot \left[V^*(\rho) - V^{\pi_\theta}(\rho) \right], \tag{240}$$

where $\bar{\mathcal{A}}^\pi(s) \coloneqq \{\bar{a}(s) \in \mathcal{A} : Q^\pi(s, \bar{a}(s)) = \max_a Q^\pi(s, a)\}$ is the greedy action set for state s given policy π.

Proof. Note $a^*(s)$ is the action that the optimal policy π^* selects under state s. We have,

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 = \left[\sum_{s,a} \left(\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)} \right)^2 \right]^{\frac{1}{2}} \tag{241}$$

$$\geq \left[\sum_s \left(\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a^*(s))} \right)^2 \right]^{\frac{1}{2}} \tag{242}$$

$$\geq \frac{1}{\sqrt{S}} \sum_s \left| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a^*(s))} \right| \qquad \left(\text{CauchySchwarz, } \|x\|_1 = |\langle \mathbf{1}, |x| \rangle| \leq \|\mathbf{1}\|_2 \cdot \|x\|_2\right) \tag{243}$$

$$= \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_s \left| d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a^*(s)|s) \cdot A^{\pi_\theta}(s, a^*(s)) \right| \qquad \text{(Lemma 1)} \tag{244}$$

$$= \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_s d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a^*(s)|s) \cdot \left| A^{\pi_\theta}(s, a^*(s)) \right|. \qquad \left(d_\mu^{\pi_\theta}(s) \geq 0,\ \pi_\theta(a^*(s)|s) \geq 0 \right) \tag{245}$$

Define the distribution mismatch coefficient as $\left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty \coloneqq \max_s \frac{d_\rho^{\pi^*}(s)}{d_\mu^{\pi_\theta}(s)}$. We have,

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \geq \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_s \frac{d_\mu^{\pi_\theta}(s)}{d_\rho^{\pi^*}(s)} \cdot d_\rho^{\pi^*}(s) \cdot \pi_\theta(a^*(s)|s) \cdot \left| A^{\pi_\theta}(s, a^*(s)) \right| \tag{246}$$

$$\geq \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot \min_s \pi_\theta(a^*(s)|s) \cdot \sum_s d_\rho^{\pi^*}(s) \cdot \left| A^{\pi_\theta}(s, a^*(s)) \right| \tag{247}$$

$$\geq \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot \min_s \pi_\theta(a^*(s)|s) \cdot \sum_s d_\rho^{\pi^*}(s) \cdot A^{\pi_\theta}(s, a^*(s)) \tag{248}$$

$$= \frac{1}{\sqrt{S}} \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot \min_s \pi_\theta(a^*(s)|s) \cdot \frac{1}{1-\gamma} \sum_s d_\rho^{\pi^*}(s) \sum_a \pi^*(a|s) \cdot A^{\pi_\theta}(s,a) \tag{249}$$

$$= \frac{1}{\sqrt{S}} \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot \min_s \pi_\theta(a^*(s)|s) \cdot \left[V^*(\rho) - V^{\pi_\theta}(\rho) \right], \tag{250}$$

where the last equation is according to the performance difference lemma of Lemma 19. Next, given any policy π, define the greedy action set for each state s,

$$\bar{\mathcal{A}}^{\pi}(s) \coloneqq \left\{ \bar{a}(s) \in \mathcal{A} : Q^{\pi}(s, \bar{a}(s)) = \max_a Q^{\pi}(s,a) \right\}. \tag{251}$$

Using similar arguments, we have,

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \ge \frac{1}{\sqrt{SA}} \sum_{s,a} \left| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)} \right| \qquad \text{(CauchySchwarz)} \tag{252}$$

$$= \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{SA}} \sum_s d_\mu^{\pi_\theta}(s) \sum_a \pi_\theta(a|s) \cdot |A^{\pi_\theta}(s,a)| \qquad \text{(Lemma 1)} \tag{253}$$

$$\ge \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{SA}} \sum_s d_\mu^{\pi_\theta}(s) \sum_{\bar{a}(s) \in \bar{\mathcal{A}}^{\pi_\theta}(s)} \pi_\theta(\bar{a}(s)|s) \cdot |A^{\pi_\theta}(s, \bar{a}(s))| \tag{254}$$

$$\ge \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{SA}} \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot \left[\min_s \sum_{\bar{a}(s) \in \bar{\mathcal{A}}^{\pi_\theta}(s)} \pi_\theta(\bar{a}(s)|s) \right] \cdot \sum_s d_\rho^{\pi^*}(s) \cdot \left| \max_a Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s) \right|, \tag{255}$$

where the last inequality is because of for all $\bar{a}(s) \in \bar{\mathcal{A}}^{\pi_\theta}(s)$,

$$A^{\pi_\theta}(s, \bar{a}(s)) = \max_a Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s), \tag{256}$$

which is the same value across all $\bar{a}(s) \in \bar{\mathcal{A}}^{\pi_\theta}(s)$. Then we have,

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \ge \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{SA}} \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot \left[\min_s \sum_{\bar{a}(s) \in \bar{\mathcal{A}}^{\pi_\theta}(s)} \pi_\theta(\bar{a}(s)|s) \right] \cdot \sum_s d_\rho^{\pi^*}(s) \cdot \left[\max_a Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s) \right] \tag{257}$$

$$\ge \frac{1}{\sqrt{SA}} \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot \left[\min_s \sum_{\bar{a}(s) \in \bar{\mathcal{A}}^{\pi_\theta}(s)} \pi_\theta(\bar{a}(s)|s) \right] \cdot \frac{1}{1-\gamma} \sum_s d_\rho^{\pi^*}(s) \cdot \left[Q^{\pi_\theta}(s, a^*(s)) - V^{\pi_\theta}(s) \right] \tag{258}$$

$$= \frac{1}{\sqrt{SA}} \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot \left[\min_s \sum_{\bar{a}(s) \in \bar{\mathcal{A}}^{\pi_\theta}(s)} \pi_\theta(\bar{a}(s)|s) \right] \cdot \frac{1}{1-\gamma} \sum_s d_\rho^{\pi^*}(s) \sum_a \pi^*(a|s) \cdot A^{\pi_\theta}(s,a) \tag{259}$$

$$= \frac{1}{\sqrt{SA}} \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot \left[\min_s \sum_{\bar{a}(s) \in \bar{\mathcal{A}}^{\pi_\theta}(s)} \pi_\theta(\bar{a}(s)|s) \right] \cdot \left[V^*(\rho) - V^{\pi_\theta}(\rho) \right], \tag{260}$$

where the last equation is again according to Lemma 19. □

Lemma 9. $\inf_{s \in \mathcal{S}, t \ge 1} \pi_{\theta_t}(a^*(s)|s) > 0.$

Proof. The proof is an extension of the proof for Lemma 5. Denote $\Delta^*(s) \coloneqq Q^*(s, a^*(s)) - \max_{a \neq a^*(s)} Q^*(s,a) > 0$ as the optimal value gap of state s, where $a^*(s)$ is the action that the optimal policy selects under state s, and $\Delta^* \coloneqq \min_{s \in \mathcal{S}} \Delta^*(s) > 0$ as the optimal value gap of the MDP. For each state $s \in \mathcal{S}$, define the following regions,

$$\mathcal{R}_1(s) \coloneqq \left\{ \theta : \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s, a^*(s))} \ge \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)}, \ \forall a \neq a^* \right\}, \tag{261}$$

$$\mathcal{R}_2(s) \coloneqq \left\{ \theta : Q^{\pi_\theta}(s, a^*(s)) \ge Q^*(s, a^*(s)) - \Delta^*(s)/2 \right\}, \tag{262}$$

$$\mathcal{R}_3(s) \coloneqq \left\{ \theta_t : V^{\pi_{\theta_t}}(s) \ge Q^{\pi_{\theta_t}}(s, a^*(s)) - \Delta^*(s)/2, \text{ for all large enough } t > 0 \right\}, \tag{263}$$

$$\mathcal{N}_c(s) \coloneqq \left\{ \theta : \pi_\theta(a^*(s)|s) \ge \frac{c(s)}{c(s)+1} \right\}, \text{ where } c(s) \coloneqq \frac{A}{(1-\gamma) \cdot \Delta^*(s)} - 1. \tag{264}$$

The proof idea also consists of similar parts.

- First, $\mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$ is a "nice" region, in the sense that, following gradient update, (i) if $\theta_t \in \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$, then $\theta_{t+1} \in \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$; (ii) $\pi_{\theta_{t+1}}(a^*(s)|s) \ge \pi_{\theta_t}(a^*(s)|s)$.
- Second, $\mathcal{N}_c(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s) \subset \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$.
- Third, there exists a finite time $t_0(s) > 0$, such that $\theta_{t_0(s)} \in \mathcal{N}_c(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$, and thus $\theta_{t_0(s)} \in \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$, which implies $\inf_{t \ge 1} \pi_{\theta_t}(a^*(s)|s) = \min_{1 \le t \le t_0(s)} \pi_{\theta_t}(a^*(s)|s)$.
- Last, define $t_0 = \max_s t_0(s)$, then we have $\inf_{s \in \mathcal{S}, t \ge 1} \pi_{\theta_t}(a^*(s)|s) = \min_{1 \le t \le t_0} \min_s \pi_{\theta_t}(a^*(s)|s)$.

First part. (i) If $\theta_t \in \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$, then $\theta_{t+1} \in \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$. Suppose $\theta_t \in \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$. We have $\theta_{t+1} \in \mathcal{R}_3(s)$ by the definition of $\mathcal{R}_3(s)$. We have,

$$Q^{\pi_{\theta_t}}(s, a^*(s)) \ge Q^*(s, a^*(s)) - \Delta^*(s)/2. \tag{265}$$

According to smoothness arguments as Eq. (309), we have $V^{\pi_{\theta_{t+1}}}(s') \ge V^{\pi_{\theta_t}}(s')$, and

$$\begin{align}
Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) &= Q^{\pi_{\theta_t}}(s, a^*(s)) + Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - Q^{\pi_{\theta_t}}(s, a^*(s)) \tag{266} \\
&= Q^{\pi_{\theta_t}}(s, a^*(s)) + \gamma \sum_{s'} \mathcal{P}(s'|s, a^*(s)) \cdot [V^{\pi_{\theta_{t+1}}}(s') - V^{\pi_{\theta_t}}(s')] \tag{267} \\
&\ge Q^{\pi_{\theta_t}}(s, a^*(s)) + 0 \ge Q^*(s, a^*(s)) - \Delta^*(s)/2, \tag{268}
\end{align}$$

which means $\theta_{t+1} \in \mathcal{R}_2(s)$. Next we prove $\theta_{t+1} \in \mathcal{R}_1(s)$. Note that $\forall a \ne a^*(s)$,

$$\begin{align}
Q^{\pi_{\theta_t}}(s, a^*(s)) - Q^{\pi_{\theta_t}}(s, a) &= Q^{\pi_{\theta_t}}(s, a^*(s)) - Q^*(s, a^*(s)) + Q^*(s, a^*(s)) - Q^{\pi_{\theta_t}}(s, a) \tag{269} \\
&\ge -\Delta^*(s)/2 + Q^*(s, a^*(s)) - Q^*(s, a) + Q^*(s, a) - Q^{\pi_{\theta_t}}(s, a) \tag{270} \\
&\ge -\Delta^*(s)/2 + Q^*(s, a^*(s)) - \max_{a \ne a^*(s)} Q^*(s, a) + Q^*(s, a) - Q^{\pi_{\theta_t}}(s, a) \tag{271} \\
&= -\Delta^*(s)/2 + \Delta^*(s) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) \cdot [V^*(s') - V^{\pi_{\theta_t}}(s')] \tag{272} \\
&\ge -\Delta^*(s)/2 + \Delta^*(s) + 0 = \Delta^*(s)/2. \tag{273}
\end{align}$$

Using similar arguments we also have $Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - Q^{\pi_{\theta_{t+1}}}(s, a) \ge \Delta^*(s)/2$. According to Lemma 1,

$$\begin{align}
\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a)} &= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a|s) \cdot A^{\pi_{\theta_t}}(s, a) \tag{274} \\
&= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a|s) \cdot \left[Q^{\pi_{\theta_t}}(s, a) - V^{\pi_{\theta_t}}(s) \right]. \tag{275}
\end{align}$$

And since $\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a^*(s))} \ge \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a)}$, we have

$$\pi_{\theta_t}(a^*(s)|s) \cdot \left[Q^{\pi_{\theta_t}}(s, a^*(s)) - V^{\pi_{\theta_t}}(s) \right] \ge \pi_{\theta_t}(a|s) \cdot \left[Q^{\pi_{\theta_t}}(s, a) - V^{\pi_{\theta_t}}(s) \right]. \tag{276}$$

Similar with the first part in the proof for Lemma 5. There are two cases.

(a) If $\pi_{\theta_t}(a^*(s)|s) \ge \pi_{\theta_t}(a|s)$, then $\theta_t(s, a^*(s)) \ge \theta_t(s, a)$. After one step gradient update,

$$\begin{align}
\theta_{t+1}(s, a^*(s)) &\leftarrow \theta_t(s, a^*(s)) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a^*(s))} \tag{277} \\
&\ge \theta_t(s, a) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a)} := \theta_{t+1}(s, a), \tag{278}
\end{align}$$

which implies $\pi_{\theta_{t+1}}(a^*(s)|s) \ge \pi_{\theta_{t+1}}(a|s)$. Since $Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - Q^{\pi_{\theta_{t+1}}}(s, a) \ge \Delta^*(s)/2 \ge 0$, $\forall a$, we have $Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - V^{\pi_{\theta_{t+1}}}(s) = Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - \sum_a \pi_{\theta_{t+1}}(a|s) \cdot Q^{\pi_{\theta_{t+1}}}(s, a) \ge 0$, and

$$\pi_{\theta_{t+1}}(a^*(s)|s) \cdot \left[Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - V^{\pi_{\theta_{t+1}}}(s) \right] \ge \pi_{\theta_{t+1}}(a|s) \cdot \left[Q^{\pi_{\theta_{t+1}}}(s, a) - V^{\pi_{\theta_{t+1}}}(s) \right]. \tag{279}$$

which is equivalent with $\frac{\partial V^{\pi_{\theta_{t+1}}}(\mu)}{\partial \theta_{t+1}(s,a^*(s))} \ge \frac{\partial V^{\pi_{\theta_{t+1}}}(\mu)}{\partial \theta_{t+1}(s,a)}$ i.e., $\theta_{t+1} \in \mathcal{R}_1(s)$.

(b) If $\pi_{\theta_t}(a^*(s)|s) < \pi_{\theta_t}(a|s)$, then by $\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} \ge \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a)}$,

$$\pi_{\theta_t}(a^*(s)|s) \cdot \left[Q^{\pi_{\theta_t}}(s,a^*(s)) - V^{\pi_{\theta_t}}(s) \right] \ge \pi_{\theta_t}(a|s) \cdot \left[Q^{\pi_{\theta_t}}(s,a) - V^{\pi_{\theta_t}}(s) \right] \tag{280}$$

$$= \pi_{\theta_t}(a|s) \cdot \left[Q^{\pi_{\theta_t}}(s,a^*(s)) - V^{\pi_{\theta_t}}(s) + Q^{\pi_{\theta_t}}(s,a) - Q^{\pi_{\theta_t}}(s,a^*(s)) \right], \tag{281}$$

which after rearranging is equivalent with

$$Q^{\pi_{\theta_t}}(s,a^*(s)) - Q^{\pi_{\theta_t}}(s,a) \ge \left(1 - \frac{\pi_{\theta_t}(a^*(s)|s)}{\pi_{\theta_t}(a|s)} \right) \cdot \left[Q^{\pi_{\theta_t}}(s,a^*(s)) - V^{\pi_{\theta_t}}(s) \right] \tag{282}$$

$$= \left(1 - \exp\left\{ \theta_t(s,a^*(s)) - \theta_t(s,a) \right\} \right) \cdot \left[Q^{\pi_{\theta_t}}(s,a^*(s)) - V^{\pi_{\theta_t}}(s) \right]. \tag{283}$$

Since $\theta_{t+1} \in \mathcal{R}_3(s)$, we have,

$$Q^{\pi_{\theta_{t+1}}}(s,a^*(s)) - V^{\pi_{\theta_{t+1}}}(s) \le \Delta^*(s)/2 \le Q^{\pi_{\theta_{t+1}}}(s,a^*(s)) - Q^{\pi_{\theta_{t+1}}}(s,a). \tag{284}$$

On the other hand,

$$\theta_{t+1}(s,a^*(s)) - \theta_{t+1}(s,a) = \theta_t(s,a^*(s)) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} - \theta_t(s,a) - \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a)} \tag{285}$$

$$\ge \theta_t(s,a^*(s)) - \theta_t(s,a), \tag{286}$$

which implies

$$1 - \exp\left\{ \theta_{t+1}(s,a^*(s)) - \theta_{t+1}(s,a) \right\} \le 1 - \exp\left\{ \theta_t(s,a^*(s)) - \theta_t(s,a) \right\}. \tag{287}$$

And since $1 - \exp\left\{ \theta_t(s,a^*(s)) - \theta_t(s,a) \right\} = 1 - \frac{\pi_{\theta_t}(a^*(s)|s)}{\pi_{\theta_t}(a|s)} > 0$ (in this case $\pi_{\theta_t}(a^*(s)|s) < \pi_{\theta_t}(a|s)$),

$$\left(1 - \exp\left\{ \theta_{t+1}(s,a^*(s)) - \theta_{t+1}(s,a) \right\} \right) \cdot \left[Q^{\pi_{\theta_{t+1}}}(s,a^*(s)) - V^{\pi_{\theta_{t+1}}}(s) \right] \le Q^{\pi_{\theta_{t+1}}}(s,a^*(s)) - Q^{\pi_{\theta_{t+1}}}(s,a), \tag{288}$$

which after rearranging is equivalent with

$$\pi_{\theta_{t+1}}(a^*(s)|s) \cdot \left[Q^{\pi_{\theta_{t+1}}}(s,a^*(s)) - V^{\pi_{\theta_{t+1}}}(s) \right] \ge \pi_{\theta_{t+1}}(a|s) \cdot \left[Q^{\pi_{\theta_{t+1}}}(s,a) - V^{\pi_{\theta_{t+1}}}(s) \right], \tag{289}$$

which means $\frac{\partial V^{\pi_{\theta_{t+1}}}(\mu)}{\partial \theta_{t+1}(s,a^*(s))} \ge \frac{\partial V^{\pi_{\theta_{t+1}}}(\mu)}{\partial \theta_{t+1}(s,a)}$ i.e., $\theta_{t+1} \in \mathcal{R}_1(s)$. Now we have (i) if $\theta_t \in \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$, then $\theta_{t+1} \in \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$.

Next we prove (ii) $\pi_{\theta_{t+1}}(a^*(s)|s) \ge \pi_{\theta_t}(a^*(s)|s)$. If $\theta_t \in \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$, then $\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} \ge \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a)}$, $\forall a \ne a^*$. After one step gradient update,

$$\pi_{\theta_{t+1}}(a^*(s)|s) = \frac{\exp\left\{ \theta_{t+1}(s,a^*(s)) \right\}}{\sum_a \exp\left\{ \theta_{t+1}(s,a) \right\}} \tag{290}$$

$$= \frac{\exp\left\{ \theta_t(s,a^*(s)) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} \right\}}{\sum_a \exp\left\{ \theta_t(s,a) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a)} \right\}} \tag{291}$$

$$\ge \frac{\exp\left\{ \theta_t(s,a^*(s)) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} \right\}}{\sum_a \exp\left\{ \theta_t(s,a) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} \right\}} \qquad \left(\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a^*(s))} \ge \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s,a)} \right) \tag{292}$$

$$= \frac{\exp\left\{ \theta_t(s,a^*(s)) \right\}}{\sum_a \exp\left\{ \theta_t(s,a) \right\}} = \pi_{\theta_t}(a^*(s)|s). \tag{293}$$

Second part. $\mathcal{N}_c(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s) \subset \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$. Suppose $\theta \in \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$ and $\pi_\theta(a^*(s)|s) \geq \frac{c(s)}{c(s)+1}$. There are two cases.

(a) If $\pi_\theta(a^*(s)|s) \geq \max_{a \neq a^*(s)}\{\pi_\theta(a|s)\}$, then we have,

$$\begin{aligned}
\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s, a^*(s))} &= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a^*(s)|s) \cdot \left[Q^{\pi_\theta}(s, a^*(s)) - V^{\pi_\theta}(s)\right] && (294)\\
&> \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot \left[Q^{\pi_\theta}(s, a) - V^{\pi_\theta}(s)\right] && (295)\\
&\coloneqq \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s, a)}, && (296)
\end{aligned}$$

where the inequality is since $Q^{\pi_\theta}(s, a^*(s)) - Q^{\pi_\theta}(s, a) \geq \Delta^*(s)/2 > 0$, $\forall a \neq a^*(s)$, similar with Eq. (269)

(b) $\pi_\theta(a^*(s)|s) < \max_{a \neq a^*(s)}\{\pi_\theta(a|s)\}$, which is not possible. Suppose there exists an $a \neq a^*(s)$, such that $\pi_\theta(a^*(s)|s) < \pi_\theta(a|s)$. Then we have the following contradiction,

$$\pi_\theta(a^*(s)|s) + \pi_\theta(a|s) > \frac{2 \cdot c(s)}{c(s)+1} = 2 - \frac{2 \cdot (1-\gamma) \cdot \Delta^*(s)}{A} > 1, \tag{297}$$

where the last inequality is according to $A \geq 2$ (there are at least two actions), and $\Delta^*(s) \leq 1/(1-\gamma)$.

Third part. (1) According to the asymptotic convergence results of Agarwal et al. (2019, Theorem 5.1), $\pi_{\theta_t}(a^*(s)|s) \to 1$. Hence, there exists $t_1(s) > 0$, such that $\pi_{\theta_{t_1(s)}}(a^*(s)|s) \geq \frac{c(s)}{c(s)+1}$. (2) $Q^{\pi_{\theta_t}}(s, a^*(s)) \to Q^*(s, a^*(s))$, as $t \to \infty$. There exists $t_2(s) > 0$, such that $Q^{\pi_{\theta_{t_2(s)}}}(s, a^*(s)) \geq Q^*(s, a^*(s)) - \Delta^*(s)/2$. (3) $Q^{\pi_{\theta_t}}(s, a^*(s)) \to V^*(s)$, and $V^{\pi_{\theta_t}}(s) \to V^*(s)$, as $t \to \infty$. There exists $t_3(s) > 0$, such that $\forall t \geq t_3(s)$, $Q^{\pi_{\theta_t}}(s, a^*(s)) - V^{\pi_{\theta_t}}(s) \leq \Delta^*(s)/2$.

Define $t_0(s) \coloneqq \max\{t_1(s), t_2(s), t_3(s)\}$. We have $\theta_{t_0(s)} \in \mathcal{N}_c(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$, and thus $\theta_{t_0(s)} \in \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$. According to the first part in our proof, i.e., once θ_t is in $\mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$, following gradient update θ_{t+1} will be in $\mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$, and $\pi_{\theta_t}(a^*(s)|s)$ is increasing in $\mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$, we have $\inf_t \pi_{\theta_t}(a^*(s)|s) = \min_{1 \leq t \leq t_0(s)} \pi_{\theta_t}(a^*(s)|s)$. $t_0(s)$ depends on initialization and $c(s)$, which only depends on the MDP and state s.

Last part. Define $t_0 = \max_s t_0(s)$. Then we have $\inf_{s \in \mathcal{S}, t \geq 1} \pi_{\theta_t}(a^*(s)|s) = \min_{1 \leq t \leq t_0} \min_s \pi_{\theta_t}(a^*(s)|s) \in \Omega(1)$. □

Theorem 4. Suppose $\mu(s) > 0$ for all state s. Using Algorithm 1 with $\eta = (1-\gamma)^3/8$ and $\pi_{\theta_1}(a^*(s)|s) \in \Omega(1)$ for every $s \in \mathcal{S}$, with some constant $C > 0$, for all $t > 0$,

$$V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \leq \frac{16SC}{(1-\gamma)^6 t} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^2 \cdot \left\| \frac{1}{\mu} \right\|_\infty. \tag{298}$$

Proof. According to the value sub-optimality lemma of Lemma 20,

$$\begin{aligned}
V^*(\rho) - V^{\pi_\theta}(\rho) &= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\theta}(s) \sum_a \left(\pi^*(a|s) - \pi_\theta(a|s)\right) \cdot Q^*(s, a) && (299)\\
&= \frac{1}{1-\gamma} \sum_s \frac{d_\rho^{\pi_\theta}(s)}{d_\mu^{\pi_\theta}(s)} \cdot d_\mu^{\pi_\theta}(s) \sum_a \left(\pi^*(a|s) - \pi_\theta(a|s)\right) \cdot Q^*(s, a) && (300)\\
&\leq \frac{1}{1-\gamma} \cdot \left\| \frac{1}{d_\mu^{\pi_\theta}} \right\|_\infty \sum_s d_\mu^{\pi_\theta}(s) \sum_a \left(\pi^*(a|s) - \pi_\theta(a|s)\right) \cdot Q^*(s, a) \qquad \left(\sum_a \left(\pi^*(a|s) - \pi_\theta(a|s)\right) \cdot Q^*(s, a) \geq 0\right) && (301)\\
&\leq \frac{1}{(1-\gamma)^2} \cdot \left\| \frac{1}{\mu} \right\|_\infty \sum_s d_\mu^{\pi_\theta}(s) \sum_a \left(\pi^*(a|s) - \pi_\theta(a|s)\right) \cdot Q^*(s, a) && (302)\\
&= \frac{1}{1-\gamma} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot \left[V^*(\mu) - V^{\pi_\theta}(\mu)\right], && (303)
\end{aligned}$$

where the last equation is again by Lemma 20, and the last inequality is according to

$$\begin{aligned} d_\mu^{\pi_\theta}(s) &\coloneqq \mathop{\mathbb{E}}_{s_0\sim\mu}\left[d_\mu^{\pi_\theta}(s)\right] && (304)\\ &= \mathop{\mathbb{E}}_{s_0\sim\mu}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr(s_t=s|s_0,\pi_\theta,\mathcal{P})\right] && (305)\\ &\ge \mathop{\mathbb{E}}_{s_0\sim\mu}\left[(1-\gamma)\Pr(s_0=s|s_0)\right] && (306)\\ &= (1-\gamma)\cdot\mu(s). && (307)\end{aligned}$$

According to Lemma 7,

$$\left|V^{\pi_{\theta_{t+1}}}(\mu) - V^{\pi_{\theta_t}}(\mu) - \left\langle \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}, \theta_{t+1}-\theta_t\right\rangle\right| \le \frac{4}{(1-\gamma)^3}\cdot\|\theta_{t+1}-\theta_t\|_2^2. \tag{308}$$

Denote $\delta_t \coloneqq V^*(\mu) - V^{\pi_{\theta_t}}(\mu)$. Then we have,

$$\begin{aligned} \delta_{t+1}-\delta_t &= V^{\pi_{\theta_t}}(\mu) - V^{\pi_{\theta_{t+1}}}(\mu) && (309)\\ &\le -\left\langle \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}, \theta_{t+1}-\theta_t\right\rangle + \frac{4}{(1-\gamma)^3}\cdot\|\theta_{t+1}-\theta_t\|_2^2 && (310)\\ &= \left(-\eta + \frac{4\eta^2}{(1-\gamma)^3}\right)\cdot\left\|\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}\right\|_2^2 \qquad \left(\theta_{t+1}=\theta_t+\eta\cdot\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}\right) && (311)\\ &= -\frac{(1-\gamma)^3}{16}\cdot\left\|\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}\right\|_2^2 \qquad \left(\eta=\frac{(1-\gamma)^3}{8}\right) && (312)\\ &\le -\frac{(1-\gamma)^3}{16S}\cdot\left\|\frac{d_\mu^{\pi^*}}{d_\mu^{\pi_{\theta_t}}}\right\|_\infty^{-2}\cdot\left[\min_s \pi_{\theta_t}(a^*(s)|s)\right]^2\cdot\left[V^*(\mu)-V^{\pi_{\theta_t}}(\mu)\right]^2 \qquad \text{(Lemma 8)} && (313)\\ &\le -\frac{(1-\gamma)^5}{16S}\cdot\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^{-2}\cdot\left[\min_s \pi_{\theta_t}(a^*(s)|s)\right]^2\cdot\delta_t^2 && (314)\\ &\le -\frac{(1-\gamma)^5}{16S}\cdot\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^{-2}\cdot\left[\inf_{s\in\mathcal{S},t\ge1} \pi_{\theta_t}(a^*(s)|s)\right]^2\cdot\delta_t^2, && (315)\end{aligned}$$

where the second last inequality is by $d_\mu^{\pi_{\theta_t}}(s) \ge (1-\gamma)\cdot\mu(s)$ similar with Eq. (304). According to Lemma 9, $\inf_{s\in\mathcal{S},t\ge1}\pi_{\theta_t}(a^*(s)|s)\in\Omega(1)>0$. Using similar induction arguments as in Eq. (87), for some constant $C>0$,

$$V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \le \frac{16SC}{(1-\gamma)^5 t}\cdot\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^2, \tag{316}$$

which leads to the final result,

$$V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \le \frac{1}{1-\gamma}\cdot\left\|\frac{1}{\mu}\right\|_\infty\cdot\left[V^*(\mu)-V^{\pi_{\theta_t}}(\mu)\right] \le \frac{16SC}{(1-\gamma)^6 t}\cdot\left\|\frac{d_\mu^{\pi^*}}{\mu}\right\|_\infty^2\cdot\left\|\frac{1}{\mu}\right\|_\infty, \tag{317}$$

where $1/C = \left[\inf_{s\in\mathcal{S},t\ge1}\pi_{\theta_t}(a^*(s)|s)\right]^2\in\Omega(1)>0$. □

A.2. Proofs for Section 4

Lemma 10. Entropy regularized policy gradient w.r.t. θ is

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot \tilde{A}^{\pi_\theta}(s,a) \tag{318}$$

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(\cdot|s) \right] \tag{319}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \theta(s,\cdot) \right], \ \forall s \tag{320}$$

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

$$\tilde{A}^{\pi_\theta}(s,a) \coloneqq \tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) - \tilde{V}^{\pi_\theta}(s) \tag{321}$$

$$\tilde{Q}^{\pi_\theta}(s,a) \coloneqq r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \tilde{V}^{\pi_\theta}(s'). \tag{322}$$

Proof. According to the definition of $\tilde{V}^{\pi_\theta}$,

$$\tilde{V}^{\pi_\theta}(\mu) = \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) \right]. \tag{323}$$

Taking derivative w.r.t. θ,

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} = \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) \right] + \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \pi_\theta(a|s) \cdot \left[\frac{\partial \tilde{Q}^{\pi_\theta}(s,a)}{\partial \theta} - \tau \frac{1}{\pi_\theta(a|s)} \frac{\partial \pi_\theta(a|s)}{\partial \theta} \right] \tag{324}$$

$$= \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) \right] + \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \pi_\theta(a|s) \cdot \frac{\partial \tilde{Q}^{\pi_\theta}(s,a)}{\partial \theta} \tag{325}$$

$$= \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) \right] + \gamma \cdot \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \pi_\theta(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \cdot \frac{\partial \tilde{V}^{\pi_\theta}(s')}{\partial \theta} \tag{326}$$

$$= \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) \right], \tag{327}$$

where the second equation is because of

$$\sum_a \pi_\theta(a|s) \cdot \left[\frac{1}{\pi_\theta(a|s)} \frac{\partial \pi_\theta(a|s)}{\partial \theta} \right] = \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} = \frac{\partial}{\partial \theta} \sum_a \pi_\theta(a|s) = \frac{\partial 1}{\partial \theta} = 0. \tag{328}$$

Using similar arguments as in the proof for Lemma 1, i.e., for $s' \neq s$, $\frac{\partial \pi_\theta(a|s)}{\partial \theta(s',\cdot)} = \mathbf{0}$,

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \left[\sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) \right] \right] \tag{329}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \left(\frac{d\pi(\cdot|s)}{d\theta(s,\cdot)} \right)^\top \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(\cdot|s) \right] \tag{330}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(\cdot|s) \right] \qquad \text{(Eq. (6))} \tag{331}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \theta(\cdot|s) + \tau \log \sum_a \exp\{\theta(s,a)\} \cdot \mathbf{1} \right] \tag{332}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \theta(\cdot|s) \right]. \qquad (H(\pi_\theta(\cdot|s))\mathbf{1} = \mathbf{0}, \text{ Lemma 21}) \tag{333}$$

For each component a, we have

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,a)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) - \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) \right] \right] \tag{334}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) - \tilde{V}^{\pi_\theta}(s) \right] \tag{335}$$

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot \tilde{A}^{\pi_\theta}(s,a). \qquad \square$$

Lemma 11. If $\theta_1 = \tilde{\theta}_1 + c \cdot \mathbf{1}$ for some constant $c \in \mathbb{R}$, then $\pi_{\theta_t} = \pi_{\tilde{\theta}_t}, \forall t \geq 1$.

Proof. Duplicate Updates 2 and 3 here for convenience.

$$\text{Update 2: } \theta_{t+1} \leftarrow \theta_t + \eta \cdot H(\pi_{\theta_t})(r - \tau \log \pi_{\theta_t}), \tag{336}$$

$$\text{Update 3: } \tilde{\theta}_{t+1} \leftarrow \tilde{\theta}_t + \eta \cdot H(\pi_{\tilde{\theta}_t})(r - \tau \log \pi_{\tilde{\theta}_t}) - \frac{\tilde{\theta}_t^\top \mathbf{1}}{K} \cdot \mathbf{1}. \tag{337}$$

Since $\theta_1 = \tilde{\theta}_1 + c \cdot \mathbf{1}$, $\pi_{\tilde{\theta}_1} = \text{softmax}(\tilde{\theta}_1) = \text{softmax}(\tilde{\theta}_1 + c \cdot \mathbf{1}) = \pi_{\theta_1}$. We prove by induction on t. Suppose $\theta_t = \tilde{\theta}_t + c_t \cdot \mathbf{1}$ for some constant $c_t \in \mathbb{R}$, for some $t \geq 1$. We have $\pi_{\theta_t} = \pi_{\tilde{\theta}_t}$. According to Update 2,

$$\theta_{t+1} \leftarrow \theta_t + \eta \cdot H(\pi_{\theta_t})(r - \tau \log \pi_{\theta_t}) \tag{338}$$

$$= \tilde{\theta}_t + c_t \cdot \mathbf{1} + \eta \cdot H(\pi_{\tilde{\theta}_t})(r - \tau \log \pi_{\tilde{\theta}_t}) \tag{339}$$

$$= \tilde{\theta}_t + \eta \cdot H(\pi_{\tilde{\theta}_t})(r - \tau \log \pi_{\tilde{\theta}_t}) - \frac{\tilde{\theta}_t^\top \mathbf{1}}{K} \cdot \mathbf{1} + \left(c_t + \frac{\tilde{\theta}_t^\top \mathbf{1}}{K} \right) \cdot \mathbf{1} \tag{340}$$

$$= \tilde{\theta}_{t+1} + \frac{\theta_t^\top \mathbf{1}}{K} \cdot \mathbf{1}, \tag{341}$$

which means $\theta_{t+1} = \tilde{\theta}_{t+1} + c_{t+1} \cdot \mathbf{1}$ for constant $c_{t+1} := \frac{\theta_t^\top \mathbf{1}}{K}$, and $\pi_{\theta_{t+1}} = \pi_{\tilde{\theta}_{t+1}}$. $\square$

Lemma 12 (Non-uniform contraction)**.** Using Update 3 with $\tau\eta \leq 1, \forall t > 0$,

$$\|\zeta_{t+1}\|_2 \leq \left(1 - \tau\eta \cdot \min_a \pi_{\tilde{\theta}_t}(a)\right) \cdot \|\zeta_t\|_2, \tag{342}$$

where $\zeta_t := \tau\tilde{\theta}_t - r - \frac{(\tau\tilde{\theta}_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}$.

Proof. Update 3 can be written as

$$\tilde{\theta}_{t+1} \leftarrow \tilde{\theta}_t - \eta \cdot H(\pi_{\tilde{\theta}_t})(\tau \log \pi_{\tilde{\theta}_t} - r) - \frac{\tilde{\theta}_t^\top \mathbf{1}}{K} \cdot \mathbf{1} \tag{343}$$

$$= \tilde{\theta}_t - \eta \cdot H(\pi_{\tilde{\theta}_t}) \left[\tau\tilde{\theta}_t - r - \left(\log \sum_a \exp\{\tilde{\theta}_t(a)\} \right) \cdot \mathbf{1} \right] - \frac{\tilde{\theta}_t^\top \mathbf{1}}{K} \cdot \mathbf{1} \tag{344}$$

$$= \tilde{\theta}_t - \eta \cdot H(\pi_{\tilde{\theta}_t})(\tau\tilde{\theta}_t - r) - \frac{\tilde{\theta}_t^\top \mathbf{1}}{K} \cdot \mathbf{1} \tag{345}$$

$$= \tilde{\theta}_t - \eta \cdot H(\pi_{\tilde{\theta}_t}) \left(\tau\tilde{\theta}_t - r - \frac{(\tau\tilde{\theta}_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right) - \frac{\tilde{\theta}_t^\top \mathbf{1}}{K} \cdot \mathbf{1}, \tag{346}$$

where the last two equations are from $H(\pi_{\tilde{\theta}_t})\mathbf{1} = \mathbf{0}$ as shown in Lemma 21. For all $t \ge 1$,

$$\zeta_{t+1} \coloneqq \tau \tilde{\theta}_{t+1} - r - \frac{(\tau\tilde{\theta}_{t+1} - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \tag{347}$$

$$= \tau \tilde{\theta}_{t} - r - \frac{(\tau\tilde{\theta}_{t} - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} + \tau(\tilde{\theta}_{t+1} - \tilde{\theta}_t) + \left(\frac{(\tau\tilde{\theta}_{t} - r)^\top \mathbf{1}}{K} - \frac{(\tau\tilde{\theta}_{t+1} - r)^\top \mathbf{1}}{K} \right) \cdot \mathbf{1} \tag{348}$$

$$= \tau \tilde{\theta}_{t} - r - \frac{(\tau\tilde{\theta}_{t} - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} + \tau(\tilde{\theta}_{t+1} - \tilde{\theta}_t) + \frac{\tau(\tilde{\theta}_t - \tilde{\theta}_{t+1})^\top \mathbf{1}}{K} \cdot \mathbf{1}. \tag{349}$$

For the last term,

$$\frac{\tau(\tilde{\theta}_t - \tilde{\theta}_{t+1})^\top \mathbf{1}}{K} \cdot \mathbf{1} = \frac{\tau}{K} \cdot \left(\eta \cdot H(\pi_{\tilde{\theta}_t}) \left(\tau\tilde{\theta}_t - r - \frac{(\tau\tilde{\theta}_{t} - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right) + \frac{\tilde{\theta}_t^\top \mathbf{1}}{K} \cdot \mathbf{1} \right)^\top \mathbf{1} \cdot \mathbf{1} \tag{350}$$

$$= \frac{\tau}{K} \cdot \frac{\tilde{\theta}_t^\top \mathbf{1}}{K} \cdot \mathbf{1}^\top \mathbf{1} \cdot \mathbf{1} = \tau \cdot \frac{\tilde{\theta}_t^\top \mathbf{1}}{K} \cdot \mathbf{1}, \tag{351}$$

where the first equation is again by $H(\pi_{\tilde{\theta}_t})^\top \mathbf{1} = H(\pi_{\tilde{\theta}_t})\mathbf{1} = \mathbf{0}$. Using the update rule and combining the above,

$$\zeta_{t+1} = \left(\mathbf{Id} - \tau\eta \cdot H(\pi_{\tilde{\theta}_t}) \right) \left(\tau\tilde{\theta}_t - r - \frac{(\tau\tilde{\theta}_{t} - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right) - \tau \cdot \frac{\tilde{\theta}_t^\top \mathbf{1}}{K} \cdot \mathbf{1} + \tau \cdot \frac{\tilde{\theta}_t^\top \mathbf{1}}{K} \cdot \mathbf{1} \tag{352}$$

$$\coloneqq \left(\mathbf{Id} - \tau\eta \cdot H(\pi_{\tilde{\theta}_t}) \right) \zeta_t. \tag{353}$$

According to Lemma 22, with $\tau\eta \le 1$,

$$\|\zeta_{t+1}\|_2 = \left\| \left(\mathbf{Id} - \tau\eta \cdot H(\pi_{\tilde{\theta}_t}) \right) \zeta_t \right\|_2 \tag{354}$$

$$\le \left(1 - \tau\eta \cdot \min_a \pi_{\tilde{\theta}_t}(a) \right) \cdot \|\zeta_t\|_2. \qquad \square$$

Lemma 13. Let $\pi_{\tilde{\theta}_t} \coloneqq \mathrm{softmax}(\tilde{\theta}_t)$. Using Update 3 with $\tau\eta \le 1$, $\forall t > 0$,

$$\|\zeta_t\|_2 \le \frac{2(\tau C + 1)\sqrt{K}}{\exp\left\{ \tau\eta \sum_{s=1}^{t-1} \left[\min_a \pi_{\tilde{\theta}_s}(a) \right] \right\}}, \tag{355}$$

where we assume $\|\tilde{\theta}_1\|_\infty \le C$ for some constant $C > 0$.

Proof. According to Lemma 12, for all $t > 0$,

$$\|\zeta_{t+1}\|_2 \le \left(1 - \tau\eta \cdot \min_a \pi_{\tilde{\theta}_t}(a) \right) \cdot \|\zeta_t\|_2 \tag{356}$$

$$\le \frac{1}{\exp\left\{ \tau\eta \cdot \min_a \pi_{\tilde{\theta}_t}(a) \right\}} \cdot \|\zeta_t\|_2 \tag{357}$$

$$\le \frac{1}{\exp\left\{ \tau\eta \cdot \min_a \pi_{\tilde{\theta}_t}(a) \right\}} \cdot \left(1 - \tau\eta \cdot \min_a \pi_{\tilde{\theta}_{t-1}}(a) \right) \cdot \|\zeta_{t-1}\|_2 \tag{358}$$

$$\le \frac{1}{\exp\left\{ \tau\eta \sum_{s=t-1}^{t} \left[\min_a \pi_{\tilde{\theta}_s}(a) \right] \right\}} \cdot \|\zeta_{t-1}\|_2 \tag{359}$$

$$\le \frac{1}{\exp\left\{ \tau\eta \sum_{s=1}^{t} \left[\min_a \pi_{\tilde{\theta}_s}(a) \right] \right\}} \cdot \|\zeta_1\|_2. \tag{360}$$

For initialized logit $\tilde{\theta}_1$,

$$\|\zeta_1\|_2 \coloneqq \left\| \tau\tilde{\theta}_1 - r - \frac{(\tau\tilde{\theta}_1 - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_2 \tag{361}$$

$$\le \|\tau\tilde{\theta}_1 - r\|_2 + \left\| \frac{(\tau\tilde{\theta}_1 - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_2 \tag{362}$$

$$= \|\tau\tilde{\theta}_1 - r\|_2 + \frac{\left|(\tau\tilde{\theta}_1 - r)^\top \mathbf{1}\right|}{\sqrt{K}} \tag{363}$$

$$\le \|\tau\tilde{\theta}_1 - r\|_2 + \frac{\|\tau\tilde{\theta}_1 - r\|_2 \cdot \|\mathbf{1}\|_2}{\sqrt{K}} \tag{364}$$

$$= 2 \cdot \|\tau\tilde{\theta}_1 - r\|_2 \tag{365}$$

$$\le 2 \cdot \left(\|\tau\tilde{\theta}_1\|_2 + \|r\|_2 \right) \tag{366}$$

$$\le 2(\tau C + 1)\sqrt{K}, \tag{367}$$

where the last inequality is by assuming $\|\tilde{\theta}_1\|_\infty \le C$ for some constant $C > 0$. □

Lemma 14. $\min_a \pi_{\tilde{\theta}_t}(a) / \min_a \pi_{\tilde{\theta}_1}(a) \in \Omega(1)$, for $t > 0$. Thus, $\sum_{s=1}^{t-1} \left[\min_a \pi_{\tilde{\theta}_s}(a)\right] \in \Omega(t)$.

Proof. We prove $\min_a \pi_{\tilde{\theta}_t}(a) \in \Omega(1)$ by induction. Suppose $\|\tilde{\theta}_1\|_\infty \le C$ for constant $C > 0$. According to Eq. (361),

$$\|\zeta_1\|_2 \le 2(\tau C + 1)\sqrt{K}, \tag{368}$$

where $\zeta_t \coloneqq \tau\tilde{\theta}_t - r - \frac{(\tau\tilde{\theta}_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}$, $\forall t \ge 1$. Suppose $\|\zeta_t\|_2 \le 2(\tau C + 1)\sqrt{K}$ for some $t \ge 1$. We have, $\forall a$,

$$\left| \tilde{\theta}_t(a) - \frac{r(a)}{\tau} - \frac{(\tilde{\theta}_t - r/\tau)^\top \mathbf{1}}{K} \right| = \frac{1}{\tau} \cdot \left| \tau\tilde{\theta}_t(a) - r(a) - \frac{(\tau\tilde{\theta}_t - r)^\top \mathbf{1}}{K} \right| \tag{369}$$

$$\le \frac{1}{\tau} \cdot \left\| \tau\tilde{\theta}_t - r - \frac{(\tau\tilde{\theta}_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_2 \tag{370}$$

$$= \frac{1}{\tau} \cdot \|\zeta_t\|_2 \tag{371}$$

$$\le 2(C + 1/\tau)\sqrt{K}. \tag{372}$$

Denote $a_1 \coloneqq \arg\min_a \tilde{\theta}_t(a)$, and $a_2 \coloneqq \arg\max_a \tilde{\theta}_t(a)$. According to the above, we have the following results,

$$\tilde{\theta}_t(a_1) \ge \frac{r(a_1)}{\tau} + \frac{(\tau\tilde{\theta}_t - r)^\top \mathbf{1}}{K} - 2(C + 1/\tau)\sqrt{K} \tag{373}$$

$$-\tilde{\theta}_t(a_2) \ge -\frac{r(a_2)}{\tau} - \frac{(\tau\tilde{\theta}_t - r)^\top \mathbf{1}}{K} - 2(C + 1/\tau)\sqrt{K}, \tag{374}$$

which can be used to lower bound the minimum probability as,

$$\min_a \pi_{\tilde{\theta}_t}(a) = \frac{\exp\{\tilde{\theta}_t(a_1)\}}{\sum_a \exp\{\tilde{\theta}_t(a)\}} \ge \frac{\exp\{\tilde{\theta}_t(a_1)\}}{\sum_a \exp\{\tilde{\theta}_t(a_2)\}} = \frac{1}{K} \cdot \exp\left\{\tilde{\theta}_t(a_1) - \tilde{\theta}_t(a_2)\right\}, \quad \left(\tilde{\theta}_t(a) \le \tilde{\theta}_t(a_2),\ \forall a\right) \tag{375}$$

which can be further lower bounded using the above results,

$$\begin{aligned}\min_a \pi_{\tilde{\theta}_t}(a) &\ge \frac{1}{K} \cdot \exp\left\{ \tilde{\theta}_t(a_1) - \tilde{\theta}_t(a_2) \right\} && (376)\\ &\ge \frac{1}{K} \cdot \exp\left\{ \frac{r(a_1)}{\tau} + \frac{(\tau\tilde{\theta}_t - r)^\top \mathbf{1}}{K} - 2(C + 1/\tau)\sqrt{K} - \frac{r(a_2)}{\tau} - \frac{(\tau\tilde{\theta}_t - r)^\top \mathbf{1}}{K} - 2(C + 1/\tau)\sqrt{K} \right\} && (377)\\ &= \frac{1}{K} \cdot \exp\left\{ \frac{r(a_1) - r(a_2)}{\tau} - 4(C + 1/\tau)\sqrt{K} \right\} && (378)\\ &\ge \frac{1}{K} \cdot \exp\left\{ -\frac{1}{\tau} - 4(C + 1/\tau)\sqrt{K} \right\} \qquad \left(r \in [0,1]^K,\ r(a_1) - r(a_2) \ge -1 \right) && (379)\\ &= \frac{1}{K} \cdot \frac{1}{\exp\{1/\tau\}} \cdot \frac{1}{\exp\{4(C + 1/\tau)\sqrt{K}\}} > 0. && (380)\end{aligned}$$

According to Lemma 12, with $\tau\eta \le 1$,

$$\|\zeta_{t+1}\|_2 \le \left(1 - \tau\eta \cdot \min_a \pi_{\tilde{\theta}_t}(a) \right) \cdot \|\zeta_t\|_2 < 2(\tau C + 1)\sqrt{K}. \tag{381}$$

Therefore, for all $t > 0$, we have

$$\min_a \pi_{\tilde{\theta}_t}(a) \ge \frac{1}{K} \cdot \frac{1}{\exp\{1/\tau\}} \cdot \frac{1}{\exp\{4(C + 1/\tau)\sqrt{K}\}} \in \Omega(1), \tag{382}$$

and thus $\sum_{s=1}^{t-1} \left[\min_a \pi_{\tilde{\theta}_s}(a) \right] \in \Omega(t)$. □

Theorem 5. Let $\pi_{\theta_t} \coloneqq \mathrm{softmax}(\theta_t)$. Using Update 2 with $\eta \le 1/\tau$, for all $t > 0$,

$$\left(\pi_\tau^* - \pi_{\theta_t}\right)^\top r \le \frac{4K^{3/2}(C + 1/\tau)}{\exp\left\{\tau\eta \cdot \Omega(1) \cdot t\right\}}, \tag{383}$$

$$\tilde{\delta}_t \le \frac{2(\tau C + 1)^2 K/\tau}{\exp\left\{2\tau\eta \cdot \Omega(1) \cdot t\right\}}, \tag{384}$$

where $\tilde{\delta}_t \coloneqq {\pi_\tau^*}^\top \left(r - \tau \log \pi_\tau^* \right) - {\pi_{\theta_t}}^\top \left(r - \tau \log \pi_{\theta_t} \right)$.

Proof. According to Hölder's inequality,

$$\begin{aligned}\left(\pi_\tau^* - \pi_{\theta_t}\right)^\top r &\le \left\|\pi_\tau^* - \pi_{\theta_t}\right\|_\infty \cdot \|r\|_1 && (385)\\ &\le K \cdot \left\|\pi_\tau^* - \pi_{\theta_t}\right\|_\infty \qquad \left(r \in [0,1]^K \right) && (386)\\ &= K \cdot \left\|\pi_\tau^* - \pi_{\tilde{\theta}_t}\right\|_\infty \qquad (\text{Lemma 11}) && (387)\\ &= K \cdot \left\| \mathrm{softmax}\left(\frac{r}{\tau}\right) - \mathrm{softmax}\left(\tilde{\theta}_t + \frac{(\tau\tilde{\theta}_t - r)^\top \mathbf{1}}{\tau K} \cdot \mathbf{1} \right) \right\|_\infty && (388)\\ &\le \frac{2K}{\tau} \cdot \left\| \tau\tilde{\theta}_t - r - \frac{(\tau\tilde{\theta}_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty \qquad (\text{Lemma 23}) && (389)\\ &\le \frac{2K}{\tau} \cdot \left\| \tau\tilde{\theta}_t - r - \frac{(\tau\tilde{\theta}_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_2 && (390)\\ &\le \frac{2K}{\tau} \cdot \frac{2(\tau C + 1)\sqrt{K}}{\exp\left\{ \tau\eta \sum_{s=1}^{t-1} \left[\min_a \pi_{\tilde{\theta}_s}(a) \right] \right\}} \qquad (\text{Lemma 13}) && (391)\\ &\le \frac{4K^{3/2}}{\tau} \cdot \frac{\tau C + 1}{\exp\left\{\tau\eta \cdot \Omega(1) \cdot t\right\}}. \qquad (\text{Lemma 14}) && (392)\end{aligned}$$

On the other hand, we have,

$$\begin{aligned}
{\pi_\tau^*}^\top \left(r - \tau \log \pi_\tau^* \right) - {\pi_{\theta_t}}^\top \left(r - \tau \log \pi_{\theta_t} \right) &= {\pi_\tau^*}^\top \left(r - \tau \log \pi_\tau^* \right) - {\pi_{\theta_t}}^\top \left(r - \tau \log \pi_\tau^* + \tau \log \pi_\tau^* - \tau \log \pi_{\theta_t} \right) && (393) \\
&= \left(\pi_\tau^* - \pi_{\theta_t} \right)^\top \left(r - \tau \log \pi_\tau^* \right) + \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t} \| \pi_\tau^*) && (394) \\
&= \left(\pi_\tau^* - \pi_{\theta_t} \right)^\top \mathbf{1} \cdot \tau \cdot \log \sum_a \exp\{ r(a)/\tau \} + \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t} \| \pi_\tau^*) && (395) \\
&= \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t} \| \pi_\tau^*) && (396) \\
&= \tau \cdot D_{\mathrm{KL}}(\pi_{\tilde{\theta}_t} \| \pi_\tau^*) \qquad \text{(Lemma 11)} && (397) \\
&\le \frac{\tau}{2} \cdot \left\| \tilde{\theta}_t - \frac{r}{\tau} - \frac{(\tau \tilde{\theta}_t - r)^\top \mathbf{1}}{\tau K} \cdot \mathbf{1} \right\|_\infty^2 \qquad \text{(Lemma 25)} && (398) \\
&= \frac{1}{2\tau} \cdot \left\| \tau \tilde{\theta}_t - r - \frac{(\tau \tilde{\theta}_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty^2 && (399) \\
&\le \frac{1}{2\tau} \cdot \left\| \tau \tilde{\theta}_t - r - \frac{(\tau \tilde{\theta}_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_2^2 && (400) \\
&\le \frac{1}{2\tau} \cdot \frac{4(\tau C + 1)^2 K}{\exp\left\{ 2 \tau \eta \sum_{s=1}^{t-1} \left[\min_a \pi_{\tilde{\theta}_s}(a) \right] \right\}} \qquad \text{(Lemma 13)} && (401) \\
&\le \frac{1}{\tau} \cdot \frac{2(\tau C + 1)^2 K}{\exp\left\{ 2 \tau \eta \cdot \Omega(1) \cdot t \right\}}. \qquad \text{(Lemma 14)} && \square
\end{aligned}$$

Lemma 15 **(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.

Proof. Denote $\mathbb{H}^{\pi_\theta}(s) \coloneqq \mathbb{H}(s, \pi_\theta)$. Also denote $\theta_\alpha \coloneqq \theta + \alpha u$, where $\alpha \in \mathbb{R}$ and $u \in \mathbb{R}^{SA}$. According to Eq. (17),

$$\begin{aligned}
\mathbb{H}^{\pi_{\theta_\alpha}}(s) &= \mathop{\mathbb{E}}_{\substack{s_0 = s, a_t \sim \pi_{\theta_\alpha}(\cdot | s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot | s_t, a_t)}} \left[\sum_{t=0}^\infty -\gamma^t \log \pi_{\theta_\alpha}(a_t | s_t) \right] && (402) \\
&= - \sum_a \pi_{\theta_\alpha}(a|s) \cdot \log \pi_{\theta_\alpha}(a|s) + \gamma \sum_a \pi_{\theta_\alpha}(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \cdot \mathbb{H}^{\pi_{\theta_\alpha}}(s'), && (403)
\end{aligned}$$

which implies,

$$\mathbb{H}^{\pi_{\theta_\alpha}}(s) = e_s^\top M(\alpha) h_{\theta_\alpha}, \tag{404}$$

where $M(\alpha) \coloneqq \left(\mathbf{Id} - \gamma P(\alpha) \right)^{-1}$ is defined in Eq. (184), $P(\alpha)$ is defined in Eq. (170), and $h_{\theta_\alpha} \in \mathbb{R}^S, \forall s$,

$$h_{\theta_\alpha}(s) \coloneqq - \sum_a \pi_{\theta_\alpha}(a|s) \cdot \log \pi_{\theta_\alpha}(a|s). \tag{405}$$

According to Eq. (405), $h_{\theta_\alpha}(s) \in [0, \log A], \forall s$. Then we have,

$$\| h_{\theta_\alpha} \|_\infty = \max_s | h_{\theta_\alpha}(s) | \le \log A. \tag{406}$$

For any state $s \in \mathcal{S}$,

$$\begin{aligned}
\left| \frac{\partial h_{\theta_\alpha}(s)}{\partial \alpha} \right| &= \left| \left\langle \frac{\partial h_{\theta_\alpha}(s)}{\partial \theta_\alpha}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| && (407) \\
&= \left| \left\langle \frac{\partial h_{\theta_\alpha}(s)}{\partial \theta_\alpha(\cdot|s)}, u(s, \cdot) \right\rangle \right| && (408) \\
&= \left| \left\langle H(\pi_{\theta_\alpha}(\cdot|s)) \log \pi_{\theta_\alpha}(\cdot|s), u(s,\cdot) \right\rangle \right| && (409) \\
&\le \left\| H(\pi_{\theta_\alpha}(\cdot|s)) \log \pi_{\theta_\alpha}(\cdot|s) \right\|_1 \cdot \| u(s, \cdot) \|_\infty . && (410)
\end{aligned}$$

The ℓ_1 norm is upper bounded as

$$\left\| H(\pi_{\theta_\alpha}(\cdot|s)) \log \pi_{\theta_\alpha}(\cdot|s) \right\|_1 = \sum_a \pi_{\theta_\alpha}(a|s) \cdot \left| \log \pi_{\theta_\alpha}(a|s) - \pi_{\theta_\alpha}(\cdot|s)^\top \log \pi_{\theta_\alpha}(\cdot|s) \right| \tag{411}$$

$$\leq \sum_a \pi_{\theta_\alpha}(a|s) \cdot \left(\left| \log \pi_{\theta_\alpha}(a|s) \right| + \left| \pi_{\theta_\alpha}(\cdot|s)^\top \log \pi_{\theta_\alpha}(\cdot|s) \right| \right) \tag{412}$$

$$= -2 \cdot \sum_a \pi_{\theta_\alpha}(a|s) \cdot \log \pi_{\theta_\alpha}(a|s) \leq 2 \cdot \log A. \tag{413}$$

Therefore we have,

$$\left\| \frac{\partial h_{\theta_\alpha}}{\partial \alpha} \right\|_\infty = \max_s \left| \frac{\partial h_{\theta_\alpha}(s)}{\partial \alpha} \right| \tag{414}$$

$$\leq \max_s \left\| H(\pi_{\theta_\alpha}(\cdot|s)) \log \pi_{\theta_\alpha}(\cdot|s) \right\|_1 \cdot \| u(s, \cdot) \|_\infty \tag{415}$$

$$\leq 2 \cdot \log A \cdot \| u \|_2. \tag{416}$$

The second derivative w.r.t. α is

$$\left| \frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \alpha^2} \right| = \left| \left(\frac{\partial}{\partial \theta_\alpha} \left\{ \frac{\partial h_{\theta_\alpha}(s)}{\partial \alpha} \right\} \right)^\top \frac{\partial \theta_\alpha}{\partial \alpha} \right| \tag{417}$$

$$= \left| \left(\frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \theta_\alpha^2} \frac{\partial \theta_\alpha}{\partial \alpha} \right)^\top \frac{\partial \theta_\alpha}{\partial \alpha} \right| \tag{418}$$

$$= \left| u(s, \cdot)^\top \frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \theta_\alpha^2(s, \cdot)} u(s, \cdot) \right|. \tag{419}$$

Denote the Hessian $T(s, \theta_\alpha) \coloneqq \frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \theta^2(s, \cdot)}$.

$$T(s, \theta_\alpha) \coloneqq \frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \theta_\alpha^2(s, \cdot)} = \frac{\partial}{\partial \theta_\alpha(s, \cdot)} \left\{ \frac{\partial h_{\theta_\alpha}(s)}{\partial \theta_\alpha(s, \cdot)} \right\} \tag{420}$$

$$= \frac{\partial}{\partial \theta_\alpha(s, \cdot)} \left\{ \left(\frac{\partial \pi_{\theta_\alpha}(\cdot|s)}{\partial \theta_\alpha(s, \cdot)} \right)^\top \frac{\partial h_{\theta_\alpha}(s)}{\partial \pi_{\theta_\alpha}(\cdot|s)} \right\} \tag{421}$$

$$= \frac{\partial}{\partial \theta_\alpha(s, \cdot)} \left\{ H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s)) \right\}. \tag{422}$$

Note $T(s, \theta_\alpha) \in \mathbb{R}^{A \times A}$, and $\forall i, j \in \mathcal{A}$, the value of $T(s, \theta_\alpha)$ is,

$$T_{i,j} = \frac{d\{\pi_{\theta_\alpha}(i|s) \cdot (-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s))\}}{d\theta_\alpha(s, j)} \tag{423}$$

$$= \frac{d\pi_{\theta_\alpha}(i|s)}{d\theta_\alpha(s, j)} \cdot (-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s)) + \pi_{\theta_\alpha}(i|s) \cdot \frac{d\{-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s)\}}{d\theta_\alpha(s, j)} \tag{424}$$

$$= (\delta_{ij} \pi_{\theta_\alpha}(j|s) - \pi_{\theta_\alpha}(i|s) \pi_{\theta_\alpha}(j|s)) \cdot (-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s)) \tag{425}$$

$$+ \pi_{\theta_\alpha}(i|s) \cdot \left(-\frac{1}{\pi_{\theta_\alpha}(i|s)} \cdot (\delta_{ij} \pi_{\theta_\alpha}(j|s) - \pi_{\theta_\alpha}(i|s) \pi_{\theta_\alpha}(j|s)) - \pi_{\theta_\alpha}(j|s) \cdot (-\log \pi_{\theta_\alpha}(j|s) - h_{\theta_\alpha}(s)) \right) \tag{426}$$

$$= \delta_{ij} \pi_{\theta_\alpha}(j|s) \cdot (-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s) - 1) - \pi_{\theta_\alpha}(i|s) \pi_{\theta_\alpha}(j|s) \cdot (-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s) - 1) \tag{427}$$

$$- \pi_{\theta_\alpha}(i|s) \pi_{\theta_\alpha}(j|s) \cdot (-\log \pi_{\theta_\alpha}(j|s) - h_{\theta_\alpha}(s)). \tag{428}$$

For any vector $y \in \mathbb{R}^A$,

$$\begin{aligned}
\left| y^\top T(s, \theta_\alpha) y \right| &= \left| \sum_{i=1}^{A} \sum_{j=1}^{A} T_{i,j} y(i) y(j) \right| \qquad (429) \\
&\le \left| \sum_i \pi_{\theta_\alpha}(i|s) \cdot \left(- \log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s) - 1 \right) \cdot y(i)^2 \right| \qquad (430) \\
&\qquad + 2 \cdot \left| \sum_i \pi_{\theta_\alpha}(i|s) \cdot y(i) \sum_j \pi_{\theta_\alpha}(j|s) \cdot \left(- \log \pi_{\theta_\alpha}(j|s) - h_{\theta_\alpha}(s) \right) \cdot y(j) \right| + \left(\pi_{\theta_\alpha}(\cdot|s)^\top y \right)^2 \qquad (431) \\
&= \left| \left(H(\pi_{\theta_\alpha}(\cdot|s)) (- \log \pi_{\theta_\alpha}(\cdot|s)) - \pi_{\theta_\alpha}(\cdot|s) \right)^\top (y \odot y) \right| \qquad (432) \\
&\qquad + 2 \cdot \left| \left(\pi_{\theta_\alpha}(\cdot|s)^\top y \right) \cdot \left(H(\pi_{\theta_\alpha}(\cdot|s)) (- \log \pi_{\theta_\alpha}(\cdot|s)) \right)^\top y \right| + \left(\pi_{\theta_\alpha}(\cdot|s)^\top y \right)^2 \qquad (433) \\
&\le \left\| H(\pi_{\theta_\alpha}(\cdot|s)) (- \log \pi_{\theta_\alpha}(\cdot|s)) \right\|_\infty \cdot \left\| y \odot y \right\|_1 + \left\| \pi_{\theta_\alpha}(\cdot|s) \right\|_\infty \cdot \left\| y \odot y \right\|_1 \qquad (434) \\
&\qquad + 2 \cdot \left\| \pi_{\theta_\alpha}(\cdot|s) \right\|_1 \cdot \| y \|_\infty \cdot \left\| H(\pi_{\theta_\alpha}(\cdot|s)) (- \log \pi_{\theta_\alpha}(\cdot|s)) \right\|_1 \cdot \| y \|_\infty + \left\| \pi_{\theta_\alpha}(\cdot|s) \right\|_2^2 \cdot \| y \|_2^2, \qquad (435)
\end{aligned}$$

where the last inequality is by Hölder's inequality. Note that $\| y \odot y \|_1 = \| y \|_2^2$, $\| \pi_{\theta_\alpha}(\cdot|s) \|_\infty \le \| \pi_{\theta_\alpha}(\cdot|s) \|_1$, $\| \pi_{\theta_\alpha}(\cdot|s) \|_2 \le \| \pi_{\theta_\alpha}(\cdot|s) \|_1 = 1$, and $\| y \|_\infty \le \| y \|_2$. The ℓ_∞ norm is upper bounded as

$$\begin{aligned}
\left\| H(\pi_{\theta_\alpha}(\cdot|s)) (- \log \pi_{\theta_\alpha}(\cdot|s)) \right\|_\infty &= \max_a \left| \pi_{\theta_\alpha}(a|s) \cdot \left(- \log \pi_{\theta_\alpha}(a|s) + \pi_{\theta_\alpha}(\cdot|s)^\top \log \pi_{\theta_\alpha}(\cdot|s) \right) \right| \qquad (436) \\
&\le \max_a - \pi_{\theta_\alpha}(a|s) \cdot \log \pi_{\theta_\alpha}(a|s) - \pi_{\theta_\alpha}(\cdot|s)^\top \log \pi_{\theta_\alpha}(\cdot|s) \qquad (437) \\
&\le \frac{1}{e} + \log A. \qquad \left(- x \cdot \log x \le \frac{1}{e} \text{ for } x \in [0, 1] \right) \qquad (438)
\end{aligned}$$

Therefore we have,

$$\left| y^\top T(s, \theta_\alpha) y \right| \le \left\| H(\pi_{\theta_\alpha}(\cdot|s)) (- \log \pi_{\theta_\alpha}(\cdot|s)) \right\|_\infty \cdot \| y \|_2^2 + \| y \|_2^2 + 2 \cdot \left\| H(\pi_{\theta_\alpha}(\cdot|s)) (- \log \pi_{\theta_\alpha}(\cdot|s)) \right\|_1 \cdot \| y \|_2^2 + \| y \|_2^2 \tag{439}$$

$$\begin{aligned}
&\le \left(\frac{1}{e} + \log A + 2 \right) \cdot \| y \|_2^2 + 2 \cdot \left\| H(\pi_{\theta_\alpha}(\cdot|s)) (- \log \pi_{\theta_\alpha}(\cdot|s)) \right\|_1 \cdot \| y \|_2^2 \qquad (\text{Eq. } (436)) \qquad (440) \\
&\le \left(\frac{1}{e} + \log A + 2 + 2 \cdot \log A \right) \cdot \| y \|_2^2 \qquad (\text{Eq. } (411)) \qquad (441) \\
&\le 3 \cdot (1 + \log A) \cdot \| y \|_2^2. \qquad (442)
\end{aligned}$$

According to the above results,

$$\begin{aligned}
\left\| \frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2} \right\|_\infty &= \max_s \left| \frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \alpha^2} \right| \qquad (443) \\
&= \max_s \left| u(s, \cdot)^\top \frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \theta_\alpha^2(s, \cdot)} u(s, \cdot) \right| \qquad (444) \\
&= \max_s \left| u(s, \cdot)^\top T(s, \theta_\alpha) u(s, \cdot) \right| \qquad (445) \\
&\le 3 \cdot (1 + \log A) \cdot \max_s \| u(s, \cdot) \|_2^2 \qquad (446) \\
&\le 3 \cdot (1 + \log A) \cdot \| u \|_2^2. \qquad (447)
\end{aligned}$$

Taking derivative w.r.t. α in Eq. (404),

$$\frac{\partial \mathbb{H}^{\pi_{\theta_\alpha}}(s)}{\partial \alpha} = \gamma \cdot e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) h_{\theta_\alpha} + e_s^\top M(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial \alpha}. \tag{448}$$

Taking second derivative w.r.t. α,

$$
\begin{align}
\frac{\partial^2 \mathbb{H}^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2} &= 2\gamma^2 \cdot e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) h_{\theta_\alpha} + \gamma \cdot e_s^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) h_{\theta_\alpha} \tag{449} \\
&\quad + 2\gamma \cdot e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial \alpha} + e_s^\top M(\alpha) \frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2}. \tag{450}
\end{align}
$$

For the last term,

$$
\begin{align}
\left| e_s^\top M(\alpha) \frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0} \right| &\le \left\| e_s \right\|_1 \cdot \left\| M(\alpha) \frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0} \right\|_\infty \tag{451} \\
&\le \frac{1}{1-\gamma} \cdot \left\| \frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0} \right\|_\infty \qquad \text{(Eq. (189))} \tag{452} \\
&\le \frac{3 \cdot (1 + \log A)}{1-\gamma} \cdot \|u\|_2^2. \qquad \text{(Eq. (443))} \tag{453}
\end{align}
$$

For the second last term,

$$
\begin{align}
\left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right| &\le \left\| M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty \tag{454} \\
&\le \frac{1}{1-\gamma} \cdot \left\| \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty \qquad \text{(Eq. (189))} \tag{455} \\
&\le \frac{2 \cdot \|u\|_2}{1-\gamma} \cdot \left\| M(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty \qquad \text{(Eq. (173))} \tag{456} \\
&\le \frac{2 \cdot \|u\|_2}{(1-\gamma)^2} \cdot \left\| \frac{\partial h_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty \qquad \text{(Eq. (189))} \tag{457} \\
&\le \frac{2 \cdot \|u\|_2}{(1-\gamma)^2} \cdot 2 \cdot \log A \cdot \|u\|_2 = \frac{4 \cdot \log A}{(1-\gamma)^2} \cdot \|u\|_2^2. \qquad \text{(Eq. (414))} \tag{458}
\end{align}
$$

For the second term,

$$
\begin{align}
\left| e_s^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right| &\le \left\| M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \tag{459} \\
&\le \frac{1}{1-\gamma} \cdot \left\| \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \qquad \text{(Eq. (189))} \tag{460} \\
&\le \frac{6 \cdot \|u\|_2^2}{1-\gamma} \cdot \left\| M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \qquad \text{(Eq. (178))} \tag{461} \\
&\le \frac{6 \cdot \|u\|_2^2}{(1-\gamma)^2} \cdot \left\| h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \qquad \text{(Eq. (189))} \tag{462} \\
&\le \frac{6 \cdot \log A}{(1-\gamma)^2} \cdot \|u\|_2^2. \qquad \text{(Eq. (406))} \tag{463}
\end{align}
$$

For the first term, according to Eqs. (173), (189) and (406),

$$
\begin{align}
\left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right| &\le \left\| M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \tag{464} \\
&\le \frac{1}{1-\gamma} \cdot 2 \cdot \|u\|_2 \cdot \frac{1}{1-\gamma} \cdot 2 \cdot \|u\|_2 \cdot \frac{1}{1-\gamma} \cdot \log A \tag{465} \\
&= \frac{4 \cdot \log A}{(1-\gamma)^3} \cdot \|u\|_2^2. \tag{466}
\end{align}
$$

Combining Eqs. (451), (454), (459) and (464) with Eq. (449),

$$\left|\frac{\partial^2 \mathbb{H}^{\pi_{\theta_\alpha}}(s)}{\partial\alpha^2}\Big|_{\alpha=0}\right| \le 2\gamma^2 \cdot \left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial\alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) h_{\theta_\alpha}\Big|_{\alpha=0}\right| + \gamma\cdot \left| e_s^\top M(\alpha)\frac{\partial^2 P(\alpha)}{\partial\alpha^2} M(\alpha) h_{\theta_\alpha}\Big|_{\alpha=0}\right| \tag{467}$$

$$\begin{aligned}
&+ 2\gamma \cdot \left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial\alpha} M(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial\alpha}\Big|_{\alpha=0}\right| + \left| e_s^\top M(\alpha) \frac{\partial^2 h_{\theta_\alpha}}{\partial\alpha^2}\Big|_{\alpha=0}\right| && (468)\\
&\le \left(2\gamma^2 \cdot \frac{4\cdot\log A}{(1-\gamma)^3} + \gamma\cdot\frac{6\cdot\log A}{(1-\gamma)^2} + 2\gamma\cdot\frac{4\cdot\log A}{(1-\gamma)^2} + \frac{3\cdot(1+\log A)}{1-\gamma}\right)\cdot\|u\|_2^2 && (469)\\
&\le \left(\frac{8\cdot\log A}{(1-\gamma)^3} + \frac{3}{1-\gamma}\right)\cdot\|u\|_2^2 && (470)\\
&\le \frac{4+8\cdot\log A}{(1-\gamma)^3}\cdot\|u\|_2^2, && (471)
\end{aligned}$$

which implies for all $y\in\mathbb{R}^{SA}$ and θ,

$$\begin{aligned}
\left| y^\top \frac{\partial^2\mathbb{H}^{\pi_\theta}(s)}{\partial\theta^2} y\right| &= \left| \left(\frac{y}{\|y\|_2}\right)^\top \frac{\partial^2\mathbb{H}^{\pi_\theta}(s)}{\partial\theta^2}\left(\frac{y}{\|y\|_2}\right)\right|\cdot\|y\|_2^2 && (472)\\
&\le \max_{\|u\|_2=1}\left|\left\langle\frac{\partial^2\mathbb{H}^{\pi_\theta}(s)}{\partial\theta^2}u, u\right\rangle\right|\cdot\|y\|_2^2 && (473)\\
&= \max_{\|u\|_2=1}\left|\left\langle\frac{\partial^2\mathbb{H}^{\pi_{\theta_\alpha}}(s)}{\partial\theta_\alpha^2}\Big|_{\alpha=0}\frac{\partial\theta_\alpha}{\partial\alpha}, \frac{\partial\theta_\alpha}{\partial\alpha}\right\rangle\right|\cdot\|y\|_2^2 && (474)\\
&= \max_{\|u\|_2=1}\left|\left\langle\frac{\partial}{\partial\theta_\alpha}\left\{\frac{\partial\mathbb{H}^{\pi_{\theta_\alpha}}(s)}{\partial\alpha}\right\}\Big|_{\alpha=0}, \frac{\partial\theta_\alpha}{\partial\alpha}\right\rangle\right|\cdot\|y\|_2^2 && (475)\\
&= \max_{\|u\|_2=1}\left|\frac{\partial^2\mathbb{H}^{\pi_{\theta_\alpha}}(s)}{\partial\alpha^2}\Big|_{\alpha=0}\right|\cdot\|y\|_2^2 && (476)\\
&\le \frac{4+8\cdot\log A}{(1-\gamma)^3}\cdot\|y\|_2^2. \qquad \text{(Eq. (467))} && (477)
\end{aligned}$$

Denote $\theta_\xi = \theta + \xi(\theta'-\theta)$, where $\xi\in[0,1]$. According to Taylor's theorem, $\forall s, \forall\theta,\ \theta'$,

$$\begin{aligned}
\left|\mathbb{H}^{\pi_{\theta'}}(s) - \mathbb{H}^{\pi_\theta}(s) - \left\langle\frac{\partial\mathbb{H}^{\pi_\theta}(s)}{\partial\theta}, \theta'-\theta\right\rangle\right| &= \frac{1}{2}\cdot\left|(\theta'-\theta)^\top\frac{\partial^2\mathbb{H}^{\pi_{\theta_\xi}}(s)}{\partial\theta_\xi^2}(\theta'-\theta)\right| && (478)\\
&\le \frac{2+4\cdot\log A}{(1-\gamma)^3}\cdot\|\theta'-\theta\|_2^2. \qquad \text{(Eq. (472))} && (479)
\end{aligned}$$

Since $\mathbb{H}^{\pi_\theta}(s)$ is $(4+8\log A)/(1-\gamma)^3$-smooth, $\forall s$, $\mathbb{H}(\rho,\pi_\theta) \coloneqq \mathbb{E}_{s\sim\rho}\left[\mathbb{H}^{\pi_\theta}(s)\right]$ is also $(4+8\log A)/(1-\gamma)^3$-smooth. □

Lemma 16 (Non-uniform Łojasiewicz)**.** Suppose $\mu(s)>0$ for all state $s\in\mathcal{S}$. $\pi_\theta(\cdot|s)\coloneqq\mathrm{softmax}(\theta(s,\cdot))$, $\forall s$.

$$\left\|\frac{\partial\tilde{V}^{\pi_\theta}(\mu)}{\partial\theta}\right\|_2 \ge \frac{\sqrt{2\tau}}{\sqrt{S}}\cdot\min_s\sqrt{\mu(s)}\cdot\min_{s,a}\pi_\theta(a|s)\cdot\left\|\frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-\frac{1}{2}}\cdot\left[\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho)\right]^{\frac{1}{2}}. \tag{480}$$

Proof. According to the definition of soft value functions,

$$\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) = \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, a_t \sim \pi_\tau^*(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t) - \tau \log \pi_\tau^*(a_t|s_t)) \right] - \tilde{V}^{\pi_\theta}(\rho) \tag{481}$$

$$= \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, a_t \sim \pi_\tau^*(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t) - \tau \log \pi_\tau^*(a_t|s_t) + \tilde{V}^{\pi_\theta}(s_t) - \tilde{V}^{\pi_\theta}(s_t)) \right] - \tilde{V}^{\pi_\theta}(\rho) \tag{482}$$

$$= \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, a_t \sim \pi_\tau^*(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t) - \tau \log \pi_\tau^*(a_t|s_t) + \gamma \tilde{V}^{\pi_\theta}(s_{t+1}) - \tilde{V}^{\pi_\theta}(s_t)) \right] \tag{483}$$

$$= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \left[\sum_a \pi_\tau^*(a|s) \cdot \left[r(s, a) - \tau \log \pi_\tau^*(a|s) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) \tilde{V}^{\pi_\theta}(s') - \tilde{V}^{\pi_\theta}(s) \right] \right] \tag{484}$$

$$= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \left[\sum_a \pi_\tau^*(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s, a) - \tau \log \pi_\tau^*(a|s) \right] - \tilde{V}^{\pi_\theta}(s) \right]. \tag{485}$$

Next, define the "soft greedy policy" $\bar{\pi}_\theta(\cdot|s) \coloneqq \mathrm{softmax}(\tilde{Q}^{\pi_\theta}(s, \cdot)/\tau)$, $\forall s$, i.e.,

$$\bar{\pi}_\theta(a|s) \coloneqq \frac{\exp\left\{ \tilde{Q}^{\pi_\theta}(s, a)/\tau \right\}}{\sum_{a'} \exp\left\{ \tilde{Q}^{\pi_\theta}(s, a')/\tau \right\}}, \ \forall a. \tag{486}$$

We have, $\forall s$,

$$\sum_a \pi_\tau^*(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s, a) - \tau \log \pi_\tau^*(a|s) \right] \le \max_{\pi(\cdot|s)} \sum_a \pi(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s, a) - \tau \log \pi(a|s) \right] \tag{487}$$

$$= \sum_a \bar{\pi}_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s, a) - \tau \log \bar{\pi}_\theta(a|s) \right] \tag{488}$$

$$= \tau \log \sum_a \exp\left\{ \tilde{Q}^{\pi_\theta}(s, a)/\tau \right\}. \tag{489}$$

Also note that,

$$\tilde{V}^{\pi_\theta}(s) = \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s, a) - \tau \log \pi_\theta(a|s) \right] \tag{490}$$

$$= \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s, a) - \tau \log \bar{\pi}_\theta(a|s) + \tau \log \bar{\pi}_\theta(a|s) - \tau \log \pi_\theta(a|s) \right] \tag{491}$$

$$= \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s, a) - \tau \log \bar{\pi}_\theta(a|s) \right] - \tau D_{\mathrm{KL}}(\pi_\theta(\cdot|s) \| \bar{\pi}_\theta(\cdot|s)) \tag{492}$$

$$= \tau \log \sum_a \exp\left\{ \tilde{Q}^{\pi_\theta}(s, a)/\tau \right\} - \tau \cdot D_{\mathrm{KL}}(\pi_\theta(\cdot|s) \| \bar{\pi}_\theta(\cdot|s)). \tag{493}$$

Combining the above,

$$\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) = \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \left[\sum_a \pi_\tau^*(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\tau^*(a|s) \right] - \tilde{V}^{\pi_\theta}(s) \right] \tag{494}$$

$$\leq \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \left[\tau \log \sum_a \exp\left\{ \tilde{Q}^{\pi_\theta}(s,a)/\tau \right\} - \tilde{V}^{\pi_\theta}(s) \right] \tag{495}$$

$$= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_\theta(\cdot|s) \| \bar{\pi}_\theta(\cdot|s)) \tag{496}$$

$$\leq \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \frac{\tau}{2} \cdot \left\| \frac{\tilde{Q}^{\pi_\theta}(s,\cdot)}{\tau} - \theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot)/\tau - \theta(s,\cdot))^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty^2 \quad \text{(Lemma 25)} \tag{497}$$

$$= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \frac{1}{2\tau} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty^2 . \tag{498}$$

Taking square root of soft sub-optimality,

$$\left[\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \right]^{\frac{1}{2}} \leq \frac{1}{\sqrt{1-\gamma}} \cdot \left[\sum_s d_\rho^{\pi_\tau^*}(s) \cdot \frac{1}{2\tau} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty^2 \right]^{\frac{1}{2}} \tag{499}$$

$$= \frac{1}{\sqrt{1-\gamma}} \cdot \left[\sum_s \left(\sqrt{d_\rho^{\pi_\tau^*}(s)} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty \right)^2 \right]^{\frac{1}{2}} \tag{500}$$

$$\leq \frac{1}{\sqrt{1-\gamma}} \cdot \sum_s \sqrt{d_\rho^{\pi_\tau^*}(s)} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty \quad (\|x\|_2 \leq \|x\|_1) \tag{501}$$

$$\leq \frac{1}{\sqrt{1-\gamma}} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{\frac{1}{2}} \sum_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty . \tag{502}$$

On the other hand, the entropy regularized policy gradient norm is lower bounded as

$$\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 = \left[\sum_{s,a} \left(\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,a)} \right)^2 \right]^{\frac{1}{2}} \tag{503}$$

$$= \left[\sum_s \left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} \right\|_2^2 \right]^{\frac{1}{2}} \tag{504}$$

$$\geq \frac{1}{\sqrt{S}} \sum_s \left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} \right\|_2 , \qquad \text{(CauchySchwarz, } \|x\|_1 = |\langle \mathbf{1}, |x| \rangle| \leq \|\mathbf{1}\|_2 \cdot \|x\|_2) \tag{505}$$

which is further lower bounded as

$$\begin{aligned}
\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 &\ge \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \left\| H(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \theta(s,\cdot) \right] \right\|_2 && \text{(Eq. (319), Lemma 10)} \quad (506) \\
&= \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \left\| H(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{K} \cdot \mathbf{1} \right] \right\|_2 && \text{(Lemma 21)} \quad (507) \\
&\ge \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \min_a \pi_\theta(a|s) \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_2 && \text{(Lemma 22)} \quad (508) \\
&\ge \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \min_a \pi_\theta(a|s) \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty . && (509)
\end{aligned}$$

Denote $\zeta_\theta(s) \coloneqq \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{K} \cdot \mathbf{1}$. We have,

$$\begin{aligned}
\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 &\ge \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \min_a \pi_\theta(a|s) \cdot \| \zeta_\theta(s) \|_\infty && (510) \\
&\ge \frac{1}{\sqrt{S}} \cdot \frac{1}{\sqrt{1-\gamma}} \cdot \min_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \min_{s,a} \pi_\theta(a|s) \cdot \sqrt{2\tau} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}} \cdot \left[\frac{1}{\sqrt{1-\gamma}} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{\frac{1}{2}} \sum_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \| \zeta_\theta(s) \|_\infty \right] && (511) \\
&\ge \frac{1}{\sqrt{S}} \cdot \frac{1}{\sqrt{1-\gamma}} \cdot \min_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \min_{s,a} \pi_\theta(a|s) \cdot \sqrt{2\tau} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}} \cdot \left[\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \right]^{\frac{1}{2}} && (512) \\
&\ge \frac{\sqrt{2\tau}}{\sqrt{S}} \cdot \min_s \sqrt{\mu(s)} \cdot \min_{s,a} \pi_\theta(a|s) \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}} \cdot \left[\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \right]^{\frac{1}{2}} , && (513)
\end{aligned}$$

where the last inequality is by $d_\mu^{\pi_\theta}(s) \ge (1-\gamma) \cdot \mu(s)$ similar with Eq. (304). □

Lemma 17. Using Algorithm 1 with soft policy gradient Eq. (18), we have $\inf_{t \ge 1} \min_{s,a} \pi_{\theta_t}(a|s) > 0$.

Proof. The augmented value function $\tilde{V}^{\pi_{\theta_t}}(\rho)$ is monotonically increasing following gradient update due to smoothness, i.e., Lemmas 7 and 15. And $\tilde{V}^{\pi_{\theta_t}}(\rho)$ is upper bounded as

$$\begin{aligned}
\tilde{V}^{\pi_{\theta_t}}(\rho) &= \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, a_t \sim \pi_{\theta_t}(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\sum_{t=0}^\infty \gamma^t (r(s_t,a_t) - \tau \log \pi_{\theta_t}(a_t|s_t)) \right] && (514) \\
&= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_{\theta_t}}(s) \cdot \left[\sum_a \pi_{\theta_t}(a|s) \cdot (r(s,a) - \tau \log \pi_{\theta_t}(a|s)) \right] && (515) \\
&\le \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_{\theta_t}}(s) \cdot (1 + \tau \log A) \qquad \left(r(s,a) \le 1, \ -\sum_a \pi_{\theta_t}(a|s) \cdot \log \pi_{\theta_t}(a|s) \le \log A \right) && (516) \\
&\le \frac{1 + \tau \log A}{1-\gamma}. && (517)
\end{aligned}$$

According to monotone convergence theorem, $\tilde{V}^{\pi_{\theta_t}}(\rho)$ converges to a finite value. Suppose $\pi_{\theta_t}(a|s) \to \pi_{\theta_\infty}(a|s)$. For any state $s \in \mathcal{S}$, define the following sets,

$$\begin{aligned}
\mathcal{A}_0(s) &\coloneqq \left\{ a : \pi_{\theta_\infty}(a|s) = 0 \right\}, && (518) \\
\mathcal{A}_+(s) &\coloneqq \left\{ a : \pi_{\theta_\infty}(a|s) > 0 \right\}. && (519)
\end{aligned}$$

Note that $\mathcal{A} = \mathcal{A}_0(s) \cup \mathcal{A}_+(s)$ since $\pi_\infty(a|s) \ge 0, \forall a \in \mathcal{A}$. We prove that for any state $s \in \mathcal{S}$, $\mathcal{A}_0(s) = \emptyset$ by contradiction. Suppose $\exists s \in \mathcal{S}$, such that $\mathcal{A}_0(s)$ is non-empty. For any $a_0 \in \mathcal{A}_0(s)$, we have $\pi_{\theta_t}(a_0|s) \to \pi_{\theta_\infty}(a_0|s) = 0$, which implies $-\log \pi_{\theta_t}(a_0|s) \to \infty$. There exists $t_0 > 0$, such that $\forall t \ge t_0$,

$$-\log \pi_{\theta_t}(a_0|s) \ge \frac{1 + \tau \log A}{\tau(1-\gamma)}. \tag{520}$$

According to Lemma 10, $\forall t \ge t_0$,

$$\begin{aligned}
\frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a_0)} &= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a_0|s) \cdot \tilde{A}^{\pi_{\theta_t}}(s, a_0) &(521)\\
&= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a_0|s) \cdot \left[\tilde{Q}^{\pi_{\theta_t}}(s, a_0) - \tau \log \pi_{\theta_t}(a_0|s) - \tilde{V}^{\pi_{\theta_t}}(s) \right] &(522)\\
&\ge \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a_0|s) \cdot \left[0 - \tau \log \pi_{\theta_t}(a_0|s) - \frac{1 + \tau \log A}{1-\gamma} \right] &(523)\\
&\ge \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a_0|s) \cdot \left[0 + \tau \cdot \frac{1 + \tau \log A}{\tau(1-\gamma)} - \frac{1 + \tau \log A}{1-\gamma} \right] = 0, &(524)
\end{aligned}$$

where the first inequality is by

$$\tilde{Q}^{\pi_{\theta_t}}(s, a_0) = r(s, a_0) + \gamma \sum_{s'} \mathcal{P}(s'|s, a_0) \tilde{V}^{\pi_{\theta_t}}(s') \ge 0. \qquad \left(r(s, a_0) \ge 0, \ \tilde{V}^{\pi_{\theta_t}}(s') \ge 0 \right) \tag{525}$$

This means $\theta_t(s, a_0)$ is always increasing $\forall t \ge t_0$, which implies $\theta_\infty(s, a_0)$ is lower bounded by constant, i.e., $\theta_\infty(s, a_0) \ge c$ for some constant c, and thus $\exp\{\theta_\infty(a_0|s)\} \ge e^c > 0$. According to

$$\pi_{\theta_\infty}(a_0|s) = \frac{\exp\{\theta_\infty(a_0|s)\}}{\sum_a \exp\{\theta_\infty(a|s)\}} = 0, \tag{526}$$

we have,

$$\sum_a \exp\{\theta_\infty(a|s)\} = \infty. \tag{527}$$

On the other hand, for any $a_+ \in \mathcal{A}_+(s)$, according to

$$\pi_{\theta_\infty}(a_+|s) = \frac{\exp\{\theta_\infty(a_+|s)\}}{\sum_a \exp\{\theta_\infty(a|s)\}} > 0, \tag{528}$$

we have,

$$\exp\{\theta_\infty(a_+|s)\} = \infty, \quad \forall a_+ \in \mathcal{A}_+(s) \tag{529}$$

which implies,

$$\sum_{a_+ \in \mathcal{A}_+(s)} \theta_\infty(a_+|s) = \infty. \tag{530}$$

Note that $\forall t$, the summation of logit incremental over all actions is zero:

$$\begin{aligned}
\sum_a \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a)} &= \sum_{a_0 \in \mathcal{A}_0(s)} \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a_0)} + \sum_{a_+ \in \mathcal{A}_+(s)} \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a_+)} &(531)\\
&= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \sum_a \pi_{\theta_t}(a|s) \tilde{A}^{\pi_{\theta_t}}(s, a) &(532)\\
&= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \left[\tilde{V}^{\pi_{\theta_t}}(s) - \tilde{V}^{\pi_{\theta_t}}(s) \right] = 0. &(533)
\end{aligned}$$

According to Eq. (521), $\forall t \ge t_0$,

$$\sum_{a_0 \in \mathcal{A}_0(s)} \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a_0)} \ge 0. \tag{534}$$

According to Eq. (531), $\forall t \ge t_0$,

$$\sum_{a_+ \in \mathcal{A}_+(s)} \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a_+)} = 0 - \sum_{a_0 \in \mathcal{A}_0(s)} \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a_0)} \le 0. \tag{535}$$

which means $\sum_{a_+ \in \mathcal{A}_+(s)} \theta_t(s, a_+)$ will always decrease for all large enough $t > 0$. This is a contradiction with Eq. (530), i.e., $\sum_{a_+ \in \mathcal{A}_+(s)} \theta_t(s, a_+) \to \infty$.

To this point, we have shown that $\mathcal{A}_0(s) = \emptyset$ for any state $s \in \mathcal{S}$, i.e., $\pi_{\theta_t}(\cdot|s)$ will converge in the interior of probabilistic simplex $\Delta(\mathcal{A})$. And at the convergent point $\pi_{\theta_\infty}(\cdot|s)$, the gradient is zero, otherwise by smoothness the objective can be further improved, which is a contradiction with convergence. According to Lemma 10, $\forall s$,

$$\frac{\partial \tilde{V}^{\pi_{\theta_\infty}}(\mu)}{\partial \theta_\infty(s, \cdot)} = \frac{1}{1-\gamma} \cdot d_\mu^{\pi_{\theta_\infty}}(s) \cdot H(\pi_{\theta_\infty}(\cdot|s)) \left[\tilde{Q}^{\pi_{\theta_\infty}}(s, \cdot) - \tau \log \pi_{\theta_\infty}(\cdot|s) \right] = \mathbf{0}. \tag{536}$$

Similar with Eq. (304), we have $d_\mu^{\pi_{\theta_\infty}}(s) \ge (1-\gamma) \cdot \mu(s) > 0$ for all state s. Therefore we have, $\forall s$,

$$H(\pi_{\theta_\infty}(\cdot|s)) \left[\tilde{Q}^{\pi_{\theta_\infty}}(s, \cdot) - \tau \log \pi_{\theta_\infty}(\cdot|s) \right] = \mathbf{0}. \tag{537}$$

According to Lemma 21, $H(\pi_{\theta_\infty}(\cdot|s))$ has eigenvalue 0 with multiplicity 1, and its corresponding eigenvector is $c \cdot \mathbf{1}$ for some constant $c \in \mathbb{R}$. Therefore, the gradient is zero implies that for all state s,

$$\tilde{Q}^{\pi_{\theta_\infty}}(s, \cdot) - \tau \log \pi_{\theta_\infty}(\cdot|s) = c \cdot \mathbf{1}, \tag{538}$$

which is equivalent with

$$\pi_{\theta_\infty}(\cdot|s) = \mathrm{softmax}(\tilde{Q}^{\pi_{\theta_\infty}}(s, \cdot)/\tau), \tag{539}$$

which, according to Nachum et al. (2017, Theorem 3), is the softmax optimal policy π_τ^*. Since $\tau \in \Omega(1) > 0$ and,

$$0 \le \tilde{Q}^{\pi_{\theta_\infty}}(s, a) \le \frac{1 + \tau \log A}{1-\gamma}, \tag{540}$$

we have $\pi_{\theta_\infty}(a|s) \in \Omega(1)$, $\forall (s, a)$. Since $\pi_{\theta_t}(a|s) \to \pi_{\theta_\infty}(a|s)$, there exists $t_0 > 0$, such that $\forall t \ge t_0$,

$$0.9 \cdot \pi_{\theta_\infty}(a|s) \le \pi_{\theta_t}(a|s) \le 1.1 \cdot \pi_{\theta_\infty}(a|s), \ \forall (s, a), \tag{541}$$

which means $\inf_{t \ge t_0} \min_{s,a} \pi_{\theta_t}(a|s) \in \Omega(1)$, and thus

$$\inf_{t \ge 1} \min_{s,a} \pi_{\theta_t}(a|s) = \min \left\{ \min_{1 \le t \le t_0} \min_{s,a} \pi_{\theta_t}(a|s), \ \inf_{t \ge t_0} \min_{s,a} \pi_{\theta_t}(a|s) \right\} = \min\{\Omega(1), \ \Omega(1)\} \in \Omega(1). \qquad \square$$

Theorem 6. Suppose $\mu(s) > 0$ for all state s. Using Algorithm 1 with entropy regularized softmax policy gradient Eq. (18), $\eta = (1-\gamma)^3/(8 + \tau(4 + 8 \log A))$ and $\pi_{\theta_1}(a|s) \in \Omega(1)$, $\forall (s, a)$,

$$\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) \le \frac{\left\| 1/\mu \right\|_\infty}{\exp\{C_\tau \cdot \Omega(1) \cdot t\}} \cdot \frac{1 + \tau \log A}{(1-\gamma)^2}, \tag{542}$$

for all $t > 0$, where $C_\tau, \Omega(1) > 0$ are independent with t.

Proof. According to the soft sub-optimality lemma of Lemma 24,

$$
\begin{aligned}
\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) &= \frac{1}{1-\gamma} \sum_s \left[d_\rho^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s) \| \pi_\tau^*(\cdot|s)) \right] && (543) \\
&= \frac{1}{1-\gamma} \sum_s \frac{d_\rho^{\pi_{\theta_t}}(s)}{d_\mu^{\pi_{\theta_t}}(s)} \cdot \left[d_\mu^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s) \| \pi_\tau^*(\cdot|s)) \right] && (544) \\
&\le \frac{1}{(1-\gamma)^2} \sum_s \frac{1}{\mu(s)} \cdot \left[d_\mu^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s) \| \pi_\tau^*(\cdot|s)) \right] && (545) \\
&\le \frac{1}{(1-\gamma)^2} \cdot \left\| \frac{1}{\mu} \right\|_\infty \sum_s \left[d_\mu^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s) \| \pi_\tau^*(\cdot|s)) \right] && (546) \\
&= \frac{1}{1-\gamma} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot \left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu) \right], && (547)
\end{aligned}
$$

where the last equation is again by Lemma 24, and the first inequality is according to $d_\mu^{\pi_{\theta_t}}(s) \ge (1-\gamma) \cdot \mu(s)$ similar with Eq. (304). According to Lemmas 7 and 15, $V^{\pi_\theta}(\mu)$ is $8/(1-\gamma)^3$-smooth, and $\mathbb{H}(\mu, \pi_\theta)$ is $(4 + 8 \log A)/(1-\gamma)^3$-smooth. Therefore, $\tilde{V}^{\pi_\theta}(\mu) = V^{\pi_\theta}(\mu) + \tau \mathbb{H}(\mu, \pi_\theta)$ is β-smooth with $\beta = (8 + \tau(4 + 8 \log A))/(1-\gamma)^3$, i.e.,

$$
\left| \tilde{V}^{\pi_{\theta_{t+1}}}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu) - \left\langle \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| \le \frac{4 + \tau(2 + 4 \log A)}{(1-\gamma)^3} \cdot \| \theta_{t+1} - \theta_t \|_2^2 . \tag{548}
$$

Denote $\tilde{\delta}_t := \tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu)$. Then we have,

$$
\begin{aligned}
\tilde{\delta}_{t+1} - \tilde{\delta}_t &= \tilde{V}^{\pi_{\theta_t}}(\mu) - \tilde{V}^{\pi_{\theta_{t+1}}}(\mu) && (549) \\
&\le - \left\langle \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}, \theta_{t+1} - \theta_t \right\rangle + \frac{4 + \tau(2 + 4 \log A)}{(1-\gamma)^3} \cdot \| \theta_{t+1} - \theta_t \|_2^2 && (550) \\
&= \left(-\eta + \frac{4 + \tau(2 + 4 \log A)}{(1-\gamma)^3} \cdot \eta^2 \right) \cdot \left\| \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2^2 && (551) \\
&= - \frac{(1-\gamma)^3}{16 + \tau(8 + 16 \log A)} \cdot \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2^2 \qquad \left(\eta = \frac{(1-\gamma)^3}{8 + \tau(4 + 8 \log A)} \right) && (552) \\
&\le - \frac{(1-\gamma)^3}{16 + \tau(8 + 16 \log A)} \cdot \frac{2\tau}{S} \cdot \min_s \mu(s) \cdot \left[\min_{s,a} \pi_{\theta_t}(a|s) \right]^2 \cdot \left\| \frac{d_\mu^{\pi_\tau^*}}{d_\mu^{\pi_{\theta_t}}} \right\|_\infty^{-1} \cdot \left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu) \right] \qquad \text{(Lemma 16)} && (553) \\
&\le - \frac{(1-\gamma)^4}{(8/\tau + 4 + 8 \log A) \cdot S} \cdot \min_s \mu(s) \cdot \left[\min_{s,a} \pi_{\theta_t}(a|s) \right]^2 \cdot \left\| \frac{d_\mu^{\pi_\tau^*}}{\mu} \right\|_\infty^{-1} \cdot \tilde{\delta}_t \qquad \left(d_\mu^{\pi_{\theta_t}}(s) \ge (1-\gamma) \cdot \mu(s) \right) && (554) \\
&\le - \frac{(1-\gamma)^4}{(8/\tau + 4 + 8 \log A) \cdot S} \cdot \min_s \mu(s) \cdot \left[\inf_{t \ge 1} \min_{s,a} \pi_{\theta_t}(a|s) \right]^2 \cdot \left\| \frac{d_\mu^{\pi_\tau^*}}{\mu} \right\|_\infty^{-1} \cdot \tilde{\delta}_t, && (555)
\end{aligned}
$$

According to Lemma 17, $\inf_{t \ge 1} \min_{s,a} \pi_{\theta_t}(a|s) \in \Omega(1)$ is independent with t. We have,

$$\tilde{\delta}_t \le \left[1 - \frac{(1-\gamma)^4}{(8/\tau + 4 + 8 \log A) \cdot S} \cdot \min_s \mu(s) \cdot \Omega(1) \cdot \left\| \frac{d_\mu^{\pi_\tau^*}}{\mu} \right\|_\infty^{-1} \right] \cdot \tilde{\delta}_{t-1} \tag{556}$$

$$\le \exp\left\{ - \frac{(1-\gamma)^4}{(8/\tau + 4 + 8 \log A) \cdot S} \cdot \min_s \mu(s) \cdot \Omega(1) \cdot \left\| \frac{d_\mu^{\pi_\tau^*}}{\mu} \right\|_\infty^{-1} \right\} \cdot \tilde{\delta}_{t-1} \tag{557}$$

$$\le \exp\left\{ - \frac{(1-\gamma)^4}{(8/\tau + 4 + 8 \log A) \cdot S} \cdot \min_s \mu(s) \cdot \Omega(1) \cdot \left\| \frac{d_\mu^{\pi_\tau^*}}{\mu} \right\|_\infty^{-1} \cdot (t-1) \right\} \cdot \tilde{\delta}_1 \tag{558}$$

$$\le \exp\left\{ - \frac{(1-\gamma)^4}{(8/\tau + 4 + 8 \log A) \cdot S} \cdot \min_s \mu(s) \cdot \Omega(1) \cdot \left\| \frac{d_\mu^{\pi_\tau^*}}{\mu} \right\|_\infty^{-1} \cdot (t-1) \right\} \cdot \frac{1 + \tau \log A}{1-\gamma}, \tag{559}$$

where the last inequality is according to Eq. (514). Therefore we have the final result,

$$\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) \le \frac{1}{1-\gamma} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot \left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu) \right] \tag{560}$$

$$\le \frac{1}{\exp\left\{ C_\tau \cdot \Omega(1) \cdot t \right\}} \cdot \frac{1 + \tau \log A}{(1-\gamma)^2} \cdot \left\| \frac{1}{\mu} \right\|_\infty, \tag{561}$$

where

$$C_\tau = \frac{(1-\gamma)^4}{(8/\tau + 4 + 8 \log A) \cdot S} \cdot \min_s \mu(s) \cdot \left\| \frac{d_\mu^{\pi_\tau^*}}{\mu} \right\|_\infty^{-1} \in \Omega(1), \tag{562}$$

is independent with t. □

A.3. Proofs for Section 5

Lemma 18 (Reverse Łojasiewicz)**.** Denote $\Delta \coloneqq r(a^*) - \max_{a \ne a^*} r(a) > 0$ as the reward gap of r.

$$\left\| \frac{d \pi_\theta^\top r}{d \theta} \right\|_2 \le \frac{\sqrt{2}}{\Delta} \cdot (\pi^* - \pi_\theta)^\top r. \tag{563}$$

Proof. Note a^* is the optimal action. Denote $\Delta(a) \coloneqq r(a^*) - r(a)$, and $\Delta = \min_{a \ne a^*} \Delta(a)$.

$$(\pi^* - \pi_\theta)^\top r = \sum_a \pi_\theta(a) \cdot r(a^*) - \sum_a \pi_\theta(a) \cdot r(a) \tag{564}$$

$$= \sum_{a \ne a^*} \pi_\theta(a) \cdot r(a^*) - \sum_{a \ne a^*} \pi_\theta(a) \cdot r(a) \tag{565}$$

$$= \sum_{a \ne a^*} \pi_\theta(a) \cdot \Delta(a) \tag{566}$$

$$\ge \sum_{a \ne a^*} \pi_\theta(a) \cdot \Delta. \tag{567}$$

On the other hand,

$$0 \le r(a^*) - \pi_\theta^\top r = (\pi^* - \pi_\theta)^\top r = \sum_{a \ne a^*} \pi_\theta(a) \cdot \Delta(a) \le \sum_{a \ne a^*} \pi_\theta(a) \cdot 1 = \sum_{a \ne a^*} \pi_\theta(a). \tag{568}$$

Therefore the ℓ_2 norm of gradient can be upper bounded as

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 = \left(\pi_\theta(a^*)^2 \cdot \left[r(a^*) - \pi_\theta^\top r \right]^2 + \sum_{a \neq a^*} \left[\pi_\theta(a)^2 \cdot (r(a) - \pi_\theta^\top r)^2 \right] \right)^{\frac{1}{2}} \tag{569}$$

$$\leq \left(1^2 \cdot \left[\sum_{a \neq a^*} \pi_\theta(a) \right]^2 + \sum_{a \neq a^*} \left[\pi_\theta(a)^2 \cdot 1^2 \right] \right)^{\frac{1}{2}} \tag{570}$$

$$\leq \left(\left[\sum_{a \neq a^*} \pi_\theta(a) \right]^2 + \left[\sum_{a \neq a^*} \pi_\theta(a) \right]^2 \right)^{\frac{1}{2}} \qquad (\|x\|_2 \leq \|x\|_1) \tag{571}$$

$$= \sqrt{2} \cdot \sum_{a \neq a^*} \pi_\theta(a). \tag{572}$$

Combining the results, we have

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \leq \sqrt{2} \cdot \sum_{a \neq a^*} \pi_\theta(a) = \frac{\sqrt{2}}{\Delta} \cdot \Delta \cdot \sum_{a \neq a^*} \pi_\theta(a) \leq \frac{\sqrt{2}}{\Delta} \cdot (\pi^* - \pi_\theta)^\top r. \qquad \square$$

Theorem 7 (Lower bound)**.** For large enough $t > 0$, using Update 1 with learning rate $\eta \in (0, 1]$,

$$(\pi^* - \pi_{\theta_t})^\top r \geq \frac{\Delta^2}{6\eta \cdot t}. \tag{573}$$

Proof. According to the reverse Łojasiewicz inequality of Lemma 18,

$$\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \leq \frac{\sqrt{2}}{\Delta} \cdot \delta_t, \tag{574}$$

where $\delta_t \coloneqq (\pi^* - \pi_{\theta_t})^\top r > 0$. Let $\theta_{t+1} \leftarrow \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t}$, and $\pi_{\theta_{t+1}} = \mathrm{softmax}(\theta_{t+1})$ be the next policy after one step gradient update. Using similar arguments as smoothness property,

$$\delta_t - \delta_{t+1} \leq |\delta_t - \delta_{t+1}| \tag{575}$$

$$= \left| (\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r \right| \tag{576}$$

$$= \left| (\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle + \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| \tag{577}$$

$$\leq \left| (\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| + \left| \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| \tag{578}$$

$$\leq \frac{5}{4} \cdot \|\theta_{t+1} - \theta_t\|_2^2 + \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2 \qquad \text{(Lemma 2, and Cauchy-Schwarz)} \tag{579}$$

$$= \left(\frac{5\eta^2}{4} + \eta \right) \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \qquad \left(\theta_{t+1} = \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right) \tag{580}$$

$$\leq \frac{9\eta}{2} \cdot \frac{1}{\Delta^2} \cdot \delta_t^2. \qquad (\eta \in (0, 1], \text{ and Lemma 18}) \tag{581}$$

According to convergence result Theorem 2 we have $\delta_t > 0$, $\delta_t \to 0$ as $t \to \infty$. We prove that for all large enough $t > 0$,

$\delta_t \le \frac{10}{9} \cdot \delta_{t+1}$ by contradiction. Suppose $\delta_t > \frac{10}{9} \cdot \delta_{t+1}$.

$$\begin{aligned}\delta_{t+1} &\ge \delta_t - \frac{9\eta}{2} \cdot \frac{1}{\Delta^2} \cdot \delta_t^2 && (582)\\ &> \frac{10}{9} \cdot \delta_{t+1} - \frac{9\eta}{2} \cdot \frac{1}{\Delta^2} \cdot \left(\frac{10}{9} \cdot \delta_{t+1} \right)^2 \qquad \left(f(x) \coloneqq x - ax^2 \text{ is increasing for } x < \frac{1}{2a}, \forall a > 0 \right) && (583)\\ &= \frac{10}{9} \cdot \delta_{t+1} - \frac{50\eta}{9} \cdot \frac{1}{\Delta^2} \cdot \delta_{t+1}^2, && (584)\end{aligned}$$

which implies $\delta_{t+1} > \frac{\Delta^2}{50\eta}$ for large enough $t > 0$. This is a contradiction with $\delta_t \to 0$ as $t \to \infty$. Now we have $\delta_t \le \frac{10}{9} \cdot \delta_{t+1}$. Divide both sides of $\delta_t - \delta_{t+1} \le \frac{9\eta}{2} \cdot \frac{1}{\Delta^2} \cdot \delta_t^2$ by $\delta_t \cdot \delta_{t+1}$,

$$\frac{1}{\delta_{t+1}} - \frac{1}{\delta_t} \le \frac{9\eta}{2} \cdot \frac{1}{\Delta^2} \cdot \frac{\delta_t}{\delta_{t+1}} \le \frac{9\eta}{2} \cdot \frac{1}{\Delta^2} \cdot \frac{10}{9} = \frac{5\eta}{\Delta^2}. \tag{585}$$

Summing up from T_1 (some large enough time) to $T_1 + t$, we have

$$\frac{1}{\delta_{T_1+t}} - \frac{1}{\delta_{T_1}} \le \frac{5\eta}{\Delta^2} \cdot (t-1) \le \frac{5\eta}{\Delta^2} \cdot t. \tag{586}$$

Since T_1 is a finite time, $\delta_{T_1} \ge 1/C$ for some constant $C > 0$. Rearranging, we have

$$(\pi^* - \pi_{\theta_{T_1+t}})^\top r = \delta_{T_1+t} \ge \frac{1}{\frac{1}{\delta_{T_1}} + \frac{5\eta}{\Delta^2} \cdot t} \ge \frac{1}{C + \frac{5\eta}{\Delta^2} \cdot t} \ge \frac{1}{C + \frac{5\eta}{\Delta^2} \cdot (T_1 + t)}. \tag{587}$$

By abusing notation $t \coloneqq T_1 + t$ and $C \le \frac{\eta}{\Delta^2} \cdot t$, we have

$$(\pi^* - \pi_{\theta_t})^\top r \ge \frac{1}{C + \frac{5\eta}{\Delta^2} \cdot t} \ge \frac{1}{\frac{\eta}{\Delta^2} \cdot t + \frac{5\eta}{\Delta^2} \cdot t} = \frac{\Delta^2}{6\eta \cdot t}, \tag{588}$$

for all large enough $t > 0$. □

Theorem 8 (Lower bound)**.** For large enough $t > 0$, using softmax policy gradient Algorithm 1 with $\eta \in (0, 1]$,

$$V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \ge \frac{(1-\gamma)^5 \cdot (\Delta^*)^2}{12\eta \cdot t}, \tag{589}$$

where $\Delta^* \coloneqq \min_{s \in \mathcal{S}, a \neq a^*(s)} \{Q^*(s, a^*(s)) - Q^*(s, a)\} > 0$ is the optimal value gap of the MDP, and $a^*(s) \coloneqq \arg\max_a \pi^*(a|s)$ is the action that the optimal policy selects under state s.

Proof. Suppose Algorithm 1 can converge faster than $O(1/t)$ for general MDPs, then it can converge faster than $O(1/t)$ for any one-state MDPs, which are special cases of general MDPs. This is a contradiction with Theorem 7.

The above one-sentence argument implies a $\Omega(1/t)$ rate lower bound. To calculate the constant in the lower bound, we need results similar with Lemma 18. According to the reverse Łojasiewicz inequality of Lemma 26,

$$\left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \le \frac{1}{1-\gamma} \cdot \frac{\sqrt{2}}{\Delta^*} \cdot \delta_t, \tag{590}$$

where $\delta_t \coloneqq V^*(\mu) - V^{\pi_{\theta_t}}(\mu) > 0$. Let $\theta_{t+1} \leftarrow \theta_t + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}$, and $\pi_{\theta_{t+1}}(\cdot|s) = \mathrm{softmax}(\theta_{t+1}(s, \cdot)), \forall s \in \mathcal{S}$ be the next policy after one step gradient update. Using similar calculations as in Eq. (575),

$$
\begin{aligned}
\delta_t - \delta_{t+1} &\le \left| V^{\pi_{\theta_{t+1}}}(\mu) - V^{\pi_{\theta_t}}(\mu) \right| && (591)\\
&= \left| V^{\pi_{\theta_{t+1}}}(\mu) - V^{\pi_{\theta_t}}(\mu) - \left\langle \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}, \theta_{t+1} - \theta_t \right\rangle + \left\langle \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| && (592)\\
&\le \left| V^{\pi_{\theta_{t+1}}}(\mu) - V^{\pi_{\theta_t}}(\mu) - \left\langle \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| + \left| \left\langle \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| && (593)\\
&\le \frac{4}{(1-\gamma)^3} \cdot \| \theta_{t+1} - \theta_t \|_2^2 + \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2 \cdot \| \theta_{t+1} - \theta_t \|_2 \qquad \text{(Lemma 7, and Cauchy-Schwarz)} && (594)\\
&= \left(\frac{4\eta^2}{(1-\gamma)^3} + \eta \right) \cdot \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2^2 \qquad \left(\theta_{t+1} = \theta_t + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right) && (595)\\
&\le \frac{10\eta}{(1-\gamma)^5} \cdot \frac{1}{(\Delta^*)^2} \cdot \delta_t^2. \qquad (\eta \in (0,1], \text{ and Lemma 26}) && (596)
\end{aligned}
$$

According to Theorem 4, we have $\delta_t > 0$, $\delta_t \to 0$ as $t \to \infty$. Using similar arguments as in Eq. (582), we can show that for all large enough $t > 0$, $\delta_t \le \frac{11}{10} \cdot \delta_{t+1}$. Divide both sides of $\delta_t - \delta_{t+1} \le \frac{10\eta}{(1-\gamma)^5} \cdot \frac{1}{(\Delta^*)^2} \cdot \delta_t^2$ by $\delta_t \cdot \delta_{t+1}$,

$$
\frac{1}{\delta_{t+1}} - \frac{1}{\delta_t} \le \frac{10\eta}{(1-\gamma)^5} \cdot \frac{1}{(\Delta^*)^2} \cdot \frac{\delta_t}{\delta_{t+1}} \le \frac{10\eta}{(1-\gamma)^5} \cdot \frac{1}{(\Delta^*)^2} \cdot \frac{11}{10} = \frac{11\eta}{(1-\gamma)^5 \cdot (\Delta^*)^2}. \tag{597}
$$

Using similar calculations as in the proof of Theorem 7, we have,

$$
V^*(\mu) - V^{\pi_{\theta_t}}(\mu) = \delta_t \ge \frac{(1-\gamma)^5 \cdot (\Delta^*)^2}{12\eta \cdot t}, \tag{598}
$$

for all large enough $t > 0$. □

Proposition 3. The Łojasiewicz degree of $\mathbb{E}_{a \sim \pi_\theta}[r(a)]$ cannot be larger than 0 with $C(\theta) = \pi_\theta(a^*)$.

Proof. We prove by contradiction. Suppose the Łojasiewicz degree of $\mathbb{E}_{a \sim \pi_\theta}[r(a)]$ can be larger than 0. Then there exists $\xi > 0$, such that,

$$
\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \ge C(\theta) \cdot \left[(\pi^* - \pi_\theta)^\top r \right]^{1-\xi}. \tag{599}
$$

Consider the following example, $r = (6, 4, 2)^\top$, $\pi_\theta = (1 - 3\epsilon, 2\epsilon, \epsilon)^\top$ with small number $\epsilon > 0$.

$$
(\pi^* - \pi_\theta)^\top r = r(a^*) - \pi_\theta^\top r = 6 - (6 - 8\epsilon) = 8 \cdot \epsilon. \tag{600}
$$

According to the reverse Łojasiewicz inequality of Lemma 18,

$$
\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \le \frac{\sqrt{2}}{\Delta} \cdot (\pi^* - \pi_\theta)^\top r = \frac{\sqrt{2}}{2} \cdot (\pi^* - \pi_\theta)^\top r \le \frac{1.5}{2} \cdot (\pi^* - \pi_\theta)^\top r = 6 \cdot \epsilon. \tag{601}
$$

Also note that $\pi_\theta(a^*) = 1 - 3\epsilon > 1/4$. Then for $\xi \in (0, 1]$, we have

$$
\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \le 6 \cdot \epsilon = \frac{1}{4} \cdot 3 \cdot 8 \cdot \epsilon < \pi_\theta(a^*) \cdot 3 \cdot 8 \cdot \epsilon = C(\theta) \cdot 3 \cdot 8 \cdot \epsilon. \tag{602}
$$

Next, since $\epsilon > 0$ can be very small,

$$
\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 < C(\theta) \cdot 3 \cdot 8 \cdot \epsilon = C(\theta) \cdot 3 \cdot (8 \cdot \epsilon)^\xi \cdot (8 \cdot \epsilon)^{1-\xi} < C(\theta) \cdot (8 \cdot \epsilon)^{1-\xi} = C(\theta) \cdot \left[(\pi^* - \pi_\theta)^\top r \right]^{1-\xi}, \tag{603}
$$

where the second inequality is by $(8 \cdot \epsilon)^\xi < 1/3$ for small $\epsilon > 0$ since $\xi > 0$. This is a contradiction with the assumption. Therefore the Łojasiewicz degree ξ cannot be larger than 0. □

Proposition 4. With $C(\theta) = \sqrt{2\tau} \cdot \min_a \pi_\theta(a)$, the Łojasiewicz degree of $\mathbb{E}_{a\sim\pi_\theta}\left[r(a) - \tau \log \pi_\theta(a)\right]$ is $1/2$.

Proof. Denote $\delta_\theta \coloneqq \mathbb{E}_{a\sim\pi_\tau^*}\left[r(a) - \tau \log \pi_\tau^*(a)\right] - \mathbb{E}_{a\sim\pi_\theta}\left[r(a) - \tau \log \pi_\theta(a)\right]$ as the soft sub-optimality. We have,

$$\begin{aligned}
\delta_\theta &= \mathop{\mathbb{E}}_{a\sim\pi_\tau^*}\left[r(a) - \tau \log \pi_\tau^*(a)\right] - \mathop{\mathbb{E}}_{a\sim\pi_\theta}\left[r(a) - \tau \log \pi_\tau^*(a)\right] - \mathop{\mathbb{E}}_{a\sim\pi_\theta}\left[\tau \log \pi_\tau^*(a) - \tau \log \pi_\theta(a)\right] && (604)\\
&= \tau \log \sum_a \exp\{r(a)/\tau\} - \tau \log \sum_a \exp\{r(a)/\tau\} + \tau \cdot D_{\mathrm{KL}}(\pi_\theta \| \pi_\tau^*) \qquad (\pi_\tau^* = \mathrm{softmax}(r/\tau)) && (605)\\
&= \tau \cdot D_{\mathrm{KL}}(\pi_\theta \| \pi_\tau^*) && (606)\\
&\le \frac{\tau}{2} \cdot \left\| \frac{r}{\tau} - \theta - \frac{(r/\tau - \theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty^2 \qquad \text{(Lemma 25)} && (607)\\
&= \frac{1}{2\tau} \cdot \left\| r - \tau\theta - \frac{(r - \tau\theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty^2 . && (608)
\end{aligned}$$

Next, the entropy regularized policy gradient w.r.t. θ is

$$\begin{aligned}
\frac{d\{\pi_\theta^\top (r - \tau \log \pi_\theta)\}}{d\theta} &= H(\pi_\theta)(r - \tau \log \pi_\theta) && (609)\\
&= H(\pi_\theta) \left(r - \tau\theta + \tau \log \sum_a \exp\{\theta(a)\} \cdot \mathbf{1} \right) && (610)\\
&= H(\pi_\theta) \left(r - \tau\theta \right) && (611)\\
&= H(\pi_\theta) \left(r - \tau\theta - \frac{(r - \tau\theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right), && (612)
\end{aligned}$$

where the last two equations are by $H(\pi_\theta)\mathbf{1} = \mathbf{0}$ as shown in Lemma 21. Then we have,

$$\begin{aligned}
\left\| \frac{d\{\pi_\theta^\top (r - \tau \log \pi_\theta)\}}{d\theta} \right\|_2 &= \left\| H(\pi_\theta) \left(r - \tau\theta - \frac{(r - \tau\theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right) \right\|_2 && (613)\\
&\ge \min_a \pi_\theta(a) \cdot \left\| r - \tau\theta - \frac{(r - \tau\theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_2 \qquad \text{(Lemma 22)} && (614)\\
&\ge \min_a \pi_\theta(a) \cdot \left\| r - \tau\theta - \frac{(r - \tau\theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty && (615)\\
&\ge \min_a \pi_\theta(a) \cdot \sqrt{2\tau} \cdot \sqrt{\delta_\theta} \qquad \text{(Eq. (604))} && (616)\\
&= \sqrt{2\tau} \cdot \min_a \pi_\theta(a) \cdot \left(\mathop{\mathbb{E}}_{a\sim\pi_\tau^*}\left[r(a) - \tau \log \pi_\tau^*(a)\right] - \mathop{\mathbb{E}}_{a\sim\pi_\theta}\left[r(a) - \tau \log \pi_\theta(a)\right] \right)^{\frac{1}{2}}, && (617)
\end{aligned}$$

which means the Łojasiewicz degree of $\mathbb{E}_{a\sim\pi_\theta}\left[r(a) - \tau \log \pi_\theta(a)\right]$ is $1/2$ and $C(\theta) = \sqrt{2\tau} \cdot \min_a \pi_\theta(a)$. □

B. Supporting Lemmas

Lemma 19 (Performance difference lemma (Kakade & Langford, 2002))**.** *For any policies π and π',*

$$V^\pi(\rho) - V^{\pi'}(\rho) = \frac{1}{1-\gamma} \sum_s d_\rho^\pi(s) \sum_a \pi(a|s) \cdot A^{\pi'}(s,a). \tag{618}$$

Proof. According to the definition of value function,

$$\begin{aligned}
V^{\pi}(\rho) - V^{\pi'}(\rho) &= \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, a_t \sim \pi(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] - V^{\pi'}(\rho) && (619) \\
&= \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, a_t \sim \pi(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t) + V^{\pi'}(s_t) - V^{\pi'}(s_t)) \right] - V^{\pi'}(\rho) && (620) \\
&= \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, a_t \sim \pi(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t) + \gamma V^{\pi'}(s_{t+1}) - V^{\pi'}(s_t)) \right] && (621) \\
&= \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, a_t \sim \pi(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t A^{\pi'}(s_t, a_t) \right] && (622) \\
&= \frac{1}{1-\gamma} \sum_{s} d_{\rho}^{\pi}(s) \sum_{a} \pi(a|s) \cdot A^{\pi'}(s, a). && \square
\end{aligned}$$

Lemma 20 (Value sub-optimality lemma). *For any policy π,*

$$V^*(\rho) - V^{\pi}(\rho) = \frac{1}{1-\gamma} \sum_{s} d_{\rho}^{\pi}(s) \sum_{a} (\pi^*(a|s) - \pi(a|s)) \cdot Q^*(s, a). \tag{623}$$

Proof. According to the definition of value function,

$$\begin{aligned}
V^*(s) - V^{\pi}(s) &= \sum_{a} \pi^*(a|s) \cdot Q^*(s, a) - \sum_{a} \pi(a|s) \cdot Q^{\pi}(s, a) && (624) \\
&= \sum_{a} (\pi^*(a|s) - \pi(a|s)) \cdot Q^*(s, a) + \sum_{a} \pi(a|s) \cdot (Q^*(s, a) - Q^{\pi}(s, a)) && (625) \\
&= \sum_{a} (\pi^*(a|s) - \pi(a|s)) \cdot Q^*(s, a) + \gamma \sum_{a} \pi(a|s) \sum_{s'} \mathcal{P}(s'|s, a) \left[V^*(s') - V(s') \right] && (626) \\
&= \frac{1}{1-\gamma} \sum_{s'} d_{s}^{\pi}(s') \sum_{a'} (\pi^*(a'|s') - \pi(a'|s')) \cdot Q^*(s', a'). && \square
\end{aligned}$$

Lemma 21 (Spectrum of H matrix). *Let $\pi \in \Delta(\mathcal{A})$. Denote $H(\pi) \coloneqq diag(\pi) - \pi\pi^{\top}$. Let*

$$\pi(1) \le \pi(2) \le \cdots \le \pi(K). \tag{627}$$

Denote the eigenvalues of $H(\pi)$ as

$$\lambda_1 \le \lambda_2 \le \cdots \le \lambda_K. \tag{628}$$

Then we have,

$$\lambda_1 = 0, \tag{629}$$

$$\pi(i-1) \le \lambda_i \le \pi(i), \ i = 2, 3, \dots, K. \tag{630}$$

Proof. According to Golub (1973, Section 5),

$$\pi(1) - \pi^{\top}\pi \le \lambda_1 \le \pi(1), \tag{631}$$

$$\pi(i-1) \le \lambda_i \le \pi(i), \ i = 2, 3, \dots, K. \tag{632}$$

We show $\lambda_1 = 0$. Note

$$H(\pi)\mathbf{1} = (\text{diag}(\pi) - \pi\pi^{\top})\mathbf{1} = \pi - \pi = 0 \cdot \mathbf{1}. \tag{633}$$

Thus $\mathbf{1}$ is an eigenvector of $H(\pi)$ which corresponds to eigenvalue 0. And for any vector $x \in \mathbb{R}^K$,

$$x^\top H(\pi) x = \mathop{\mathbb{E}}_{\pi}[x \odot x] - \Big(\mathop{\mathbb{E}}_{\pi}[x] \Big)^2 = \text{Var}_\pi[x] \ge 0, \tag{634}$$

which means all the eigenvalues of $H(\pi)$ are non-negative. □

Lemma 22. *Let $\pi \in \Delta(\mathcal{A})$. Denote $H(\pi) \coloneqq diag(\pi) - \pi \pi^\top$. For any vector $x \in \mathbb{R}^K$,*

$$\left\| (\mathbf{Id} - H(\pi)) \left(x - \frac{x^\top \mathbf{1}}{K} \cdot \mathbf{1} \right) \right\|_2 \le \Big(1 - \min_a \pi(a) \Big) \cdot \left\| x - \frac{x^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_2, \tag{635}$$

$$\left\| H(\pi) \left(x - \frac{x^\top \mathbf{1}}{K} \cdot \mathbf{1} \right) \right\|_2 \ge \min_a \pi(a) \cdot \left\| x - \frac{x^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_2. \tag{636}$$

Proof. x can be written as linear combination of eigenvectors of $H(\pi)$,

$$x = a_1 \cdot \frac{\mathbf{1}}{\sqrt{K}} + a_2 v_2 + \cdots + a_K v_K \tag{637}$$

$$= \frac{x^\top \mathbf{1}}{K} \cdot \mathbf{1} + a_2 v_2 + \cdots + a_K v_K. \tag{638}$$

Since $H(\pi)$ is symmetric, $\left\{ \frac{\mathbf{1}}{\sqrt{K}}, v_2, \dots, v_K \right\}$ are orthonormal. The last equation is because representation is unique, and

$$a_1 = x^\top \frac{\mathbf{1}}{\sqrt{K}} = \frac{x^\top \mathbf{1}}{\sqrt{K}}. \tag{639}$$

Denote

$$x' \coloneqq x - \frac{x^\top \mathbf{1}}{K} \cdot \mathbf{1} = a_2 v_2 + \cdots + a_K v_K. \tag{640}$$

We have

$$\|x'\|_2^2 = a_2^2 + \cdots + a_K^2. \tag{641}$$

On the other hand,

$$(\mathbf{Id} - H(\pi)) x' = a_2 (1 - \lambda_2) v_2 + \cdots + a_K (1 - \lambda_K) v_K. \tag{642}$$

Therefore

$$\|(\mathbf{Id} - H(\pi)) x'\|_2 = \left(a_2^2 (1 - \lambda_2)^2 + \cdots + a_K^2 (1 - \lambda_K)^2 \right)^{\frac{1}{2}} \tag{643}$$

$$\le \left((a_2^2 + \cdots + a_K^2) \cdot (1 - \lambda_2)^2 \right)^{\frac{1}{2}} \tag{644}$$

$$= (1 - \lambda_2) \cdot \|x'\|_2 \tag{645}$$

$$\le \Big(1 - \min_a \pi(a) \Big) \cdot \|x'\|_2, \tag{646}$$

where the first inequality is by $0 \le \pi(1) \le \lambda_2 \le \cdots \le \lambda_K \le \pi(K) \le 1$, and the last inequality is according to $\lambda_2 \ge \pi(1) = \min_a \pi(a)$, and both are shown in Lemma 21. Similarly,

$$\|H(\pi) x'\|_2 = \left(a_2^2 \lambda_2^2 + \cdots + a_K^2 \lambda_K^2 \right)^{\frac{1}{2}} \tag{647}$$

$$\ge \left((a_2^2 + \cdots + a_K^2) \cdot \lambda_2^2 \right)^{\frac{1}{2}} \tag{648}$$

$$= \lambda_2 \cdot \|x'\|_2 \tag{649}$$

$$\ge \min_a \pi(a) \cdot \|x'\|_2.$$

□

Lemma 23. *Let* $\pi_\theta \coloneqq \mathrm{softmax}(\theta)$ *and* $\pi_{\theta'} \coloneqq \mathrm{softmax}(\theta')$. *Then*

$$\left\| \pi_\theta - \pi_{\theta'} \right\|_\infty \le 2 \cdot \left\| \theta - \theta' \right\|_\infty . \tag{650}$$

Proof. See Xiao et al. (2019, Lemma 5), our proof is for completeness. Since $\pi_\theta, \pi_{\theta'} \in [0,1]^K$,

$$\begin{aligned}
\left\| \pi_\theta - \pi_{\theta'} \right\|_\infty &\le \left\| \log \pi_\theta - \log \pi_{\theta'} \right\|_\infty \qquad (651) \\
&= \left\| \theta - \theta' - \left(\log \sum_a \exp\{\theta(a)\} - \log \sum_a \exp\{\theta'(a)\} \right) \cdot \mathbf{1} \right\|_\infty \qquad (652) \\
&\le \left\| \theta - \theta' \right\|_\infty + \left| \log \sum_a \exp\{\theta(a)\} - \log \sum_a \exp\{\theta'(a)\} \right| \qquad (653) \\
&\le 2 \cdot \left\| \theta - \theta' \right\|_\infty , \qquad (654)
\end{aligned}$$

where the last inequality is according to Nachum et al. (2017, Lemma 8). □

Lemma 24 (Soft sub-optimality lemma)**.** *For any policy* π,

$$\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi}(\rho) = \frac{1}{1-\gamma} \sum_s \left[d_\rho^{\pi}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi(\cdot|s) \| \pi_\tau^*(\cdot|s)) \right]. \tag{655}$$

Proof. According to Nachum et al. (2017, Theorem 1), $\forall (s,a)$,

$$\pi_\tau^*(a|s) = \exp\left\{ (\tilde{Q}^{\pi_\tau^*}(s,a) - \tilde{V}^{\pi_\tau^*}(s))/\tau \right\}. \tag{656}$$

According to the definition of soft value function,

$$\begin{aligned}
\tilde{V}^{\pi_\tau^*}(s) - \tilde{V}^{\pi}(s) &= \tilde{V}^{\pi_\tau^*}(s) - \sum_a \pi(a|s) \cdot \left[\tilde{Q}^{\pi}(s,a) - \tau \log \pi(a|s) \right] \qquad (657) \\
&= \tilde{V}^{\pi_\tau^*}(s) - \sum_a \pi(a|s) \cdot \left[\tilde{Q}^{\pi}(s,a) - \tau \log \pi^*(a|s) + \tau \log \pi^*(a|s) - \tau \log \pi(a|s) \right] \qquad (658) \\
&= \tilde{V}^{\pi_\tau^*}(s) - \sum_a \pi(a|s) \cdot \left[\tilde{Q}^{\pi}(s,a) - \tilde{Q}^{\pi_\tau^*}(s,a) + \tilde{V}^{\pi_\tau^*}(s) \right] + \tau \cdot D_{\mathrm{KL}}(\pi(\cdot|s) \| \pi_\tau^*(\cdot|s)) \qquad (\text{Eq. } (656)) \\
& \qquad (659) \\
&= \tau \cdot D_{\mathrm{KL}}(\pi(\cdot|s) \| \pi_\tau^*(\cdot|s)) + \gamma \sum_a \pi(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \cdot \left[\tilde{V}^{\pi_\tau^*}(s') - \tilde{V}^{\pi}(s') \right] \qquad (660) \\
&= \frac{1}{1-\gamma} \sum_{s'} \left[d_s^{\pi}(s') \cdot \tau \cdot D_{\mathrm{KL}}(\pi(\cdot|s') \| \pi_\tau^*(\cdot|s')) \right].
\end{aligned}$$

□

Lemma 25 (KL-Logit inequality)**.** *Let* $\pi_\theta \coloneqq \mathrm{softmax}(\theta)$ *and* $\pi_{\theta'} \coloneqq \mathrm{softmax}(\theta')$. *Then for any constant* $c \in \mathbb{R}$,

$$D_{\mathrm{KL}}(\pi_\theta \| \pi_{\theta'}) \le \frac{1}{2} \cdot \left\| \theta' - \theta - c \cdot \mathbf{1} \right\|_\infty^2 . \tag{661}$$

In particular, let $c \coloneqq \frac{(\theta' - \theta)^\top \mathbf{1}}{K}$, *we have*

$$D_{\mathrm{KL}}(\pi_\theta \| \pi_{\theta'}) \le \frac{1}{2} \cdot \left\| \theta' - \theta - \frac{(\theta' - \theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty^2 . \tag{662}$$

Proof. According to the ℓ_1 norm strong convexity of negative entropy over probabilistic simplex, i.e., for any policies π, π',

$${\pi'}^\top \log \pi' \ge \pi^\top \log \pi + (\pi' - \pi)^\top \log \pi + \frac{1}{2} \cdot \left\| \pi - \pi' \right\|_1^2 , \tag{663}$$

we have (let $\pi = \pi_\theta$, and $\pi' = \pi_{\theta'}$),

$$\begin{aligned}
D_{\mathrm{KL}}(\pi_\theta \| \pi_{\theta'}) &= \pi_\theta^\top \log \pi_\theta - {\pi_{\theta'}}^\top \log \pi_{\theta'} - (\pi_\theta - \pi_{\theta'})^\top \log \pi_{\theta'} && (664)\\
&\le (\pi_\theta - \pi_{\theta'})^\top \log \pi_\theta - \frac{1}{2} \cdot \| \pi_\theta - \pi_{\theta'} \|_1^2 - (\pi_\theta - \pi_{\theta'})^\top \log \pi_{\theta'} && (665)\\
&= (\pi_\theta - \pi_{\theta'})^\top \left(\log \pi_\theta - \log \pi_{\theta'} \right) - \frac{1}{2} \cdot \| \pi_\theta - \pi_{\theta'} \|_1^2 && (666)\\
&= (\pi_\theta - \pi_{\theta'})^\top \left[\theta - \theta' - \left(\log \sum_a \exp\{\theta(a)\} - \log \sum_a \exp\{\theta'(a)\} \right) \cdot \mathbf{1} \right] - \frac{1}{2} \cdot \| \pi_\theta - \pi_{\theta'} \|_1^2 && (667)\\
&= (\pi_\theta - \pi_{\theta'})^\top \left(\theta - \theta' \right) - \frac{1}{2} \cdot \| \pi_\theta - \pi_{\theta'} \|_1^2 && (668)\\
&= (\pi_\theta - \pi_{\theta'})^\top \left(\theta - \theta' - c \cdot \mathbf{1} \right) - \frac{1}{2} \cdot \| \pi_\theta - \pi_{\theta'} \|_1^2 \qquad \left((\pi_\theta - \pi_{\theta'})^\top c \cdot \mathbf{1} = 0, \ \forall c \in \mathbb{R} \right) && (669)\\
&\le \| \theta - \theta' - c \cdot \mathbf{1} \|_\infty \cdot \| \pi_\theta - \pi_{\theta'} \|_1 - \frac{1}{2} \cdot \| \pi_\theta - \pi_{\theta'} \|_1^2 \qquad \text{(Hölder's inequality)} && (670)\\
&\le \frac{1}{2} \cdot \| \theta - \theta' - c \cdot \mathbf{1} \|_\infty^2 , && (671)
\end{aligned}$$

where the last inequality is according to $ax - bx^2 \le \frac{a^2}{4b}, \forall a, b > 0$. □

Lemma 26 (Reverse Łojasiewicz). *Denote $\Delta^*(s) \coloneqq Q^*(s, a^*(s)) - \max_{a \neq a^*(s)} Q^*(s, a) > 0$ as the optimal value gap of state s, where $a^*(s)$ is the action that the optimal policy selects under state s, and $\Delta^* \coloneqq \min_{s \in \mathcal{S}} \Delta^*(s) > 0$ as the optimal value gap of the MDP. Then we have,*

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \le \frac{1}{1-\gamma} \cdot \frac{\sqrt{2}}{\Delta^*} \cdot \left[V^*(\mu) - V^{\pi_\theta}(\mu) \right]. \tag{672}$$

Proof. Denote $\Delta^*(s, a) \coloneqq Q^*(s, a^*(s)) - Q^*(s, a)$, and $\Delta^*(s) = \min_{a \neq a^*(s)} \Delta^*(s, a)$. We have,

$$\begin{aligned}
V^*(\mu) - V^{\pi_\theta}(\mu) &= \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \sum_a \left(\pi^*(a|s) - \pi_\theta(a|s) \right) \cdot Q^*(s, a) \qquad \text{(Lemma 20)} && (673)\\
&= \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \left[\sum_a \pi_\theta(a|s) \cdot Q^*(s, a^*(s)) - \sum_a \pi_\theta(a|s) \cdot Q^*(s, a) \right] && (674)\\
&= \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \left[\sum_{a \neq a^*(s)} \pi_\theta(a|s) \cdot Q^*(s, a^*(s)) - \sum_{a \neq a^*(s)} \pi_\theta(a|s) \cdot Q^*(s, a) \right] && (675)\\
&= \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \left[\sum_{a \neq a^*(s)} \pi_\theta(a|s) \cdot \Delta^*(s, a) \right] && (676)\\
&\ge \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \left[\sum_{a \neq a^*(s)} \pi_\theta(a|s) \right] \cdot \Delta^*(s). && (677)
\end{aligned}$$

Since $Q^{\pi_\theta}(s,a) \in [0, 1/(1-\gamma)]$, and $V^{\pi_\theta}(s) \in [0, 1/(1-\gamma)]$, we have $|A^{\pi_\theta}(s,a)| \in [0, 1/(1-\gamma)]$. Also,

$$\begin{aligned}
|A^{\pi_\theta}(s, a^*(s))| &= \left| Q^{\pi_\theta}(s, a^*(s)) - \sum_a \pi_\theta(a|s) \cdot Q^{\pi_\theta}(s,a) \right| && (678) \\
&= \left| \sum_{a \neq a^*(s)} \pi_\theta(a|s) \cdot [Q^{\pi_\theta}(s, a^*(s)) - Q^{\pi_\theta}(s,a)] \right| && (679) \\
&\le \sum_{a \neq a^*(s)} \pi_\theta(a|s) \cdot |Q^{\pi_\theta}(s, a^*(s)) - Q^{\pi_\theta}(s,a)| \qquad \text{(triangle inequality)} && (680) \\
&\le \frac{1}{1-\gamma} \sum_{a \neq a^*(s)} \pi_\theta(a|s). \qquad (Q^{\pi_\theta}(s,a) \in [0, 1/(1-\gamma)]) && (681)
\end{aligned}$$

Therefore the ℓ_2 norm of gradient can be upper bounded as

$$\begin{aligned}
\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 &= \frac{1}{1-\gamma} \cdot \left[\sum_s d_\mu^{\pi_\theta}(s)^2 \sum_a \pi_\theta(a|s)^2 \cdot A^{\pi_\theta}(s,a)^2 \right]^{\frac{1}{2}} && (682) \\
&= \frac{1}{1-\gamma} \cdot \left[\sum_s d_\mu^{\pi_\theta}(s)^2 \cdot \left(\pi_\theta(a^*(s)|s)^2 \cdot A^{\pi_\theta}(s, a^*(s))^2 + \sum_{a \neq a^*(s)} \pi_\theta(a|s)^2 \cdot A^{\pi_\theta}(s,a)^2 \right) \right]^{\frac{1}{2}} && (683) \\
&\le \frac{1}{1-\gamma} \cdot \left[\sum_s d_\mu^{\pi_\theta}(s)^2 \cdot \left(1 \cdot \frac{1}{(1-\gamma)^2} \cdot \left[\sum_{a \neq a^*(s)} \pi_\theta(a|s) \right]^2 + \sum_{a \neq a^*(s)} \pi_\theta(a|s)^2 \cdot \frac{1}{(1-\gamma)^2} \right) \right]^{\frac{1}{2}} && (684) \\
&\le \frac{1}{(1-\gamma)^2} \cdot \left[\sum_s d_\mu^{\pi_\theta}(s)^2 \cdot 2 \cdot \left[\sum_{a \neq a^*(s)} \pi_\theta(a|s) \right]^2 \right]^{\frac{1}{2}} \qquad (\|x\|_2 \le \|x\|_1) && (685) \\
&\le \frac{1}{(1-\gamma)^2} \cdot \sqrt{2} \cdot \sum_s d_\mu^{\pi_\theta}(s) \cdot \left[\sum_{a \neq a^*(s)} \pi_\theta(a|s) \right]. \qquad (\|x\|_2 \le \|x\|_1) && (686)
\end{aligned}$$

Combining the results, we have

$$\begin{aligned}
\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 &\le \frac{1}{(1-\gamma)^2} \cdot \sqrt{2} \cdot \sum_s d_\mu^{\pi_\theta}(s) \cdot \left[\sum_{a \neq a^*(s)} \pi_\theta(a|s) \right] && (687) \\
&= \frac{1}{1-\gamma} \cdot \frac{\sqrt{2}}{\Delta^*} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \left[\sum_{a \neq a^*(s)} \pi_\theta(a|s) \right] \cdot \Delta^* && (688) \\
&\le \frac{1}{1-\gamma} \cdot \frac{\sqrt{2}}{\Delta^*} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \left[\sum_{a \neq a^*(s)} \pi_\theta(a|s) \right] \cdot \Delta^*(s) \qquad (\Delta^* \le \Delta^*(s), \ \forall s) && (689) \\
&\le \frac{1}{1-\gamma} \cdot \frac{\sqrt{2}}{\Delta^*} \cdot \left[V^*(\mu) - V^{\pi_\theta}(\mu) \right]. && \square
\end{aligned}$$

C. Sub-optimality Guarantees for Other Entropy-Based RL Methods

Some interesting insight worth mentioning in the proof of Lemma 16 is that the intermediate results provides sub-optimality guarantees for existing entropy regularized RL methods. In particular, Eqs. (486) and (496) provides policy improvement guarantee for Soft Actor-Critic (Haarnoja et al., 2018, SAC), and Eqs. (497) and (502) provide sub-optimality guarantees for Patch Consistency Learning (Nachum et al., 2017, PCL).

Remark 6 (Soft policy improvement inequality)**.** *In Haarnoja et al. (2018, Eq. (4) and Lemma 2), the policy is updated by*

$$\pi_{\theta_{t+1}} = \arg\min_{\pi_\theta} D_{\mathrm{KL}} \left(\pi_\theta(\cdot|s) \middle\| \frac{\exp\left\{ Q^{\pi_{\theta_t}}(s, \cdot) \right\}}{\sum_a \exp\left\{ Q^{\pi_{\theta_t}}(s, a) \right\}} \right), \tag{690}$$

which is exactly the KL divergence in Eq. (496), with $\bar{\pi}_\theta(\cdot|s)$ defined in Eq. (486). The soft policy improvement inequality of Eq. (496) guarantees that if the soft policy improvement is small, then the sub-optimality is small.

Remark 7 (Path inconsistency inequality)**.** *In Nachum et al. (2017, Theorems 1 and 3), it is shown that*

- *(i) soft optimal policy π_τ^* satisfies the consistency conditions Eqs. (26) and (27);*
- *(ii) for any policy π that satisfies the consistency conditions, i.e., if $\forall s, a$,*

$$\pi(a|s) = \exp\left\{ (\tilde{Q}^\pi(s, a) - \tilde{V}^\pi(s))/\tau \right\} \tag{691}$$

$$\tilde{V}^\pi(s) = \tau \log \sum_a \exp\left\{ \tilde{Q}^\pi(s, a)/\tau \right\}, \tag{692}$$

then $\pi = \pi_\tau^*$, and $\tilde{V}^\pi = \tilde{V}^{\pi_\tau^*}$.

However, Nachum et al. (2017) does not show if the consistency is violated during learning, how the violation is related to the sub-optimality. To see why Lemma 16 provides insight, define the following "path inconsistency",

$$r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) \tilde{V}^\pi(s') - \tilde{V}^\pi(s) - \tau \log \pi(a|s) = \tilde{Q}^\pi(s, a) - \tilde{V}^\pi(s) - \tau \log \pi(a|s), \tag{693}$$

which captures the violation of consistency conditions during learning. Note that for softmax policy $\pi_\theta(\cdot|s) := \mathrm{softmax}(\theta(s, \cdot))$, the r.h.s. of Eq. (693) can be written in vector form as

$$\tilde{Q}^{\pi_\theta}(s, \cdot) - \tilde{V}^{\pi_\theta}(s) \cdot \mathbf{1} - \tau \log \pi_\theta(\cdot|s) = \tilde{Q}^{\pi_\theta}(s, \cdot) - \tilde{V}^{\pi_\theta}(s) \cdot \mathbf{1} - \tau \theta(s, \cdot) + \tau \log \sum_a \exp\{\theta(s, a)\} \cdot \mathbf{1}. \tag{694}$$

Denote $c_\theta(s) := \frac{\tilde{V}^{\pi_\theta}(s)}{\tau} - \log \sum_a \exp\{\theta(s, a)\}$, and using Lemma 25 in the proof of Lemma 16, in particular, Eq. (497),

$$D_{\mathrm{KL}}(\pi_\theta(\cdot|s) \| \bar{\pi}_\theta(\cdot|s)) \le \frac{1}{2} \cdot \left\| \frac{\tilde{Q}^{\pi_\theta}(s, \cdot)}{\tau} - \theta(s, \cdot) - c_\theta(s) \cdot \mathbf{1} \right\|_\infty^2 \tag{695}$$

$$= \frac{1}{2\tau^2} \cdot \left\| \tilde{Q}^{\pi_\theta}(s, \cdot) - \tilde{V}^{\pi_\theta}(s) \cdot \mathbf{1} - \tau \log \pi_\theta(\cdot|s) \right\|_\infty^2 . \tag{696}$$

Using the above results in Eq. (502),

$$\left[\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \right]^{\frac{1}{2}} \le \frac{1}{\sqrt{1-\gamma}} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{\frac{1}{2}} \sum_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \left\| \tilde{Q}^{\pi_\theta}(s, \cdot) - \tilde{V}^{\pi_\theta}(s) \cdot \mathbf{1} - \tau \log \pi_\theta(\cdot|s) \right\|_\infty \tag{697}$$

$$= \frac{1}{\sqrt{1-\gamma}} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{\frac{1}{2}} \sum_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \max_a \left| r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) \tilde{V}^{\pi_\theta}(s') - \tau \log \pi_\theta(a|s) - \tilde{V}^{\pi_\theta}(s) \right|, \tag{698}$$

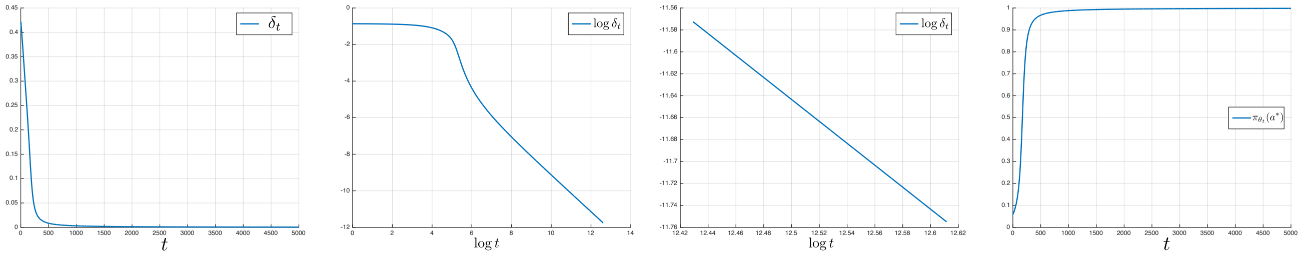
where (square of) $\left| r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) \tilde{V}^{\pi_\theta}(s') - \tau \log \pi_\theta(a|s) - \tilde{V}^{\pi_\theta}(s) \right|$ is exactly the (one-step) path inconsistency objective used in PCL (Nachum et al., 2017, Eq. (14)). Therefore, minimizing path inconsistency guarantees small sub-optimality. The path inconsistency inequality of Eq. (697) implies path consistency of Nachum et al. (2017).

D. Simulation Results

To verify the convergence rates in the main paper, we conducted experiments on one-state MDPs, which have K actions, with randomly generated reward $r \in [0, 1]^K$, and randomly initialized policy π_{θ_1}.

D.1. Softmax Policy Gradient

$K = 20$, $r \in [0,1]^K$ is randomly generated, and π_{θ_1} is randomly initialized. Softmax policy gradient, i.e., Update 1 is used with learning rate $\eta = 2/5$ and $T = 3 \times 10^5$. As shown in Fig. 2(a), the sub-optimality $\delta_t := (\pi^* - \pi_{\theta_t})^\top r$ approaches 0. Subfigures (b) and (c) show $\log \delta_t$ as a function of $\log t$. As $\log t$ increases, the slope is approaching -1, indicating that $\log \delta_t = -\log t + C$, which is equivalent with $\delta_t = C'/t$. Subfigure (d) shows $\pi_{\theta_t}(a^*)$ as a function of t.

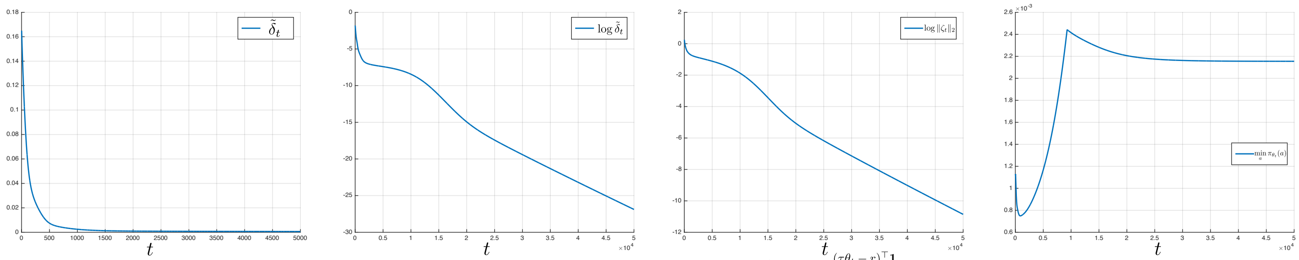


(a) $\delta_t \triangleq (\pi^* - \pi_{\theta_t})^\top r$ (b) (c) slope ≈ -1.0005 (d)

Figure 2. Softmax policy gradient, Update 1.

D.2. Entropy Regularized Softmax Policy Gradient

$K = 20$, $r \in [0,1]^K$ and π_{θ_1} are the same as above. Entropy regularized softmax policy gradient, i.e., Update 2 is used with temperature $\tau = 0.2$, learning rate $\eta = 2/5$ and $T = 5 \times 10^4$. As shown in Fig. 3(a), the soft sub-optimality $\tilde{\delta}_t := \pi_\tau^{*\top}(r - \tau \log \pi_\tau^*) - \pi_{\theta_t}^\top (r - \tau \log \pi_{\theta_t})$ approaches 0. Subfigure (b) shows $\log \tilde{\delta}_t$ as a function of t. As t increases, the curve approaches a straight line, indicating that $\log \tilde{\delta}_t = -C_1 \cdot t + C_2$, which is equivalent with $\tilde{\delta}_t = C_2' / \exp\{C_1' \cdot t\}$. Subfigure (c) shows ζ_t as defined in Lemma 12 as a function of t, which verifies Lemma 13. Subfigure (d) shows $\min_a \pi_{\theta_t}(a)$ as a function of t. As t increases, $\min_a \pi_{\theta_t}(a)$ approaches constant values, which verifies Lemma 14.

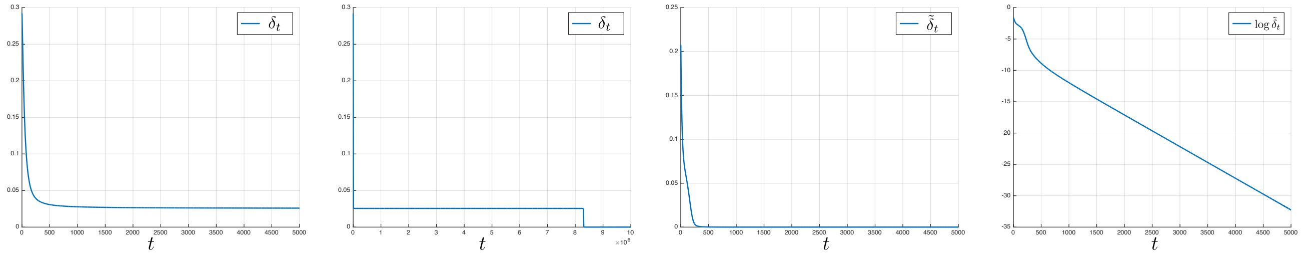


(a) $\tilde{\delta}_t \triangleq \pi_\tau^{*\top}(r - \tau \log \pi_\tau^*) - \pi_{\theta_t}^\top (r - \tau \log \pi_{\theta_t})$ (b) (c) $\zeta_t \triangleq \tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}$ (d)

Figure 3. Entropy regularized softmax policy gradient, Update 2.

D.3. "Bad" Initializations for Softmax Policy Gradient (PG)

As illustrated in Fig. 1, "bad" initializations lead to attraction toward sub-optimal corners and slowly escaping for softmax policy gradient. Fig. 4 shows one example with $K = 5$. Softmax policy gradient takes about 8×10^6 iterations around a sub-optimal corner. While with entropy regularization ($\tau = 0.2$), the convergence is significantly faster.



(a) Softmax policy gradient. (b) Softmax policy gradient. (c) Entropy regularized softmax PG. (d) Entropy regularized softmax PG.

Figure 4. Bad initialization for softmax policy gradient.