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

This paper is concerned with the analysis of the convergence rate of *policy gradient methods* in reinforcement learning (RL). In a policy gradient method the user chooses a policy parametrization. Every policy is assigned a unique numerical score, the value of the policy, through the interaction of the policy with the environment. Policy gradient methods incrementally update the policy parameters by following the gradient of the value. The appeal of policy gradient

methods is that they are conceptually straightforward and under some regularity conditions they guarantee monotonic improvement of the value. A secondary appeal is that policy gradient methods were shown to be able to find good policies for a number of challenging control problems   (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 nonconvexity of the value landscape with respect to standard policy parametrizations. As a result, little has been known about the global convergence behavior of policy gradient ascent. Recently, important new progress in understanding the convergence behavior of policy gradient has been achieved. As in this paper we will restrict ourselves to the tabular setting, we analyze the part of the literature that also deals with this setting. While the tabular setting is clearly limiting, this is the setting where so far the cleanest results have been achieved and we think that understanding this setting is a necessary first step towards the bigger problem of understanding RL algorithms. Returning to the discussion of recent work, in their recent paper Bhandari & Russo (2019) have shown that, without parametrization, projected gradient ascent on the simplex does not suffer from spurious local optima. In concurrent work, Agarwal et al. (2019) complemented the above result by showing that *(i)* without parametrization, projected gradient ascent converges at rate $O(1/\sqrt{t})$ to a global optimum; and *(ii)* with softmax parametrization, policy gradient converges asymptotically. Agarwal et al. also analyze other variants of policy gradient, and show that policy gradient with relative entropy regularization converges at rate $O(1/\sqrt{t})$, natural policy gradient (mirror descent) converges at rate $O(1/t)$, and given a "compatible" function approximation (thus, going beyond the tabular case) natural policy gradient converges at rate $O(1/\sqrt{t})$. Shani et al. (2020) obtains the slower rate $O(1/\sqrt{t})$ for mirror descent. They also proposed a variant that adds entropy regularization and prove a rate of $O(1/t)$ for this modified problem.

Despite these advances, many open questions remain in understanding the behavior of policy gradient methods, even in

---

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

the tabular setting and even when the true gradient is available in the updates. In this paper, we provide answers to the following three questions left open by previous work in this area: *(i)* What is the convergence rate of policy gradient methods with softmax parametrization? The best previous result, due to Agarwal et al. (2019), established asymptotic convergence but gave no rates. *(ii)* What is the convergence rate of entropy regularized softmax policy gradient? Figuring out the answer to this question was explicitly stated as an open problem by Agarwal et al. (2019). *(iii)* Empirical results suggest that entropy helps optimization (Ahmed et al., 2019). Can this empirical observation be turned into a rigorous theoretical result?[1]

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

Third, we provide a theoretical understanding of entropy regularization in policy gradient methods. *(i)* We prove a new lower bound of $\Omega(1/t)$ for softmax policy gradient, implying that the upper bound of $O(1/t)$ that we established for this case, apart from constant factors, is unimprovable. This result also provides a theoretical explanation of the optimization advantage of entropy regularization: even with access to the true gradient, entropy helps policy gradient *converge faster than any achievable rate of softmax policy*

gradient ascent without regularization. *(ii)* We study the concept of non-uniform Łojasiewicz degree and show that, without regularization, the Łojasiewicz degree of expected reward cannot be positive, which 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 $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, the value of state s under π is defined as

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

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

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

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

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

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

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

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

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

---

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

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

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

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

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

H matrix. Given any distribution π over $[K]$, let $H(\pi) := \mathrm{diag}(\pi) - \pi\pi^\top \in \mathbb{R}^{K \times K}$, where $\mathrm{diag}(x) \in \mathbb{R}^{K \times K}$ is the diagonal matrix that has $x \in \mathbb{R}^K$ at its diagonal. The H matrix, which will play a central role in our analysis, e.g., in Section 4, is important because $H(\pi_\theta)$ is the Jacobian of the $\theta \mapsto \pi_\theta := \mathrm{softmax}(\theta)$ map that maps $\mathbb{R}^{[K]}$ to the $(K-1)$-simplex:

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

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

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

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

3. Policy Gradient

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

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

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

3.1. Vanilla Softmax Policy Gradient

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

Algorithm 1 Policy Gradient Method

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

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

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

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

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

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

3.2. Convergence Rate

3.2.1. THE INSTRUCTIVE CASE OF BANDITS

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

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

With the 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 $\mathbb{R}^K$.*

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

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

Here, and in what follows, we will slightly abuse notation in that we identify π_θ and r with vectors in $\mathbb{R}^K$ (say, $r_i = r(i)$, and $(\pi_\theta)_i = \pi_\theta(i)$, $i \in [K]$). Putting things together, we see that in this case the update in Algorithm 1 takes the following form:

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

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

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

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

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

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

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 $\pi_{\theta_t}(a^*)$, the optimal action's probability, becomes very small during the updates.

Nevertheless, the inequality still 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). *Let $c_t = \min_{1 \leq s \leq t} \pi_{\theta_s}(a^*)$. Using Update 1 with $\eta = 2/5$, for all $t \geq 1$,*

$$(\pi^* - \pi_{\theta_t})^\top r \leq 5/(t \cdot c_t^2), \quad \text{and}$$

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

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

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

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

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

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

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

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 $t_0 \geq 1$ such that for any $t \geq t_0$, $t \mapsto \pi_{\theta_t}(a^*)$ 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 all $t \geq 1$,*

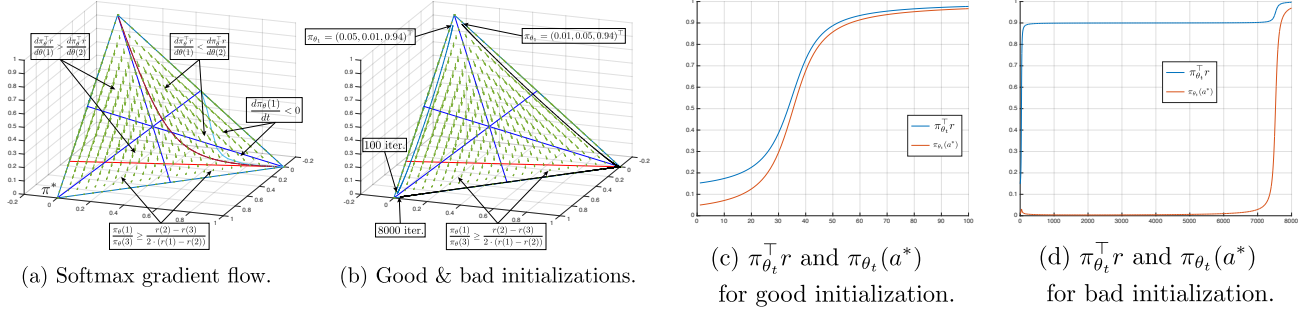$$(\pi^* - \pi_{\theta_t})^\top r \leq 1/(c^2 \cdot t), \quad (13)$$

(a) Softmax gradient flow.

(b) Good & bad initializations.

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

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

Figure 1. Visualization of proof idea for Lemma 5.

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

Proposition 2 suggests that one should set θ_1 so that π_{θ_1} is the uniform distribution. Using this initialization, we can show that $\inf_{t \geq 1} \pi_{\theta_t}(a^*) \geq 1/K$, 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 θ_1 such that $\pi_{\theta_1}(a) = 1/K$, $\forall a$, for all $t \geq 1$,*

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

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

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 the constant C in Theorem 2 depends on the problem and initialization. For the simple 3-armed case, this dependence is relatively clear:

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

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

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

Visualization. Let $r = (1.0, 0.9, 0.1)^\top$. In Fig. 1(a), the region below the red line corresponds to $\mathcal{R} = \{\pi_\theta : \pi_\theta(1)/\pi_\theta(3) \geq (r(2) - r(3))/(2 \cdot (r(1) - r(2)))\}$. Any globally convergent iteration will enter $\mathcal{R}$ within finite time (the closure of $\mathcal{R}$ contains π^*) and never leaves $\mathcal{R}$ (this is the main idea in Lemma 5). Subfigure (b) shows the behavior of the gradient updates with "good" ($\pi_{\theta_1} = (0.05, 0.01, 0.94)^\top$) and "bad" ($\pi_{\theta_1} = (0.01, 0.05, 0.94)^\top$) initial policies. While these are close to each other, the iterates behave quite differently (in both cases $\eta = 2/5$). From the good initialization, the

iterates converge quickly 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! In subfigure (c), we see that $\pi_{\theta_t}(a^*)$ increases for the good initialization, while in subfigure (d), for the bad initialization, we see that it initially decreases. These experiments confirm that the dependence of the error bound in Theorem 2 on the initial values cannot be removed.

Non-unique optimal actions. When the optimal action is non-unique, the 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. Details are given in the appendix.

3.2.2. GENERAL MDPS

For general MDPs, the optimization problem takes the form

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

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

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

This assumption was also adapted by Agarwal et al. (2019). The assumption ensures "sufficient exploration" in the sense that the discounted occupancy measure d_μ^π of any policy π when started from μ will be guaranteed to be positive

over the whole state space. Agarwal et al. (2019) asked whether this assumption is necessary for convergence to global optimality.

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

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

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

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

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

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

Lemma 8 (Non-uniform Łojasiewicz). *Let Assumption 2 hold. Then,*

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

As noted before, by Assumption 2, $d_\mu^{\pi_\theta}$ is also bounded away from zero on the whole state space and thus the multiplier of the sub-optimality on the right-hand side of the above inequality is positive.

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

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

Using Lemmas 7 to 9, we prove that softmax policy gradient

converges to an optimal policy at a $O(1/t)$ rate in MDPs, just like what we have seen in the bandit case:

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

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

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

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

4. Entropy Regularized Policy Gradient

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

4.1. Maximum Entropy RL

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

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

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

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

and $\tau \geq 0$, the "temperature", determines the strength of the penalty.[2] Clearly, the value of any policy can be obtained by adding an entropy penalty to the rewards (as proposed originally by Williams & Peng (1991)). Hence, similarly to Lemma 1, one can obtain the following expression for the gradient of the entropy regularized objective under the softmax policy parametrization:

Lemma 10. *It holds that*

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

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

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

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

4.2. Convergence Rate

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

4.2.1. BANDIT CASE

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

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

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

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

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

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

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

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

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

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

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

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

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

This lemma immediately implies the following bound:

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

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

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

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

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

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

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

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

4.2.2. GENERAL MDPs

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

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

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

---

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

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

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

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

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

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

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

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

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

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

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

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

4.2.3. Controlling the Bias

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

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

$O(\log{(\tau/\Delta)})$ iterations (where Δ is the reward gap) and then turn off regularization. The error bound becomes

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

where $t_1 + t_2 = t$, $t_1 \in O(\log{(\tau/\Delta)})$ and $C \in [1/K, 1]$. This approach removes the nasty dependence on the choice of the initial parameters. While this dependence is also removed if we initialize with the uniform policy, uniform initialization is insufficient if only noisy estimates of the gradients are available. However, we leave the study of this case for future work. An obvious problem with this approach is that Δ is unknown. This can be helped by exiting the first phase when we detect "convergence" e.g. by detecting that the relative change of the policy is small.

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

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

The rationale for the scaling factor is that it allows one to prove a variant of Lemma 11 where ζ_t is redefined as $\zeta_t = \tau_t \theta_t - r - \frac{(\tau_t \theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}$. While this is promising, the proof cannot be finished as before. The difficulty is that $\pi_{\theta_t} \to \pi^*$ (which is what we want to achieve) implies that $\min_a \pi_{\theta_t}(a) \to 0$, which prevents the use of our previous proof technique. We conjecture that the rate in this case degenerates to $O(1/t)$, but the dependence on initialization will be better than beforehand.

5. Does Entropy Regularization Really Help?

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

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

provide a satisfactory solution to the exploration problem (e.g., O'Donoghue et al., 2020).

5.1. Lower Bounds

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

Lemma 17 (Reversed Łojasiewicz). *Take any $r \in [0,1]^K$. Let $\Delta = r(a^*) - \max_{a \neq a^*} r(a) > 0$ be the reward gap of r. Then,*

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

Combining this result with smoothness gives the desired lower bound:

Theorem 7 (Lower bound). *Take any $r \in [0,1]^K$. For large enough $t \geq 1$, using Update 1 with learning rate $\eta_t \in (0,1]$,*

$$(\pi^* - \pi_{\theta_t})^\top r \geq \frac{\Delta^2}{6 \cdot t}. \quad (30)$$

Note that Theorem 7 is a special case of general MDPs. Next, we strengthen this result and show that the $\Omega(1/t)$ lower bound also holds for *any* MDP:

Theorem 8 (Lower bound). *Take any MDP. For large enough $t \geq 1$, using softmax policy gradient Algorithm 1 with $\eta_t \in (0,1]$,*

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

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

Remark 5. *Our convergence rates in Section 3 match the lower bounds up to constant. However, the constant gap is large, e.g., K^2 in Theorem 3, and Δ^2 in Theorem 7. The gap is because the reversed Łojasiewicz inequality of Lemma 17 uses Δ. Note that the appearance of Δ here is unavoidable when π_θ is close to π^*. We leave it as an open problem to close this gap.*

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

5.2. Non-uniform Łojasiewicz Degree

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

Definition 1 (Non-uniform Łojasiewicz degree). *A function $f : \mathcal{X} \to \mathbb{R}$ has Łojasiewicz degree $\xi \in [0,1]$ if[3]*

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

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

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

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

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

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

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

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

6. Conclusions and Future Work

In this paper we set out to study the convergence speed of softmax policy gradient methods with and without entropy regularization in the tabular setting. Here, the error is measured in terms of the sub-optimality of the policy obtained after some number of updates. Our main findings is that without entropy regularization, the rate is $\Theta(1/t)$, which is faster than rates previously obtained. Our analysis also uncovered an unpleasant dependence on the initial parameter values. With entropy regularization, the rate becomes linear,

---

[3]Note that in literature (Łojasiewicz, 1963), C cannot depend on x. Based on the examples we have seen, we relax this requirement.

where now the constant in the exponent is influenced by the initial choice of parameters. Thus, our analysis shows that entropy regularization substantially changes the rate at which gradient methods converge. Our main technical innovation is the introduction of a non-uniform variant of the Łojasiewicz inequality. Our work leaves open a number of interesting questions: While we have some lower bounds, there remains some gaps to be filled between the lower and upper bounds. While entropy regularization makes policy gradient methods converge faster, regularization introduces "bias" and it remains open as to whether removing this bias is possible without degrading the rates. Other interesting directions are extending the results for alternative (e.g., restricted) policy parameterizations or studying policy gradient when the gradient must be estimated from data. One also expects that non-uniform Łojasiewicz inequalities and the Łojasiewicz degree could also be put to good use in other areas of non-convex optimization.

Acknowledgements

References

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

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

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

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

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.

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

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

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

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

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

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.

A. Proofs

A.1. Proofs for Section 3: softmax parametrization

A.1.1. PRELIMINARIES

Lemma 1. Consider the map $\theta \mapsto V^{\pi_\theta}(\mu)$ where $\theta \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ and $\pi_\theta(\cdot|s) = \text{softmax}(\theta(s, \cdot))$. The derivative of this map satisfies

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

Note that this is given as Agarwal et al. (2019, Lemma C.1); we include a proof for completeness.

Proof. According to the policy gradient theorem (Theorem 1),

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

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,

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

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

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

A.1.2. Proofs for softmax parametrization in bandits

Proposition 1. On some problems, $\theta \mapsto \mathbb{E}_{a \sim \pi_\theta}[r(a)]$ 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,

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

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

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

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

Lemma 2 (Smoothness). Let $\pi_\theta = \text{softmax}(\theta)$ and $\pi_{\theta'} = \text{softmax}(\theta')$. For any $r \in [0, 1]^K$, $\theta \mapsto \pi_\theta^\top r$ is 5/2-smooth, i.e.,

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

Proof. Let $S := S(r, \theta) \in \mathbb{R}^{K \times K}$ be the second derivative of the value map $\theta \mapsto \pi_\theta^\top r$. By Taylor's theorem, it suffices to show that the spectral radius of S (regardless of r and θ) is bounded by 5/2. Now, by its definition we have

$$S = \frac{d}{d\theta} \left\{ \frac{d\pi_\theta^\top r}{d\theta} \right\}$$

$$= \frac{d}{d\theta} \{ H(\pi_\theta) r \} \qquad \text{(using Eq. (6))}$$

$$= \frac{d}{d\theta} \left\{ (\text{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top) r \right\}.$$

Continuing with our calculation fix $i, j \in [K]$. Then,

$$S_{i,j} = \frac{d\{\pi_\theta(i) \cdot (r(i) - \pi_\theta^\top r)\}}{d\theta(j)}$$

$$= \frac{d\pi_\theta(i)}{d\theta(j)} \cdot (r(i) - \pi_\theta^\top r) + \pi_\theta(i) \cdot \frac{d\{r(i) - \pi_\theta^\top r\}}{d\theta(j)}$$

$$= (\delta_{ij} \pi_\theta(j) - \pi_\theta(i) \pi_\theta(j)) \cdot (r(i) - \pi_\theta^\top r) - \pi_\theta(i) \cdot (\pi_\theta(j) r(j) - \pi_\theta(j) \pi_\theta^\top r)$$

$$= \delta_{ij} \pi_\theta(j) \cdot (r(i) - \pi_\theta^\top r) - \pi_\theta(i) \pi_\theta(j) \cdot (r(i) - \pi_\theta^\top r) - \pi_\theta(i) \pi_\theta(j) \cdot (r(j) - \pi_\theta^\top r),$$

where

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise} \end{cases} \qquad (33)$$

is Kronecker's δ-function. To show the bound on the spectral radius of S, pick $y \in \mathbb{R}^K$. Then,

$$\left| y^\top S y \right| = \left| \sum_{i=1}^K \sum_{j=1}^K S_{i,j} y(i) y(j) \right|$$

$$= \left| \sum_i \pi_\theta(i)(r(i) - \pi_\theta^\top r) y(i)^2 - 2 \sum_i \pi_\theta(i)(r(i) - \pi_\theta^\top r) y(i) \sum_j \pi_\theta(j) y(j) \right|$$

$$= \left| (H(\pi_\theta) r)^\top (y \odot y) - 2 \cdot (H(\pi_\theta) r)^\top y \cdot (\pi_\theta^\top y) \right|$$

$$\le \|H(\pi_\theta) r\|_\infty \cdot \|y \odot y\|_1 + 2 \cdot \|H(\pi_\theta) r\|_1 \cdot \|y\|_\infty \cdot \|\pi_\theta\|_1 \cdot \|y\|_\infty,$$

where $\odot$ is Hadamard (component-wise) product, and the last inequality uses Hölder's inequality together with the triangle inequality. Note that $\|y \odot y\|_1 = \|y\|_2^2$, $\|\pi_\theta\|_1 = 1$, and $\|y\|_\infty \leq \|y\|_2$. For $i \in [K]$, denote by $H_{i,:}(\pi_\theta)$ the i-th row of $H(\pi_\theta)$ as a row vector. Then,

$$
\begin{aligned}
\|H_{i,:}(\pi_\theta)\|_1 &= \pi_\theta(i) - \pi_\theta(i)^2 + \pi_\theta(i) \cdot \sum_{j \neq i} \pi_\theta(j) \\
&= \pi_\theta(i) - \pi_\theta(i)^2 + \pi_\theta(i) \cdot (1 - \pi_\theta(i)) \\
&= 2 \cdot \pi_\theta(i) \cdot (1 - \pi_\theta(i)) \\
&\leq 1/2. \qquad\qquad \text{(using that } x \cdot (1 - x) \leq 1/4 \text{ holds for } x \in [0, 1])
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
\|H(\pi_\theta)r\|_1 &= \sum_i \pi_\theta(i) \cdot \left| r(i) - \pi_\theta^\top r \right| \\
&\leq \max_i \left| r(i) - \pi_\theta^\top r \right| \\
&\leq 1. \qquad\qquad\qquad\qquad\qquad \text{(using } r \in [0, 1]^K)
\end{aligned}
\tag{34}
$$

Therefore we have,

$$
\begin{aligned}
\left| y^\top S(r, \theta) y \right| &\leq \|H(\pi_\theta)r\|_\infty \cdot \|y\|_2^2 + 2 \cdot \|H(\pi_\theta)r\|_1 \cdot \|y\|_2^2 \\
&= \max_i \left| (H_{i,:}(\pi_\theta))^\top r \right| \cdot \|y\|_2^2 + 2 \cdot \|H(\pi_\theta)r\|_1 \cdot \|y\|_2^2 \\
&\leq \max_i \|H_{i,:}(\pi_\theta)\|_1 \cdot \|r\|_\infty \cdot \|y\|_2^2 + 2 \cdot 1 \cdot \|y\|_2^2 \\
&\leq (1/2 + 2) \cdot \|y\|_2^2 = 5/2 \cdot \|y\|_2^2,
\end{aligned}
\tag{35}
$$

finishing the proof. $\qquad\square$

Lemma 3 (Non-uniform Łojasiewicz). Assume r has a single maximizing action a^*. Let $\pi^* := \arg\max_{\pi \in \Delta} \pi^\top r$, and $\pi_\theta = \mathrm{softmax}(\theta)$. Then, for any θ,

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

When there are multiple optimal actions, we have

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

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

Proof. We give the proof for the general case, as the case of a single maximizing action is a corollary to this case. ¡¡¡¡¡¡¡ HEAD Using the expression we got for the gradient earlier, ======= Using the expression we got for the gradient earlier,

¿¿¿¿¿¿¿ b74314479f0188591a53fac803079c0fa81e5a64

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

For the remaining results in this section, for simplicity, we assume that $\mathcal{A}^* = \{a^*\}$.

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

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

$$\sum_{t=1}^{T} (\pi^* - \pi_{\theta_t})^\top r \le \min\left\{ \sqrt{5T}/c_T,\ (5\log T)/c_T^2 + 1 \right\}. \tag{36}$$

Proof. According to Lemma 2,

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

which implies

$$
\begin{aligned}
\pi_{\theta_t}^\top r - \pi_{\theta_{t+1}}^\top r &\le -\left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle + \frac{5}{4} \cdot \|\theta_{t+1} - \theta_t\|_2^2 \\
&= -\eta \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 + \frac{5}{4} \cdot \eta^2 \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 && \left(\text{using } \theta_{t+1} = \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right) \\
&= -\frac{1}{5} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 && (\text{using } \eta = 2/5) \\
&\le -\frac{1}{5} \cdot \left[\pi_{\theta_t}(a^*) \cdot (\pi^* - \pi_{\theta_t})^\top r \right]^2 && (\text{by Lemma 3}) \\
&\le -\frac{c_t^2}{5} \cdot \left[(\pi^* - \pi_{\theta_t})^\top r \right]^2, && (\text{by the definition of } c_t) \quad (37)
\end{aligned}
$$

which is equivalent to

$$(\pi^* - \pi_{\theta_{t+1}})^\top r - (\pi^* - \pi_{\theta_t})^\top r \le -\frac{c_t^2}{5} \cdot \left[(\pi^* - \pi_{\theta_t})^\top r \right]^2. \tag{38}$$

Let $\delta_t = (\pi^* - \pi_{\theta_t})^\top r$. To prove the first part, we need to show that $\delta_t \le \frac{5}{c_t^2} \cdot \frac{1}{t}$ holds for any $t \ge 1$. We prove this by induction on t.

Base case: Since $\delta_t \le 1$ and $c_t \in (0, 1)$, the result trivially holds up to $t \le 5$.

Inductive step: Now, let $t \ge 2$ and suppose that $\delta_t \le \frac{5}{c_t^2} \cdot \frac{1}{t}$. Consider $f_t : \mathbb{R} \to \mathbb{R}$ defined using $f_t(x) = x - \frac{c_t^2}{5} \cdot x^2$. We have that f_t is monotonically increasing in $\left[0, \frac{5}{2 \cdot c_t^2}\right]$. Hence,

$$
\begin{aligned}
\delta_{t+1} &\le f_t(\delta_t) && (\text{by Eq. (38)}) \\
&\le f_t\left(\frac{5}{c_t^2} \cdot \frac{1}{t} \right) && \left(\text{using } \delta_t \le \frac{5}{c_t^2} \cdot \frac{1}{t} \le \frac{5}{2 \cdot c_t^2},\ t \ge 2\right) \\
&= \frac{5}{c_t^2} \cdot \left(\frac{1}{t} - \frac{1}{t^2} \right) \\
&\le \frac{5}{c_t^2} \cdot \frac{1}{t+1} \\
&\le \frac{5}{c_{t+1}^2} \cdot \frac{1}{t+1}, && (\text{using } c_t \ge c_{t+1} > 0) \quad (39)
\end{aligned}
$$

which completes the induction and the proof of the first part of the lemma.

For the second part, summing up $\delta_t \leq \frac{5}{c_t^2} \cdot \frac{1}{t} \leq \frac{5}{c_T^2} \cdot \frac{1}{t}$, we have

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

On the other hand, rearranging Eq. (38) and summing up $\delta_t^2 \leq \frac{5}{c_t^2} \cdot (\delta_t - \delta_{t+1}) \leq \frac{5}{c_T^2} \cdot (\delta_t - \delta_{t+1})$ from $t = 1$ to T,

$$\begin{aligned}
\sum_{t=1}^{T} \delta_t^2 &\leq \frac{5}{c_T^2} \sum_{t=1}^{T} (\delta_t - \delta_{t+1}) \\
&= \frac{5}{c_T^2} \cdot (\delta_1 - \delta_{T+1}) \\
&\leq \frac{5}{c_T^2}. \quad\quad\quad\quad\quad\quad\quad\text{(since } \delta_{T+1} \geq 0, \delta_1 \leq 1)
\end{aligned}$$

Therefore, by Cauchy-Schwarz,

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

Lemma 5. For $\eta = 2/5$, we have $\inf_{t \geq 1} \pi_{\theta_t}(a^*) > 0$.

Proof. Let

$$c = \frac{K}{2\Delta} \cdot \left(1 - \frac{\Delta}{K}\right)$$

and

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

denote the reward gap of r. We will prove that $\inf_{t \geq 1} \pi_{\theta_t}(a^*) = \min_{1 \leq t \leq t_0} \pi_{\theta_t}(a^*)$, where $t_0 = \min\{t : \pi_{\theta_t}(a^*) \geq \frac{c}{c+1}\}$. Note that t_0 depends only on θ_1 and c, and c depends only on the problem. Define the following regions,

$$\begin{aligned}
\mathcal{R}_1 &= \left\{\theta : \frac{d\pi_\theta^\top r}{d\theta(a^*)} \geq \frac{d\pi_\theta^\top r}{d\theta(a)}, \ \forall a \neq a^*\right\}, \\
\mathcal{R}_2 &= \{\theta : \pi_\theta(a^*) \geq \pi_\theta(a), \ \forall a \neq a^*\}, \\
\mathcal{N}_c &= \left\{\theta : \pi_\theta(a^*) \geq \frac{c}{c+1}\right\}.
\end{aligned}$$

We make the following three-part claim.

Claim 1. *The following hold:*

a) $\mathcal{R}_1$ *is a "nice" region, in the sense that if $\theta_t \in \mathcal{R}_1$ then, with any $\eta > 0$, following a gradient update (i) $\theta_{t+1} \in \mathcal{R}_1$ and (ii) $\pi_{\theta_{t+1}}(a^*) \geq \pi_{\theta_t}(a^*)$.*

b) *We have $\mathcal{R}_2 \subset \mathcal{R}_1$ and $\mathcal{N}_c \subset \mathcal{R}_1$.*

c) *For $\eta = 2/5$, there exists a finite time $t_0 \geq 1$, such that $\theta_{t_0} \in \mathcal{N}_c$, and thus $\theta_{t_0} \in \mathcal{R}_1$, which implies that $\inf_{t \geq 1} \pi_{\theta_t}(a^*) = \min_{1 \leq t \leq t_0} \pi_{\theta_t}(a^*)$.*

Claim a) Part (i): We want to show that if $\theta_t \in \mathcal{R}_1$, then $\theta_{t+1} \in \mathcal{R}_1$. Let

$$\mathcal{R}_1(a) = \left\{ \theta : \frac{d\pi_\theta^\top r}{d\theta(a^*)} \geq \frac{d\pi_\theta^\top r}{d\theta(a)} \right\} .$$

Note that $\mathcal{R}_1 = \cap_{a \neq a^*} \mathcal{R}_1(a)$. Pick $a \neq a^*$. Clearly, it suffices to show that if $\theta_t \in \mathcal{R}_1(a)$ then $\theta_{t+1} \in \mathcal{R}_1(a)$. Hence, suppose that $\theta_t \in \mathcal{R}_1(a)$. We consider two cases.

Case (a): $\pi_{\theta_t}(a^*) \geq \pi_{\theta_t}(a)$. Since $\pi_{\theta_t}(a^*) \geq \pi_{\theta_t}(a)$, we also have $\theta_t(a^*) \geq \theta_t(a)$. After an update of the parameters,

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

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

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

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

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

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

Case (b): Suppose now that $\pi_{\theta_t}(a^*) < \pi_{\theta_t}(a)$. First note that for any θ and $a \neq a^*$, $\theta \in \mathcal{R}_1(a)$ holds if and only if

$$r(a^*) - r(a) \geq \left(1 - \frac{\pi_\theta(a^*)}{\pi_\theta(a)} \right) \cdot \left[r(a^*) - \pi_\theta^\top r \right] . \tag{40}$$

Indeed, from the condition $\frac{d\pi_\theta^\top r}{d\theta(a^*)} \geq \frac{d\pi_\theta^\top r}{d\theta(a)}$, we get

$$\pi_\theta(a^*) \cdot \left[r(a^*) - \pi_\theta^\top r \right] \geq \pi_\theta(a) \cdot \left[r(a) - \pi_\theta^\top r \right]$$

$$= \pi_\theta(a) \cdot \left[r(a^*) - \pi_\theta^\top r \right] - \pi_\theta(a) \cdot \left[r(a^*) - r(a) \right],$$

which, after rearranging, is equivalent to Eq. (40). Hence, it suffices to show that Eq. (40) holds for θ_{t+1} provided it holds for θ_t.

From the latter condition, we get

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

After an update of the parameters, according to the ascent lemma for smooth function (Lemma 18), $\pi_{\theta_{t+1}}^\top r \geq \pi_{\theta_t}^\top r$, i.e.,

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

On the other hand,

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

$$\geq \theta_t(a^*) - \theta_t(a),$$

which implies that

$$1 - \exp\{\theta_{t+1}(a^*) - \theta_{t+1}(a)\} \leq 1 - \exp\{\theta_t(a^*) - \theta_t(a)\}.$$

Furthermore, by our assumption that $\pi_{\theta_t}(a^*) < \pi_{\theta_t}(a)$, we have $1 - \exp\{\theta_t(a^*) - \theta_t(a)\} = 1 - \frac{\pi_{\theta_t}(a^*)}{\pi_{\theta_t}(a)} > 0$. Putting things together, we get

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

$$\leq r(a^*) - r(a),$$

which is equivalent to

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

and thus by our previous remark, $\theta_{t+1} \in \mathcal{R}_1(a)$, thus, finishing the proof of part (i).

Part (ii): Assume again that $\theta_t \in \mathcal{R}_1$. We want to show that $\pi_{\theta_{t+1}}(a^*) \ge \pi_{\theta_t}(a^*)$. Since $\theta_t \in \mathcal{R}_1$, we have $\frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \ge \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)}$, $\forall a \ne a^*$. Hence,

$$
\begin{aligned}
\pi_{\theta_{t+1}}(a^*) &= \frac{\exp\{\theta_{t+1}(a^*)\}}{\sum_a \exp\{\theta_{t+1}(a)\}} \\
&= \frac{\exp\left\{\theta_t(a^*) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)}\right\}}{\sum_a \exp\left\{\theta_t(a) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)}\right\}} \\
&\ge \frac{\exp\left\{\theta_t(a^*) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)}\right\}}{\sum_a \exp\left\{\theta_t(a) + \eta \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)}\right\}} \qquad \left(\text{using } \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a^*)} \ge \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)}\right) \\
&= \frac{\exp\{\theta_t(a^*)\}}{\sum_a \exp\{\theta_t(a)\}} = \pi_{\theta_t}(a^*).
\end{aligned}
$$

Claim b) We start by showing that $\mathcal{R}_2 \subset \mathcal{R}_1$. For this, let $\theta \in \mathcal{R}_2$, i.e., $\pi_\theta(a^*) \ge \pi_\theta(a)$. Then,

$$
\begin{aligned}
\frac{d\pi_\theta^\top r}{d\theta(a^*)} &= \pi_\theta(a^*) \cdot \left[r(a^*) - \pi_\theta^\top r\right] \\
&> \pi_\theta(a) \cdot \left[r(a) - \pi_\theta^\top r\right] \qquad (\text{using } r(a^*) - \pi_\theta^\top r > 0 \text{ and } r(a^*) > r(a)) \\
&= \frac{d\pi_\theta^\top r}{d\theta(a)}.
\end{aligned}
$$

Hence, $\theta \in \mathcal{R}_1$ and thus $\mathcal{R}_2 \subset \mathcal{R}_1$ as desired.

Now, let us prove that $\mathcal{N}_c \subset \mathcal{R}_1$. Take $\theta \in \mathcal{N}_c$. We want to show that $\theta \in \mathcal{R}_1$. If $\theta \in \mathcal{R}_2$, by $\mathcal{R}_2 \subset \mathcal{R}_1$, we also have that $\theta \in \mathcal{R}_1$. Hence, it remains to show that $\theta \in \mathcal{R}_1$ holds when $\theta \in \mathcal{N}_c$ and $\theta \notin \mathcal{R}_2$.

Thus, take any θ that satisfies these two conditions. Pick $a \ne a^*$. It suffices to show that $\theta \in \mathcal{R}_1(a)$. Without loss of generality, assume that $a^* = 1$ and $a = 2$. Then, we have,

$$
\begin{aligned}
\frac{d\pi_\theta^\top r}{d\theta(a^*)} - \frac{d\pi_\theta^\top r}{d\theta(a)} &= \frac{d\pi_\theta^\top r}{d\theta(1)} - \frac{d\pi_\theta^\top r}{d\theta(2)} = \pi_\theta(1) \cdot \left[r(1) - \pi_\theta^\top r\right] - \pi_\theta(2) \cdot \left[r(2) - \pi_\theta^\top r\right] \\
&= 2\pi_\theta(1) \cdot \left[r(1) - \pi_\theta^\top r\right] + \sum_{i=3}^K \pi_\theta(i) \cdot \left[r(i) - \pi_\theta^\top r\right] \qquad (\text{see below}) \\
&= \left(2\pi_\theta(1) + \sum_{i=3}^K \pi_\theta(i)\right) \cdot \left[r(1) - \pi_\theta^\top r\right] - \sum_{i=3}^K \pi_\theta(i) \cdot \left[r(1) - r(i)\right] \\
&\ge \left(2\pi_\theta(1) + \sum_{i=3}^K \pi_\theta(i)\right) \cdot \left[r(1) - \pi_\theta^\top r\right] - \sum_{i=3}^K \pi_\theta(i) \\
&\ge \left(2\pi_\theta(1) + \sum_{i=3}^K \pi_\theta(i)\right) \cdot \frac{\Delta}{K} - \sum_{i=3}^K \pi_\theta(i), \qquad (41)
\end{aligned}
$$

where the second equation is because

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

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

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

$$\geq \sum_{i=2}^{K} \pi_\theta(i) \cdot \Delta \geq \max_{a \neq a^*}\{\pi_\theta(a)\} \cdot \Delta$$

$$\geq \frac{\Delta}{K}. \qquad\qquad \text{(using } \pi_\theta(a^*) < \max_{a \neq a^*}\{\pi_\theta(a)\}, \ \max_{a \neq a^*}\{\pi_\theta(a)\} = \max_a\{\pi_\theta(a)\} \geq \tfrac{1}{K})$$

Plugging $\sum_{i=3}^{K} \pi_\theta(i) = 1 - \pi_\theta(1) - \pi_\theta(2)$ into Eq. (41) and rearranging the resulting expression we get

$$\frac{d\pi_\theta^\top r}{d\theta(a^*)} - \frac{d\pi_\theta^\top r}{d\theta(a)} \geq \pi_\theta(1) \cdot \left(1 + \frac{\Delta}{K}\right) - \left(1 - \frac{\Delta}{K}\right) + \pi_\theta(2) \cdot \left(1 - \frac{\Delta}{K}\right)$$

$$\geq \pi_\theta(2) \cdot \left(1 - \frac{\Delta}{K}\right) \geq 0, \qquad\qquad \text{(using } \theta \in \mathcal{N}_c, \text{ i.e., } \pi_\theta(1) \geq c/(c+1))$$

which implies that $\theta \in \mathcal{R}_1(a)$, thus, finishing the proof.

Claim c) We claim that $\pi_{\theta_t}(a^*) \to 1$ as $t \to \infty$. For this, we wish to use the asymptotic convergence results of Agarwal et al. (2019, Theorem 5.1), which states this, but the stepsize there is $\eta \leq 1/5$ while here we have $\eta = 2/5$. We claim that their asymptotic result still hold with the larger η. In fact, the restriction on η comes from that they can only prove the ascent lemma (Lemma 18) for $\eta \leq 1/5$. Other than this, their proof does not rely on the choice of η. Since we can prove the ascent lemma with $\eta \leq 2/5$ (and in particular with $\eta = 2/5$), their result continues to hold even with $\eta = 2/5$.

Thus, $\pi_{\theta_t}(a^*) \to 1$ as $t \to \infty$. Hence, there exists $t_0 \geq 1$, such that $\pi_{\theta_{t_0}}(a^*) \geq \frac{c}{c+1}$, 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 $\inf_t \pi_{\theta_t}(a^*) = \min_{1 \leq t \leq t_0} \pi_{\theta_t}(a^*)$. t_0 depends on initialization and c, which only depends on the problem. $\qquad\square$

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

Proof. We have $t_0 = \min\{t \geq 1 : \pi_{\theta_t}(a^*) \geq \frac{c}{c+1}\}$, where $c = \frac{K}{2\Delta} \cdot \left(1 - \frac{\Delta}{K}\right)$ in the proof for Lemma 5 satisfies for any $t \geq t_0$, $t \mapsto \pi_{\theta_t}(a^*)$ is increasing.

Now, let θ_1 be so that π_{θ_1} is the uniform distribution. We show that $t_0 = 1$. Recall from Claim 1 that $\mathcal{R}_2$ is the region where the probability of the optimal action exceeds that of the suboptimal ones and $\mathcal{R}_1$ is the region where the gradient of the optimal action exceeds those of the suboptimal ones and that $\mathcal{R}_2 \subset \mathcal{R}_1$. Clearly, $\theta_1 \in \mathcal{R}_2$ and hence also $\theta_1 \in \mathcal{R}_1$. Now, by Part a) of Claim 1, $\mathcal{R}_1$ is invariant under the updates, showing that $t_0 = 1$ holds as required. $\qquad\square$

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

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

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

Proof. According to Lemmas 4 and 5, the claim immediately holds, with $c = \inf_{t \geq 1} \pi_{\theta_t}(a^*) > 0$. $\qquad\square$

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

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

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

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

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

we have $(\pi^* - \pi_{\theta_t})^\top r \leq 5K^2/t, \forall t \geq 1$. The remaining results follow from Eq. (36) and $c_T \geq 1/K$. $\qquad\qquad \square$

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

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

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

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

By Part a) of Claim 1, it suffices to prove that $\theta \in \mathcal{R}_1$. Thus, our goal is to show that any θ such that $\frac{\pi_\theta(1)}{\pi_\theta(3)} \geq \frac{r(2)-r(3)}{2 \cdot (r(1)-r(2))}$ is in fact an element of $\mathcal{R}_1$. Suppose $\frac{\pi_\theta(1)}{\pi_\theta(3)} \geq \frac{r(2)-r(3)}{2 \cdot (r(1)-r(2))}$. There are two cases.

Case (a): If $\frac{\pi_\theta(1)}{\pi_\theta(3)} \geq \frac{r(2)-r(3)}{r(1)-r(2)}$, then we have,

$$r(2) - \pi_\theta^\top r = -\pi_\theta(1) \cdot (r(1) - r(2)) + \pi_\theta(3) \cdot (r(2) - r(3))$$
$$= \pi_\theta(3) \cdot (r(1) - r(2)) \cdot \left[-\frac{\pi_\theta(1)}{\pi_\theta(3)} + \frac{r(2) - r(3)}{r(1) - r(2)}\right]$$
$$\leq 0, \qquad\qquad\qquad (\tfrac{\pi_\theta(1)}{\pi_\theta(3)} \geq \tfrac{r(2)-r(3)}{r(1)-r(2)})$$

which implies,

$$\frac{d\pi_\theta^\top r}{d\theta(1)} - \frac{d\pi_\theta^\top r}{d\theta(2)} = \pi_\theta(1) \cdot \left(r(1) - \pi_\theta^\top r\right) - \pi_\theta(2) \cdot \left(r(2) - \pi_\theta^\top r\right)$$
$$\geq 0 - 0 = 0. \qquad\qquad\qquad (r(1) - \pi_\theta^\top r > 0)$$

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

$$\frac{d\pi_\theta^\top r}{d\theta(1)} - \frac{d\pi_\theta^\top r}{d\theta(3)} = \pi_\theta(1) \cdot \left(r(1) - \pi_\theta^\top r\right) - \pi_\theta(3) \cdot \left(r(3) - \pi_\theta^\top r\right)$$
$$\geq 0 - 0 = 0. \qquad\qquad\qquad (42)$$

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

Case (b): If $\frac{r(2)-r(3)}{2\cdot(r(1)-r(2))} \leq \frac{\pi_\theta(1)}{\pi_\theta(3)} < \frac{r(2)-r(3)}{r(1)-r(2)}$, then we have,

$$
\begin{aligned}
\frac{d\pi_\theta^\top r}{d\theta(1)} - \frac{d\pi_\theta^\top r}{d\theta(2)} &= \pi_\theta(1)\cdot\left(r(1)-\pi_\theta^\top r\right) - \pi_\theta(2)\cdot\left(r(2)-\pi_\theta^\top r\right) \\
&= 2\cdot\pi_\theta(1)\cdot\left(r(1)-\pi_\theta^\top r\right) + \pi_\theta(3)\cdot\left(r(3)-\pi_\theta^\top r\right) \\
&\geq \pi_\theta(3)\cdot\left[\frac{r(2)-r(3)}{r(1)-r(2)}\cdot\left(r(1)-\pi_\theta^\top r\right) + \left(r(3)-\pi_\theta^\top r\right)\right] \qquad (\tfrac{\pi_\theta(1)}{\pi_\theta(3)} \geq \tfrac{r(2)-r(3)}{2\cdot(r(1)-r(2))}) \\
&\geq \pi_\theta(3)\cdot\left[\frac{r(2)-r(3)}{r(1)-r(2)}\cdot\left(r(1)-r(2)\right) + \left(r(3)-\pi_\theta^\top r\right)\right] \\
&= \pi_\theta(3)\cdot\left(r(2)-\pi_\theta^\top r\right) \geq 0,
\end{aligned}
$$

where the second equation is according to

$$
\pi_\theta(1)\cdot\left(r(1)-\pi_\theta^\top r\right) + \pi_\theta(2)\cdot\left(r(2)-\pi_\theta^\top r\right) + \pi_\theta(3)\cdot\left(r(3)-\pi_\theta^\top r\right) = \pi_\theta^\top r - \pi_\theta^\top r = 0,
$$

and the second inequality is because of

$$
\begin{aligned}
r(1)-\pi_\theta^\top r &= (1-\pi_\theta(1))\cdot r(1) - (\pi_\theta(2)\cdot r(2)+\pi_\theta(3)\cdot r(3)) \\
&= \pi_\theta(2)\cdot(r(1)-r(2)) + \pi_\theta(3)\cdot(r(1)-r(3)) \\
&= (\pi_\theta(2)+\pi_\theta(3))\cdot(r(1)-r(2)) + \pi_\theta(3)\cdot(r(2)-r(3)) \\
&> (\pi_\theta(2)+\pi_\theta(3))\cdot(r(1)-r(2)) + \pi_\theta(1)\cdot(r(1)-r(2)) \qquad (\tfrac{\pi_\theta(1)}{\pi_\theta(3)} < \tfrac{r(2)-r(3)}{r(1)-r(2)}) \\
&= r(1)-r(2),
\end{aligned}
$$

and the last inequality is from

$$
\begin{aligned}
r(2)-\pi_\theta^\top r &= \pi_\theta(3)\cdot(r(1)-r(2))\cdot\left[-\frac{\pi_\theta(1)}{\pi_\theta(3)} + \frac{r(2)-r(3)}{r(1)-r(2)}\right] \\
&> 0. \qquad\qquad\qquad\qquad\qquad\qquad\qquad (\tfrac{\pi_\theta(1)}{\pi_\theta(3)} < \tfrac{r(2)-r(3)}{r(1)-r(2)})
\end{aligned}
$$

Now we have $\frac{d\pi_\theta^\top r}{d\theta(1)} \geq \frac{d\pi_\theta^\top r}{d\theta(2)}$. According to Eq. (42), we have $\frac{d\pi_\theta^\top r}{d\theta(1)} \geq \frac{d\pi_\theta^\top r}{d\theta(3)}$. Therefore we have $\theta \in \mathcal{R}_1$. $\qquad\square$

A.1.3. PROOFS FOR SOFTMAX PARAMETRIZATION IN MDPS

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 $s \in \mathcal{S}$,

$$
\begin{aligned}
\sum_a \left|\frac{\partial\pi_{\theta_\alpha}(a|s)}{\partial\alpha}\Big|_{\alpha=0}\right| &= \sum_a \left|\left\langle\frac{\partial\pi_{\theta_\alpha}(a|s)}{\partial\theta_\alpha}\Big|_{\alpha=0}, \frac{\partial\theta_\alpha}{\partial\alpha}\right\rangle\right| \\
&= \sum_a \left|\left\langle\frac{\partial\pi_\theta(a|s)}{\partial\theta}, u\right\rangle\right|.
\end{aligned}
$$

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

$$
\begin{aligned}
\sum_a \left|\frac{\partial\pi_{\theta_\alpha}(a|s)}{\partial\alpha}\Big|_{\alpha=0}\right| &= \sum_a \left|\left\langle\frac{\partial\pi_\theta(a|s)}{\partial\theta(s,\cdot)}, u(s,\cdot)\right\rangle\right| \\
&= \sum_a \pi_\theta(a|s)\cdot\left|u(s,a) - \pi_\theta(\cdot|s)^\top u(s,\cdot)\right| \\
&\leq \max_a |u(s,a)| + \left|\pi_\theta(\cdot|s)^\top u(s,\cdot)\right| \leq 2\cdot\|u\|_2. \qquad (43)
\end{aligned}
$$

Similarly,

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

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

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

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

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

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

where the δ notation is as defined in Eq. (33). Then we have,

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

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

Therefore we have,

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

$$\leq \|u(s,\cdot)\|_2^2 + 2 \cdot \|u(s,\cdot)\|_2^2 + \|u(s,\cdot)\|_2^2 + 2 \cdot \|u(s,\cdot)\|_2^2 \leq 6 \cdot \|u\|_2^2. \tag{44}$$

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

$$[P(\alpha)]_{(s,s')} = \sum_a \pi_{\theta_\alpha}(a|s) \cdot \mathcal{P}(s'|s,a). \tag{45}$$

The derivative w.r.t. α is

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

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

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

The ℓ_∞ norm is upper bounded as

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

$$\leq \max_s \sum_a \sum_{s'} \mathcal{P}(s'|s,a) \cdot \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \Big|_{\alpha=0} \right| \cdot \|x\|_\infty$$

$$= \max_s \sum_a \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \Big|_{\alpha=0} \right| \cdot \|x\|_\infty$$

$$\leq 2 \cdot \|u\|_2 \cdot \|x\|_\infty. \qquad \text{(by Eq. (43))} \tag{46}$$

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

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

The ℓ_∞ norm is upper bounded as

$$
\begin{aligned}
\left\|\frac{\partial^2 P(\alpha)}{\partial \alpha^2}\bigg|_{\alpha=0} x\right\|_\infty &= \max_s \left|\sum_{s'}\sum_a \left[\frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\bigg|_{\alpha=0}\right] \cdot \mathcal{P}(s'|s,a) \cdot x(s')\right| \\
&\leq \max_s \sum_a \sum_{s'} \mathcal{P}(s'|s,a) \cdot \left|\frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\bigg|_{\alpha=0}\right| \cdot \|x\|_\infty \\
&= \max_s \sum_a \left|\frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\bigg|_{\alpha=0}\right| \cdot \|x\|_\infty \\
&\leq 6 \cdot \|u\|_2^2 \cdot \|x\|_\infty. \quad\quad\quad\quad \text{(by Eq. (44))}
\end{aligned}
\tag{47}
$$

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

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

which implies,

$$V^{\pi_{\theta_\alpha}}(s) = e_s^\top M(\alpha) r_{\theta_\alpha}, \tag{48}$$

where

$$M(\alpha) = \left(\mathbf{Id} - \gamma P(\alpha)\right)^{-1}, \tag{49}$$

and $r_{\theta_\alpha} \in \mathbb{R}^S$ for $s \in \mathcal{S}$ is given by

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

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

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

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

$$\mathbf{1} = \frac{1}{1-\gamma} \cdot \left(\mathbf{Id} - \gamma P(\alpha)\right) \mathbf{1} \implies M(\alpha)\mathbf{1} = \frac{1}{1-\gamma} \cdot \mathbf{1},$$

which implies, $\forall i$,

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

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

$$
\begin{aligned}
\|M(\alpha)x\|_\infty &= \max_i \left|[M(\alpha)]_{i,:}^\top x\right| \\
&\leq \max_i \left\|[M(\alpha)]_{i,:}\right\|_1 \cdot \|x\|_\infty \\
&= \frac{1}{1-\gamma} \cdot \|x\|_\infty.
\end{aligned}
\tag{50}
$$

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

$$\|r_{\theta_\alpha}\|_\infty = \max_s |r_{\theta_\alpha}(s)| = \max_s \left| \sum_a \pi_{\theta_\alpha}(a|s) \cdot r(s, a) \right| \leq 1. \tag{51}$$

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

$$\left| \frac{\partial r_{\theta_\alpha}(s)}{\partial \alpha} \right| = \left| \left(\frac{\partial r_{\theta_\alpha}(s)}{\partial \theta_\alpha} \right)^\top \frac{\partial \theta_\alpha}{\partial \alpha} \right|$$

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

$$= \left| \left(H \left(\pi_{\theta_\alpha}(\cdot|s) \right) r(s, \cdot) \right)^\top u(s, \cdot) \right|$$

$$\leq \| H \left(\pi_{\theta_\alpha}(\cdot|s) \right) r(s, \cdot) \|_1 \cdot \| u(s, \cdot) \|_\infty .$$

Similarly to Eq. (34), the ℓ_1 norm is upper bounded as

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

$$\leq \max_a \left| r(s, a) - \pi_{\theta_\alpha}(\cdot|s)^\top r(s, \cdot) \right|$$

$$\leq 1. \qquad \text{(since } r(s, a) \in [0, 1])$$

Therefore we have,

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

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

$$\leq \|u\|_2. \tag{52}$$

Similarly,

$$\left\| \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2} \right\|_\infty = \max_s \left| \frac{\partial^2 r_{\theta_\alpha}(s)}{\partial \alpha^2} \right|$$

$$= \max_s \left| \left(\frac{\partial}{\partial \theta_\alpha} \left\{ \frac{\partial r_{\theta_\alpha}(s)}{\partial \alpha} \right\} \right)^\top \frac{\partial \theta_\alpha}{\partial \alpha} \right|$$

$$= \max_s \left| \left(\frac{\partial^2 r_{\theta_\alpha}(s)}{\partial \theta_\alpha^2} \frac{\partial \theta_\alpha}{\partial \alpha} \right)^\top \frac{\partial \theta_\alpha}{\partial \alpha} \right|$$

$$= \max_s \left| u(s, \cdot)^\top \frac{\partial^2 \{ \pi_{\theta_\alpha}(\cdot|s)^\top r(s, \cdot) \}}{\partial \theta_\alpha(s, \cdot)^2} u(s, \cdot) \right|$$

$$\leq 5/2 \cdot \| u(s, \cdot) \|_2^2 \leq 3 \cdot \|u\|_2^2. \qquad \text{(by Eq. (35))} \tag{53}$$

Taking derivative w.r.t. α in Eq. (48),

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

Taking second derivative w.r.t. α,

$$\frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2} = 2\gamma^2 \cdot e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} + \gamma \cdot e_s^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha}$$

$$+ 2\gamma \cdot e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} + e_s^\top M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}. \tag{54}$$

For the last term,

$$
\begin{aligned}
\left| e_s^\top M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0} \right| &\leq \| e_s \|_1 \cdot \left\| M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0} \right\|_\infty \\
&\leq \frac{1}{1-\gamma} \cdot \left\| \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0} \right\|_\infty && \text{(by Eq. (50))} \\
&\leq \frac{3}{1-\gamma} \cdot \| u \|_2^2. && \text{(by Eq. (53))} \quad (55)
\end{aligned}
$$

For the second last term,

$$
\begin{aligned}
\left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right| &\leq \left\| M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty \\
&\leq \frac{1}{1-\gamma} \cdot \left\| \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty && \text{(by Eq. (50))} \\
&\leq \frac{2 \cdot \| u \|_2}{1-\gamma} \cdot \left\| M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty && \text{(by Eq. (46))} \\
&\leq \frac{2 \cdot \| u \|_2}{(1-\gamma)^2} \cdot \left\| \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty && \text{(by Eq. (50))} \\
&\leq \frac{2 \cdot \| u \|_2}{(1-\gamma)^2} \cdot \| u \|_2 = \frac{2}{(1-\gamma)^2} \cdot \| u \|_2^2. && \text{(by Eq. (52))} \quad (56)
\end{aligned}
$$

For the second term,

$$
\begin{aligned}
\left| e_s^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha} \Big|_{\alpha=0} \right| &\leq \left\| M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \\
&\leq \frac{1}{1-\gamma} \cdot \left\| \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty && \text{(by Eq. (50))} \\
&\leq \frac{6 \cdot \| u \|_2^2}{1-\gamma} \cdot \left\| M(\alpha) r_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty && \text{(by Eq. (47))} \\
&\leq \frac{6 \cdot \| u \|_2^2}{(1-\gamma)^2} \cdot \left\| r_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty && \text{(by Eq. (50))} \\
&\leq \frac{6}{(1-\gamma)^2} \cdot \| u \|_2^2. && \text{(by Eq. (51))} \quad (57)
\end{aligned}
$$

For the first term, according to Eq. (46), Eqs. (50) and (51),

$$
\begin{aligned}
\left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} \Big|_{\alpha=0} \right| &\leq \left\| M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \\
&\leq \frac{1}{1-\gamma} \cdot 2 \cdot \| u \|_2 \cdot \frac{1}{1-\gamma} \cdot 2 \cdot \| u \|_2 \cdot \frac{1}{1-\gamma} \cdot 1 \\
&= \frac{4}{(1-\gamma)^3} \cdot \| u \|_2^2. \quad (58)
\end{aligned}
$$

Combining Eqs. (55) to (58) with Eq. (54),

$$
\begin{aligned}
\left| \frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2} \Big|_{\alpha=0} \right| &\leq 2\gamma^2 \cdot \left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} \Big|_{\alpha=0} \right| + \gamma \cdot \left| e_s^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha} \Big|_{\alpha=0} \right| \\
&\quad + 2\gamma \cdot \left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right| + \left| e_s^\top M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0} \right| \\
&\leq \left(2\gamma^2 \cdot \frac{4}{(1-\gamma)^3} + \gamma \cdot \frac{6}{(1-\gamma)^2} + 2\gamma \cdot \frac{2}{(1-\gamma)^2} + \frac{3}{1-\gamma} \right) \cdot \| u \|_2^2 \\
&\leq \frac{8}{(1-\gamma)^3} \cdot \| u \|_2^2, \quad (59)
\end{aligned}
$$

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

$$
\begin{aligned}
\left| y^\top \frac{\partial^2 V^{\pi_\theta}(s)}{\partial \theta^2} y \right| &= \left| \left(\frac{y}{\|y\|_2} \right)^\top \frac{\partial^2 V^{\pi_\theta}(s)}{\partial \theta^2} \left(\frac{y}{\|y\|_2} \right) \right| \cdot \|y\|_2^2 \\
&\leq \max_{\|u\|_2=1} \left| \left\langle \frac{\partial^2 V^{\pi_\theta}(s)}{\partial \theta^2} u, u \right\rangle \right| \cdot \|y\|_2^2 \\
&= \max_{\|u\|_2=1} \left| \left\langle \frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \theta_\alpha^2} \Big|_{\alpha=0} \frac{\partial \theta_\alpha}{\partial \alpha}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \cdot \|y\|_2^2 \\
&= \max_{\|u\|_2=1} \left| \left\langle \frac{\partial}{\partial \theta_\alpha} \left\{ \frac{\partial V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha} \right\} \Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \cdot \|y\|_2^2 \\
&= \max_{\|u\|_2=1} \left| \frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2} \Big|_{\alpha=0} \right| \cdot \|y\|_2^2 \\
&\leq \frac{8}{(1-\gamma)^3} \cdot \|y\|_2^2. \qquad \text{(by Eq. (59))} \qquad\qquad (60)
\end{aligned}
$$

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

$$
\begin{aligned}
\left| V^{\pi_{\theta'}}(s) - V^{\pi_\theta}(s) - \left\langle \frac{\partial V^{\pi_\theta}(s)}{\partial \theta}, \theta' - \theta \right\rangle \right| &= \frac{1}{2} \cdot \left| (\theta' - \theta)^\top \frac{\partial^2 V^{\pi_{\theta_\xi}}(s)}{\partial \theta_\xi^2} (\theta' - \theta) \right| \\
&\leq \frac{4}{(1-\gamma)^3} \cdot \|\theta' - \theta\|_2^2. \qquad \text{(by Eq. (60))}
\end{aligned}
$$

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

Lemma 8 (Non-uniform Łojasiewicz). Let $\pi_\theta(\cdot|s) = \text{softmax}(\theta(s, \cdot))$, $s \in \mathcal{S}$ and fix an arbitrary optimal policy π^*. Suppose Assumption 2 hold, i.e., $\min_s \mu(s) > 0$. Then,

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

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

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

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

Proof. We have,

$$
\begin{aligned}
\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 &= \left[\sum_{s,a} \left(\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s, a)} \right)^2 \right]^{\frac{1}{2}} \\
&\geq \left[\sum_s \left(\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s, a^*(s))} \right)^2 \right]^{\frac{1}{2}} \\
&\geq \frac{1}{\sqrt{S}} \sum_s \left| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s, a^*(s))} \right| \qquad\qquad \text{(by Cauchy-Schwarz, } \|x\|_1 = |\langle \mathbf{1}, |x| \rangle| \leq \|\mathbf{1}\|_2 \cdot \|x\|_2) \\
&= \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_s \left| d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a^*(s)|s) \cdot A^{\pi_\theta}(s, a^*(s)) \right| \qquad\qquad \text{(by Lemma 1)} \\
&= \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_s d_\mu^{\pi_\theta}(s) \cdot \pi_\theta(a^*(s)|s) \cdot |A^{\pi_\theta}(s, a^*(s))|. \quad \text{(because } d_\mu^{\pi_\theta}(s) \geq 0 \text{ and } \pi_\theta(a^*(s)|s) \geq 0)
\end{aligned}
$$

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

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

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

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

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

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

where the one but last equality used that π^* is deterministic and in state s chooses $a^*(s)$ with probability one, and the last equality uses the performance difference formula (Lemma 19).

To prove the second claim, given a policy π, define the greedy action set for each state s,

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

By similar arguments that were used in the first part, we have,

$$\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \geq \frac{1}{\sqrt{SA}} \sum_{s,a} \left| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s, a)} \right| \qquad \text{(by Cauchy-Schwarz)}$$

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

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

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

where the last inequality is because for any $\bar{a}(s) \in \bar{\mathcal{A}}^{\pi_\theta}(s)$ we have

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

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

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

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

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

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

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

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

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

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

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

$$\mathcal{R}_3(s) = \left\{ \theta_t : V^{\pi_{\theta_t}}(s) \geq Q^{\pi_{\theta_t}}(s, a^*(s)) - \Delta^*(s)/2, \text{ for all } t \geq 1 \text{ large enough} \right\},$$

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

Similarly to the previous proof, we have the following claims:

Claim I. $\mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$ is a "nice" region, in the sense that, following a gradient update, (i) if $\theta_t \in \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$, then $\theta_{t+1} \in \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$; while we also have (ii) $\pi_{\theta_{t+1}}(a^*(s)|s) \geq \pi_{\theta_t}(a^*(s)|s)$.

Claim II. $\mathcal{N}_c(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s) \subset \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$.

Claim III. There exists a finite time $t_0(s) \geq 1$, 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 $\inf_{t \geq 1} \pi_{\theta_t}(a^*(s)|s) = \min_{1 \leq t \leq t_0(s)} \pi_{\theta_t}(a^*(s)|s)$.

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

Clearly, claim IV suffices to prove the lemma since for any θ, $\min_{s,a} \pi_\theta(a|s) > 0$. In what follows we provide the proofs of these four claims.

Claim I. First we prove part (i) of the claim. 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,

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

According to smoothness arguments as Eq. (63), we have $V^{\pi_{\theta_{t+1}}}(s') \geq V^{\pi_{\theta_t}}(s')$, and

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

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

$$\begin{aligned} Q^{\pi_{\theta_t}}(s, a^*(s)) - Q^{\pi_{\theta_t}}(s, a) &= Q^{\pi_{\theta_t}}(s, a^*(s)) - Q^*(s, a^*(s)) + Q^*(s, a^*(s)) - Q^{\pi_{\theta_t}}(s, a) \\ &\geq -\Delta^*(s)/2 + Q^*(s, a^*(s)) - Q^*(s, a) + Q^*(s, a) - Q^{\pi_{\theta_t}}(s, a) \\ &\geq -\Delta^*(s)/2 + Q^*(s, a^*(s)) - \max_{a \neq a^*(s)} Q^*(s, a) + Q^*(s, a) - Q^{\pi_{\theta_t}}(s, a) \\ &= -\Delta^*(s)/2 + \Delta^*(s) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) \cdot [V^*(s') - V^{\pi_{\theta_t}}(s')] \\ &\geq -\Delta^*(s)/2 + \Delta^*(s) + 0 \\ &= \Delta^*(s)/2. \end{aligned} \tag{61}$$

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

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

$$= \frac{1}{1 - \gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a|s) \cdot [Q^{\pi_{\theta_t}}(s, a) - V^{\pi_{\theta_t}}(s)].$$

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

$$\pi_{\theta_t}(a^*(s)|s) \cdot [Q^{\pi_{\theta_t}}(s, a^*(s)) - V^{\pi_{\theta_t}}(s)] \geq \pi_{\theta_t}(a|s) \cdot [Q^{\pi_{\theta_t}}(s, a) - V^{\pi_{\theta_t}}(s)].$$

Similarly to the first part in the proof for Lemma 5. There are two cases.

Case (a): If $\pi_{\theta_t}(a^*(s)|s) \geq \pi_{\theta_t}(a|s)$, then $\theta_t(s, a^*(s)) \geq \theta_t(s, a)$. After an update of the parameters,

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

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

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

$$\pi_{\theta_{t+1}}(a^*(s)|s) \cdot [Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - V^{\pi_{\theta_{t+1}}}(s)] \geq \pi_{\theta_{t+1}}(a|s) \cdot [Q^{\pi_{\theta_{t+1}}}(s, a) - V^{\pi_{\theta_{t+1}}}(s)],$$

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

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

$$\pi_{\theta_t}(a^*(s)|s) \cdot [Q^{\pi_{\theta_t}}(s, a^*(s)) - V^{\pi_{\theta_t}}(s)] \geq \pi_{\theta_t}(a|s) \cdot [Q^{\pi_{\theta_t}}(s, a) - V^{\pi_{\theta_t}}(s)]$$

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

which, after rearranging, is equivalent to

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

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

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

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

On the other hand,

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

$$\geq \theta_t(s, a^*(s)) - \theta_t(s, a),$$

which implies

$$1 - \exp\{\theta_{t+1}(s, a^*(s)) - \theta_{t+1}(s, a)\} \leq 1 - \exp\{\theta_t(s, a^*(s)) - \theta_t(s, a)\}.$$

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

$$(1 - \exp\{\theta_{t+1}(s, a^*(s)) - \theta_{t+1}(s, a)\}) \cdot [Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - V^{\pi_{\theta_{t+1}}}(s)] \leq Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - Q^{\pi_{\theta_{t+1}}}(s, a),$$

which after rearranging is equivalent to

$$\pi_{\theta_{t+1}}(a^*(s)|s) \cdot [Q^{\pi_{\theta_{t+1}}}(s, a^*(s)) - V^{\pi_{\theta_{t+1}}}(s)] \geq \pi_{\theta_{t+1}}(a|s) \cdot [Q^{\pi_{\theta_{t+1}}}(s, a) - V^{\pi_{\theta_{t+1}}}(s)],$$

which means $\frac{\partial V^{\pi_{\theta_{t+1}}}(\mu)}{\partial \theta_{t+1}(s, a^*(s))} \geq \frac{\partial V^{\pi_{\theta_{t+1}}}(\mu)}{\partial \theta_{t+1}(s, a)}$ 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 $\theta_{t+1} \in \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$.

Let us now turn to proving part (ii). We have $\pi_{\theta_{t+1}}(a^*(s)|s) \geq \pi_{\theta_t}(a^*(s)|s)$. If $\theta_t \in \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$, then $\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a^*(s))} \geq \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a)}, \forall a \neq a^*$. After an update of the parameters,

$$
\begin{aligned}
\pi_{\theta_{t+1}}(a^*(s)|s) &= \frac{\exp\{\theta_{t+1}(s, a^*(s))\}}{\sum_a \exp\{\theta_{t+1}(s, a)\}} \\
&= \frac{\exp\left\{\theta_t(s, a^*(s)) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a^*(s))}\right\}}{\sum_a \exp\left\{\theta_t(s, a) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a)}\right\}} \\
&\geq \frac{\exp\left\{\theta_t(s, a^*(s)) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a^*(s))}\right\}}{\sum_a \exp\left\{\theta_t(s, a) + \eta \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a^*(s))}\right\}} && \left(\text{because } \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a^*(s))} \geq \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a)}\right) \\
&= \frac{\exp\{\theta_t(s, a^*(s))\}}{\sum_a \exp\{\theta_t(s, a)\}} = \pi_{\theta_t}(a^*(s)|s).
\end{aligned}
$$

Claim II. $\mathcal{N}_c(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s) \subset \mathcal{R}_1(s) \cap \mathcal{R}_2(s) \cap \mathcal{R}_3(s)$. 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.

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

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

where the inequality is since $Q^{\pi_\theta}(s, a^*(s)) - Q^{\pi_\theta}(s, a) \geq \Delta^*(s)/2 > 0, \forall a \neq a^*(s)$, similarly to Eq. (61).

Case (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,

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

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

Claim III. (1) According to the asymptotic convergence results of Agarwal et al. (2019, Theorem 5.1), which we can use thanks to Assumption 2, $\pi_{\theta_t}(a^*(s)|s) \to 1$. Hence, there exists $t_1(s) \geq 1$, 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 $t_2(s) \geq 1$, 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 $t_3(s) \geq 1$, 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 $t_0(s) = \max\{t_1(s), t_2(s), t_3(s)\}$. 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.

Claim IV. Define $t_0 = \max_s t_0(s)$. Then we have $\inf_{s \in \mathcal{S}, t \geq 1} \pi_{\theta_t}(a^*(s)|s) = \min_{1 \leq t \leq t_0} \min_s \pi_{\theta_t}(a^*(s)|s) > 0$. $\qquad \square$

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

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

Proof. Let us first note that for any θ and μ,

$$
\begin{aligned}
d_\mu^{\pi_\theta}(s) &= \mathop{\mathbb{E}}_{s_0 \sim \mu} \left[d_\mu^{\pi_\theta}(s) \right] \\
&= \mathop{\mathbb{E}}_{s_0 \sim \mu} \left[(1-\gamma) \sum_{t=0}^\infty \gamma^t \Pr(s_t = s | s_0, \pi_\theta, \mathcal{P}) \right] \\
&\geq \mathop{\mathbb{E}}_{s_0 \sim \mu} \left[(1-\gamma) \Pr(s_0 = s | s_0) \right] \\
&= (1-\gamma) \cdot \mu(s).
\end{aligned}
\tag{62}
$$

According to the value sub-optimality lemma of Lemma 21,

$$
\begin{aligned}
V^*(\rho) - V^{\pi_\theta}(\rho) &= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\theta}(s) \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) \\
&= \frac{1}{1-\gamma} \sum_s \frac{d_\rho^{\pi_\theta}(s)}{d_\mu^{\pi_\theta}(s)} \cdot d_\mu^{\pi_\theta}(s) \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) \\
&\leq \frac{1}{1-\gamma} \cdot \left\| \frac{1}{d_\mu^{\pi_\theta}} \right\|_\infty \sum_s d_\mu^{\pi_\theta}(s) \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) \\
&\qquad \text{(because } \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) \geq 0) \\
&\leq \frac{1}{(1-\gamma)^2} \cdot \left\| \frac{1}{\mu} \right\|_\infty \sum_s d_\mu^{\pi_\theta}(s) \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s,a) \quad \text{(by Eq. (62) and } \min_s \mu(s) > 0) \\
&= \frac{1}{1-\gamma} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot [V^*(\mu) - V^{\pi_\theta}(\mu)],
\end{aligned}
$$

where the last equation is again by Lemma 21.

According to Lemma 7, $V^{\pi_\theta}(\mu)$ is β-smooth with $\beta = 8/(1-\gamma)^3$. Denote $\delta_t = V^*(\mu) - V^{\pi_{\theta_t}}(\mu)$. And note $\eta = \frac{(1-\gamma)^3}{8}$. We have,

$$
\begin{aligned}
\delta_{t+1} - \delta_t &= V^{\pi_{\theta_t}}(\mu) - V^{\pi_{\theta_{t+1}}}(\mu) \\
&\leq -\frac{(1-\gamma)^3}{16} \cdot \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2^2 &\text{(by Lemma 18)} \\
&\leq -\frac{(1-\gamma)^3}{16S} \cdot \left\| \frac{d_\mu^{\pi^*}}{d_\mu^{\pi_{\theta_t}}} \right\|_\infty^{-2} \cdot \left[\min_s \pi_{\theta_t}(a^*(s)|s) \right]^2 \cdot [V^*(\mu) - V^{\pi_{\theta_t}}(\mu)]^2 &\text{(by Lemma 8)} \\
&\leq -\frac{(1-\gamma)^5}{16S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-2} \cdot \left[\min_s \pi_{\theta_t}(a^*(s)|s) \right]^2 \cdot \delta_t^2 \\
&\leq -\frac{(1-\gamma)^5}{16S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-2} \cdot \left[\inf_{s \in \mathcal{S}, t \geq 1} \pi_{\theta_t}(a^*(s)|s) \right]^2 \cdot \delta_t^2,
\end{aligned}
\tag{63}
$$

where the second to last inequality is by $d_\mu^{\pi_{\theta_t}}(s) \geq (1-\gamma) \cdot \mu(s)$ (cf. Eq. (62)). According to Lemma 9,

$c = \inf_{s \in \mathcal{S}, t \geq 1} \pi_{\theta_t}(a^*(s)|s) > 0$. Using similar induction arguments as in Eq. (39),

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

which leads to the final result,

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

thus, finishing the proof. □

A.2. Proofs for Section 4: entropy regularized softmax policy gradient

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

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

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

$$= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) \right], \ \forall s \tag{64}$$

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

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

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

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

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

Taking derivative w.r.t. θ,

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

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

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

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

where the second equation is because of

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

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

$$
\begin{aligned}
\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s, \cdot)} &= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \left[\sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta(s, \cdot)} \cdot \left[\tilde{Q}^{\pi_\theta}(s, a) - \tau \log \pi_\theta(a|s) \right] \right] \\
&= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \left(\frac{d\pi(\cdot|s)}{d\theta(s, \cdot)} \right)^\top \left[\tilde{Q}^{\pi_\theta}(s, \cdot) - \tau \log \pi_\theta(\cdot|s) \right] \\
&= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s, \cdot) - \tau \log \pi_\theta(\cdot|s) \right] \qquad \text{(by Eq. (6))} \\
&= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s, \cdot) - \tau \theta(\cdot|s) + \tau \log \sum_a \exp\{\theta(s, a)\} \cdot \mathbf{1} \right] \\
&= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot H(\pi_\theta(\cdot|s)) \left[\tilde{Q}^{\pi_\theta}(s, \cdot) - \tau \theta(\cdot|s) \right]. \qquad (H(\pi_\theta(\cdot|s))\mathbf{1} = \mathbf{0} \text{ in Lemma 22})
\end{aligned}
$$

For each component a, we have

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

Lemma 11 (Non-uniform contraction). Using Update 2 with $\tau\eta \leq 1$, $\forall t \geq 1$,

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

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

Proof. Update 2 can be written as

$$
\begin{aligned}
\theta_{t+1} &= \theta_t - \eta \cdot H(\pi_{\theta_t})(\tau \log \pi_{\theta_t} - r) \\
&= \theta_t - \eta \cdot H(\pi_{\theta_t}) \left[\tau\theta_t - r - \left(\log \sum_a \exp\{\theta_t(a)\} \right) \cdot \mathbf{1} \right] \\
&= \theta_t - \eta \cdot H(\pi_{\theta_t})(\tau\theta_t - r) \\
&= \theta_t - \eta \cdot H(\pi_{\theta_t}) \left(\tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right),
\end{aligned}
$$

where the last two equations are from $H(\pi_{\theta_t})\mathbf{1} = \mathbf{0}$ as shown in Lemma 22. For all $t \geq 1$,

$$
\begin{aligned}
\zeta_{t+1} &= \tau\theta_{t+1} - r - \frac{(\tau\theta_{t+1} - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \\
&= \tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} + \tau(\theta_{t+1} - \theta_t) + \left(\frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} - \frac{(\tau\theta_{t+1} - r)^\top \mathbf{1}}{K} \right) \cdot \mathbf{1} \\
&= \tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} + \tau(\theta_{t+1} - \theta_t) + \frac{\tau(\theta_t - \theta_{t+1})^\top \mathbf{1}}{K} \cdot \mathbf{1}.
\end{aligned}
$$

For the last term,

$$
\frac{\tau(\theta_t - \theta_{t+1})^\top \mathbf{1}}{K} \cdot \mathbf{1} = \frac{\tau}{K} \cdot \left(\eta \cdot H(\pi_{\theta_t}) \left(\tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right) \right)^\top \mathbf{1} \cdot \mathbf{1} = \mathbf{0},
$$

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

$$\zeta_{t+1} = (\mathbf{Id} - \tau\eta \cdot H(\pi_{\theta_t})) \left(\tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right)$$

$$= (\mathbf{Id} - \tau\eta \cdot H(\pi_{\theta_t})) \zeta_t.$$

According to Lemma 23, with $\tau\eta \leq 1$,

$$\|\zeta_{t+1}\|_2 = \|(\mathbf{Id} - \tau\eta \cdot H(\pi_{\theta_t})) \zeta_t\|_2$$

$$\leq \left(1 - \tau\eta \cdot \min_a \pi_{\theta_t}(a) \right) \cdot \|\zeta_t\|_2. \qquad \Box$$

Lemma 12. Let $\pi_{\theta_t} = \mathrm{softmax}(\theta_t)$. Using Update 2 with $\tau\eta \leq 1$, $\forall t \geq 1$,

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

Proof. According to Lemma 11, for all $t \geq 1$,

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

$$\leq \frac{1}{\exp\left\{ \tau\eta \cdot \min_a \pi_{\theta_t}(a) \right\}} \cdot \|\zeta_t\|_2$$

$$\leq \frac{1}{\exp\left\{ \tau\eta \cdot \min_a \pi_{\theta_t}(a) \right\}} \cdot \left(1 - \tau\eta \cdot \min_a \pi_{\theta_{t-1}}(a) \right) \cdot \|\zeta_{t-1}\|_2$$

$$\leq \frac{1}{\exp\left\{ \tau\eta \sum_{s=t-1}^{t} [\min_a \pi_{\theta_s}(a)] \right\}} \cdot \|\zeta_{t-1}\|_2$$

$$\leq \frac{1}{\exp\left\{ \tau\eta \sum_{s=1}^{t} [\min_a \pi_{\theta_s}(a)] \right\}} \cdot \|\zeta_1\|_2.$$

For the initial logit θ_1,

$$\|\zeta_1\|_2 = \left\| \tau\theta_1 - r - \frac{(\tau\theta_1 - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_2$$

$$\leq \|\tau\theta_1 - r\|_2 + \left\| \frac{(\tau\theta_1 - r)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_2 \qquad \text{(by triangle inequality)}$$

$$= \|\tau\theta_1 - r\|_2 + \frac{|(\tau\theta_1 - r)^\top \mathbf{1}|}{\sqrt{K}}$$

$$\leq \|\tau\theta_1 - r\|_2 + \frac{\|\tau\theta_1 - r\|_2 \cdot \|\mathbf{1}\|_2}{\sqrt{K}} \qquad \text{(by Cauchy-Schwarz)}$$

$$= 2 \cdot \|\tau\theta_1 - r\|_2$$

$$\leq 2 \cdot (\|\tau\theta_1\|_2 + \|r\|_2)$$

$$\leq 2(\tau\|\theta_1\|_\infty + 1)\sqrt{K}, \qquad (65)$$

finishing the proof. $\qquad \Box$

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

Proof. Define the constant $c = c(\tau, K, \|\theta_1\|_\infty)$ as

$$c = \frac{1}{K} \cdot \frac{1}{\exp\{1/\tau\}} \cdot \frac{1}{\exp\{4(\|\theta_1\|_\infty + 1/\tau)\sqrt{K}\}}.$$

First, according to Eq. (65), we have,

$$\|\zeta_1\|_2 \le 2(\tau\|\theta_1\|_\infty + 1)\sqrt{K}.$$

Next, according to Lemma 11, with $\tau\eta \le 1$,

$$\|\zeta_{t+1}\|_2 \le \left(1 - \tau\eta \cdot \min_a \pi_{\theta_t}(a)\right) \cdot \|\zeta_t\|_2 \le 2(\tau\|\theta_1\|_\infty + 1)\sqrt{K}.$$

Therefore, for all $t \ge 1$, we have,

$$\|\zeta_t\|_2 \le 2(\tau\|\theta_1\|_\infty + 1)\sqrt{K}.$$

We now prove $\min_a \pi_{\theta_t}(a) \ge c$. We have, $\forall a$,

$$
\begin{aligned}
\left|\theta_t(a) - \frac{r(a)}{\tau} - \frac{(\theta_t - r/\tau)^\top \mathbf{1}}{K}\right| &= \frac{1}{\tau} \cdot \left|\tau\theta_t(a) - r(a) - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K}\right| \\
&\le \frac{1}{\tau} \cdot \left\|\tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} \cdot \mathbf{1}\right\|_2 \\
&= \frac{1}{\tau} \cdot \|\zeta_t\|_2 \\
&\le 2(\|\theta_1\|_\infty + 1/\tau)\sqrt{K}.
\end{aligned}
$$

Denote $a_1 = \arg\min_a \theta_t(a)$, and $a_2 = \arg\max_a \theta_t(a)$. According to the above, we have the following results,

$$\theta_t(a_1) \ge \frac{r(a_1)}{\tau} + \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} - 2(\|\theta_1\|_\infty + 1/\tau)\sqrt{K},$$

$$-\theta_t(a_2) \ge -\frac{r(a_2)}{\tau} - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} - 2(\|\theta_1\|_\infty + 1/\tau)\sqrt{K},$$

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

$$\min_a \pi_{\theta_t}(a) = \frac{\exp\{\theta_t(a_1)\}}{\sum_a \exp\{\theta_t(a)\}} \ge \frac{\exp\{\theta_t(a_1)\}}{\sum_a \exp\{\theta_t(a_2)\}} = \frac{1}{K} \cdot \exp\left\{\theta_t(a_1) - \theta_t(a_2)\right\}, \qquad (\text{since } \theta_t(a) \le \theta_t(a_2), \forall a)$$

which can be further lower bounded using the above results,

$$
\begin{aligned}
\min_a \pi_{\theta_t}(a) &\ge \frac{1}{K} \cdot \exp\left\{\theta_t(a_1) - \theta_t(a_2)\right\} \\
&\ge \frac{1}{K} \cdot \exp\left\{\frac{r(a_1)}{\tau} + \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} - 2(\|\theta_1\|_\infty + 1/\tau)\sqrt{K} - \frac{r(a_2)}{\tau} - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K} - 2(\|\theta_1\|_\infty + 1/\tau)\sqrt{K}\right\} \\
&= \frac{1}{K} \cdot \exp\left\{\frac{r(a_1) - r(a_2)}{\tau} - 4(\|\theta_1\|_\infty + 1/\tau)\sqrt{K}\right\} \\
&\ge \frac{1}{K} \cdot \exp\left\{-\frac{1}{\tau} - 4(\|\theta_1\|_\infty + 1/\tau)\sqrt{K}\right\} \qquad (\text{because } r \in [0,1]^K \text{ and } r(a_1) - r(a_2) \ge -1) \\
&= \frac{1}{K} \cdot \frac{1}{\exp\{1/\tau\}} \cdot \frac{1}{\exp\{4(\|\theta_1\|_\infty + 1/\tau)\sqrt{K}\}} = c. \qquad \square
\end{aligned}
$$

Theorem 5. Let $\pi_{\theta_t} = \text{softmax}(\theta_t)$. Using Update 2 with $\eta \le 1/\tau$, for all $t \ge 1$,

$$\left(\pi_\tau^* - \pi_{\theta_t}\right)^\top r \le \frac{4K^{3/2}(\|\theta_1\|_\infty + 1/\tau)}{\exp\{\tau\eta \cdot c \cdot t\}},$$

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

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

Proof. According to Hölder's inequality,

$$
\begin{aligned}
\left(\pi_\tau^* - \pi_{\theta_t}\right)^\top r &\le \left\|\pi_\tau^* - \pi_{\theta_t}\right\|_\infty \cdot \|r\|_1 \\
&\le K \cdot \left\|\pi_\tau^* - \pi_{\theta_t}\right\|_\infty && \text{(because } r \in [0,1]^K) \\
&= K \cdot \left\|\operatorname{softmax}\left(\frac{r}{\tau}\right) - \operatorname{softmax}\left(\theta_t + \frac{(\tau\theta_t - r)^\top \mathbf{1}}{\tau K}\cdot \mathbf{1}\right)\right\|_\infty \\
&\le \frac{2K}{\tau}\cdot \left\|\tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K}\cdot \mathbf{1}\right\|_\infty && \text{(by Lemma 24)} \\
&\le \frac{2K}{\tau}\cdot \left\|\tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K}\cdot \mathbf{1}\right\|_2 \\
&\le \frac{2K}{\tau}\cdot \frac{2(\tau\|\theta_1\|_\infty + 1)\sqrt{K}}{\exp\left\{\tau\eta \sum_{s=1}^{t-1} [\min_a \pi_{\theta_s}(a)]\right\}} && \text{(by Lemma 12)} \\
&\le \frac{4K^{3/2}}{\tau}\cdot \frac{\tau\|\theta_1\|_\infty + 1}{\exp\left\{\tau\eta\cdot c\cdot t\right\}}. && \text{(by Lemma 13)}
\end{aligned}
$$

On the other hand, we have,

$$
\begin{aligned}
{\pi_\tau^*}^\top (r - \tau\log\pi_\tau^*) - {\pi_{\theta_t}}^\top (r - \tau\log\pi_{\theta_t}) &= {\pi_\tau^*}^\top (r - \tau\log\pi_\tau^*) - {\pi_{\theta_t}}^\top (r - \tau\log\pi_\tau^* + \tau\log\pi_\tau^* - \tau\log\pi_{\theta_t}) \\
&= \left(\pi_\tau^* - \pi_{\theta_t}\right)^\top (r - \tau\log\pi_\tau^*) + \tau\cdot D_{\mathrm{KL}}(\pi_{\theta_t}\|\pi_\tau^*) \\
&= \left(\pi_\tau^* - \pi_{\theta_t}\right)^\top \mathbf{1}\cdot\tau\cdot\log\sum_a \exp\{r(a)/\tau\} + \tau\cdot D_{\mathrm{KL}}(\pi_{\theta_t}\|\pi_\tau^*) \\
&= \tau\cdot D_{\mathrm{KL}}(\pi_{\theta_t}\|\pi_\tau^*) \\
&\le \frac{\tau}{2}\cdot \left\|\theta_t - \frac{r}{\tau} - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{\tau K}\cdot \mathbf{1}\right\|_\infty^2 && \text{(by Lemma 27)} \\
&= \frac{1}{2\tau}\cdot \left\|\tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K}\cdot \mathbf{1}\right\|_\infty^2 \\
&\le \frac{1}{2\tau}\cdot \left\|\tau\theta_t - r - \frac{(\tau\theta_t - r)^\top \mathbf{1}}{K}\cdot \mathbf{1}\right\|_2^2 \\
&\le \frac{1}{2\tau}\cdot \frac{4(\tau\|\theta_1\|_\infty + 1)^2 K}{\exp\left\{2\tau\eta \sum_{s=1}^{t-1} [\min_a \pi_{\theta_s}(a)]\right\}} && \text{(by Lemma 12)} \\
&\le \frac{1}{\tau}\cdot \frac{2(\tau\|\theta_1\|_\infty + 1)^2 K}{\exp\left\{2\tau\eta\cdot c\cdot t\right\}}. && \text{(by Lemma 13)} \qquad \square
\end{aligned}
$$

A.2.1. PROOFS FOR MDPs AND ENTROPY REGULARIZATION

Lemma 14 (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.

Proof. Denote $\mathbb{H}^{\pi_\theta}(s) = \mathbb{H}(s, \pi_\theta)$. Also denote $\theta_\alpha = \theta + \alpha u$, where $\alpha \in \mathbb{R}$ and $u \in \mathbb{R}^{SA}$. According to Eq. (16),

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

which implies,

$$
\mathbb{H}^{\pi_{\theta_\alpha}}(s) = e_s^\top M(\alpha)h_{\theta_\alpha}, \tag{66}
$$

where $M(\alpha) = (\mathbf{Id} - \gamma P(\alpha))^{-1}$ is defined in Eq. (49), $P(\alpha)$ is defined in Eq. (45), and $h_{\theta_\alpha} \in \mathbb{R}^S$ for $s \in \mathcal{S}$ is given by

$$h_{\theta_\alpha}(s) = -\sum_a \pi_{\theta_\alpha}(a|s) \cdot \log \pi_{\theta_\alpha}(a|s). \tag{67}$$

According to Eq. (67), $h_{\theta_\alpha}(s) \in [0, \log A], \forall s$. Then we have,

$$\|h_{\theta_\alpha}\|_\infty = \max_s |h_{\theta_\alpha}(s)| \leq \log A. \tag{68}$$

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

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

The ℓ_1 norm is upper bounded as

$$
\begin{aligned}
\|H(\pi_{\theta_\alpha}(\cdot|s)) \log \pi_{\theta_\alpha}(\cdot|s)\|_1 &= \sum_a \pi_{\theta_\alpha}(a|s) \cdot \left| \log \pi_{\theta_\alpha}(a|s) - \pi_{\theta_\alpha}(\cdot|s)^\top \log \pi_{\theta_\alpha}(\cdot|s) \right| \\
&\leq \sum_a \pi_{\theta_\alpha}(a|s) \cdot \left(|\log \pi_{\theta_\alpha}(a|s)| + \left| \pi_{\theta_\alpha}(\cdot|s)^\top \log \pi_{\theta_\alpha}(\cdot|s) \right| \right) \\
&= -2 \cdot \sum_a \pi_{\theta_\alpha}(a|s) \cdot \log \pi_{\theta_\alpha}(a|s) \leq 2 \cdot \log A. \tag{69}
\end{aligned}
$$

Therefore we have,

$$
\begin{aligned}
\left\| \frac{\partial h_{\theta_\alpha}}{\partial \alpha} \right\|_\infty &= \max_s \left| \frac{\partial h_{\theta_\alpha}(s)}{\partial \alpha} \right| \\
&\leq \max_s \|H(\pi_{\theta_\alpha}(\cdot|s)) \log \pi_{\theta_\alpha}(\cdot|s)\|_1 \cdot \|u(s, \cdot)\|_\infty \\
&\leq 2 \cdot \log A \cdot \|u\|_2. \tag{70}
\end{aligned}
$$

The second derivative w.r.t. α is

$$
\begin{aligned}
\left| \frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \alpha^2} \right| &= \left| \left(\frac{\partial}{\partial \theta_\alpha} \left\{ \frac{\partial h_{\theta_\alpha}(s)}{\partial \alpha} \right\} \right)^\top \frac{\partial \theta_\alpha}{\partial \alpha} \right| \\
&= \left| \left(\frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \theta_\alpha^2} \frac{\partial \theta_\alpha}{\partial \alpha} \right)^\top \frac{\partial \theta_\alpha}{\partial \alpha} \right| \\
&= \left| u(s, \cdot)^\top \frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \theta_\alpha^2(s, \cdot)} u(s, \cdot) \right|.
\end{aligned}
$$

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

$$
\begin{aligned}
T(s, \theta_\alpha) = \frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \theta_\alpha^2(s, \cdot)} &= \frac{\partial}{\partial \theta_\alpha(s, \cdot)} \left\{ \frac{\partial h_{\theta_\alpha}(s)}{\partial \theta_\alpha(s, \cdot)} \right\} \\
&= \frac{\partial}{\partial \theta_\alpha(s, \cdot)} \left\{ \left(\frac{\partial \pi_{\theta_\alpha}(\cdot|s)}{\partial \theta_\alpha(s, \cdot)} \right)^\top \frac{\partial h_{\theta_\alpha}(s)}{\partial \pi_{\theta_\alpha}(\cdot|s)} \right\} \\
&= \frac{\partial}{\partial \theta_\alpha(s, \cdot)} \left\{ H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s)) \right\}.
\end{aligned}
$$

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,

$$
\begin{aligned}
T_{i,j} &= \frac{d\{\pi_{\theta_\alpha}(i|s) \cdot (-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s))\}}{d\theta_\alpha(s,j)} \\
&= \frac{d\pi_{\theta_\alpha}(i|s)}{d\theta_\alpha(s,j)} \cdot (-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s)) + \pi_{\theta_\alpha}(i|s) \cdot \frac{d\{-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s)\}}{d\theta_\alpha(s,j)} \\
&= (\delta_{ij}\pi_{\theta_\alpha}(j|s) - \pi_{\theta_\alpha}(i|s)\pi_{\theta_\alpha}(j|s)) \cdot (-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s)) \\
&\quad + \pi_{\theta_\alpha}(i|s) \cdot \left(-\frac{1}{\pi_{\theta_\alpha}(i|s)} \cdot (\delta_{ij}\pi_{\theta_\alpha}(j|s) - \pi_{\theta_\alpha}(i|s)\pi_{\theta_\alpha}(j|s)) - \pi_{\theta_\alpha}(j|s) \cdot (-\log \pi_{\theta_\alpha}(j|s) - h_{\theta_\alpha}(s))\right) \\
&= \delta_{ij}\pi_{\theta_\alpha}(j|s) \cdot (-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s) - 1) - \pi_{\theta_\alpha}(i|s)\pi_{\theta_\alpha}(j|s) \cdot (-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s) - 1) \\
&\quad - \pi_{\theta_\alpha}(i|s)\pi_{\theta_\alpha}(j|s) \cdot (-\log \pi_{\theta_\alpha}(j|s) - h_{\theta_\alpha}(s)).
\end{aligned}
$$

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

$$
\begin{aligned}
\left|y^\top T(s, \theta_\alpha) y\right| &= \left|\sum_{i=1}^{A}\sum_{j=1}^{A} T_{i,j} y(i) y(j)\right| \\
&\leq \left|\sum_i \pi_{\theta_\alpha}(i|s) \cdot (-\log \pi_{\theta_\alpha}(i|s) - h_{\theta_\alpha}(s) - 1) \cdot y(i)^2\right| \\
&\quad + 2 \cdot \left|\sum_i \pi_{\theta_\alpha}(i|s) \cdot y(i) \sum_j \pi_{\theta_\alpha}(j|s) \cdot (-\log \pi_{\theta_\alpha}(j|s) - h_{\theta_\alpha}(s)) \cdot y(j)\right| + \left(\pi_{\theta_\alpha}(\cdot|s)^\top y\right)^2 \\
&= \left|(H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s)) - \pi_{\theta_\alpha}(\cdot|s))^\top (y \odot y)\right| \\
&\quad + 2 \cdot \left|\left(\pi_{\theta_\alpha}(\cdot|s)^\top y\right) \cdot (H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s)))^\top y\right| + \left(\pi_{\theta_\alpha}(\cdot|s)^\top y\right)^2 \\
&\leq \|H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s))\|_\infty \cdot \|y \odot y\|_1 + \|\pi_{\theta_\alpha}(\cdot|s)\|_\infty \cdot \|y \odot y\|_1 \\
&\quad + 2 \cdot \|\pi_{\theta_\alpha}(\cdot|s)\|_1 \cdot \|y\|_\infty \cdot \|H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s))\|_1 \cdot \|y\|_\infty + \|\pi_{\theta_\alpha}(\cdot|s)\|_2^2 \cdot \|y\|_2^2,
\end{aligned}
$$

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

$$
\begin{aligned}
\|H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s))\|_\infty &= \max_a \left|\pi_{\theta_\alpha}(a|s) \cdot \left(-\log \pi_{\theta_\alpha}(a|s) + \pi_{\theta_\alpha}(\cdot|s)^\top \log \pi_{\theta_\alpha}(\cdot|s)\right)\right| \\
&\leq \max_a -\pi_{\theta_\alpha}(a|s) \cdot \log \pi_{\theta_\alpha}(a|s) - \pi_{\theta_\alpha}(\cdot|s)^\top \log \pi_{\theta_\alpha}(\cdot|s) \\
&\leq \frac{1}{e} + \log A. \qquad \left(\text{since } -x \cdot \log x \leq \frac{1}{e} \text{ for all } x \in [0,1]\right) \qquad (71)
\end{aligned}
$$

Therefore we have,

$$
\begin{aligned}
\left|y^\top T(s, \theta_\alpha) y\right| &\leq \|H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s))\|_\infty \cdot \|y\|_2^2 + \|y\|_2^2 + 2 \cdot \|H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s))\|_1 \cdot \|y\|_2^2 + \|y\|_2^2 \\
&\leq \left(\frac{1}{e} + \log A + 2\right) \cdot \|y\|_2^2 + 2 \cdot \|H(\pi_{\theta_\alpha}(\cdot|s))(-\log \pi_{\theta_\alpha}(\cdot|s))\|_1 \cdot \|y\|_2^2 \qquad \text{(by Eq. (71))} \\
&\leq \left(\frac{1}{e} + \log A + 2 + 2 \cdot \log A\right) \cdot \|y\|_2^2 \qquad \text{(by Eq. (69))} \\
&\leq 3 \cdot (1 + \log A) \cdot \|y\|_2^2.
\end{aligned}
$$

According to the above results,

$$
\begin{aligned}
\left\| \frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2} \right\|_\infty &= \max_s \left| \frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \alpha^2} \right| \\
&= \max_s \left| u(s,\cdot)^\top \frac{\partial^2 h_{\theta_\alpha}(s)}{\partial \theta_\alpha^2(s,\cdot)} u(s,\cdot) \right| \\
&= \max_s \left| u(s,\cdot)^\top T(s,\theta_\alpha) u(s,\cdot) \right| \\
&\leq 3 \cdot (1 + \log A) \cdot \max_s \| u(s,\cdot) \|_2^2 \\
&\leq 3 \cdot (1 + \log A) \cdot \| u \|_2^2.
\end{aligned}
\tag{72}
$$

Taking derivative w.r.t. α in Eq. (66),

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

Taking second derivative w.r.t. α,

$$
\begin{aligned}
\frac{\partial^2 \mathbb{H}^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2} &= 2\gamma^2 \cdot e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) h_{\theta_\alpha} + \gamma \cdot e_s^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) h_{\theta_\alpha} \\
&\quad + 2\gamma \cdot e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial \alpha} + e_s^\top M(\alpha) \frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2}.
\end{aligned}
\tag{73}
$$

For the last term,

$$
\begin{aligned}
\left| e_s^\top M(\alpha) \frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0} \right| &\leq \| e_s \|_1 \cdot \left\| M(\alpha) \frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0} \right\|_\infty \\
&\leq \frac{1}{1-\gamma} \cdot \left\| \frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0} \right\|_\infty && \text{(by Eq. (50))} \\
&\leq \frac{3 \cdot (1 + \log A)}{1-\gamma} \cdot \| u \|_2^2. && \text{(by Eq. (72))}
\end{aligned}
\tag{74}
$$

For the second last term,

$$
\begin{aligned}
\left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right| &\leq \left\| M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty \\
&\leq \frac{1}{1-\gamma} \cdot \left\| \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty && \text{(by Eq. (50))} \\
&\leq \frac{2 \cdot \| u \|_2}{1-\gamma} \cdot \left\| M(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty && \text{(by Eq. (46))} \\
&\leq \frac{2 \cdot \| u \|_2}{(1-\gamma)^2} \cdot \left\| \frac{\partial h_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right\|_\infty && \text{(by Eq. (50))} \\
&\leq \frac{2 \cdot \| u \|_2}{(1-\gamma)^2} \cdot 2 \cdot \log A \cdot \| u \|_2 = \frac{4 \cdot \log A}{(1-\gamma)^2} \cdot \| u \|_2^2. && \text{(by Eq. (70))}
\end{aligned}
\tag{75}
$$

For the second term,

$$
\begin{aligned}
\left| e_s^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right| &\leq \left\| M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \\
&\leq \frac{1}{1-\gamma} \cdot \left\| \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty && \text{(by Eq. (50))} \\
&\leq \frac{6 \cdot \|u\|_2^2}{1-\gamma} \cdot \left\| M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty && \text{(by Eq. (47))} \\
&\leq \frac{6 \cdot \|u\|_2^2}{(1-\gamma)^2} \cdot \left\| h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty && \text{(by Eq. (50))} \\
&\leq \frac{6 \cdot \log A}{(1-\gamma)^2} \cdot \|u\|_2^2. && \text{(by Eq. (68))} \qquad (76)
\end{aligned}
$$

For the first term, according to Eqs. (46) and (50), Eq. (68),

$$
\begin{aligned}
\left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right| &\leq \left\| M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right\|_\infty \\
&\leq \frac{1}{1-\gamma} \cdot 2 \cdot \|u\|_2 \cdot \frac{1}{1-\gamma} \cdot 2 \cdot \|u\|_2 \cdot \frac{1}{1-\gamma} \cdot \log A \\
&= \frac{4 \cdot \log A}{(1-\gamma)^3} \cdot \|u\|_2^2. \qquad (77)
\end{aligned}
$$

Combining Eqs. (74) to (77) with Eq. (73),

$$
\begin{aligned}
\left| \frac{\partial^2 \mathbb{H}^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2} \Big|_{\alpha=0} \right| &\leq 2\gamma^2 \cdot \left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right| + \gamma \cdot \left| e_s^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) h_{\theta_\alpha} \Big|_{\alpha=0} \right| \\
&\quad + 2\gamma \cdot \left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial h_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0} \right| + \left| e_s^\top M(\alpha) \frac{\partial^2 h_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0} \right| \\
&\leq \left(2\gamma^2 \cdot \frac{4 \cdot \log A}{(1-\gamma)^3} + \gamma \cdot \frac{6 \cdot \log A}{(1-\gamma)^2} + 2\gamma \cdot \frac{4 \cdot \log A}{(1-\gamma)^2} + \frac{3 \cdot (1 + \log A)}{1-\gamma} \right) \cdot \|u\|_2^2 \\
&\leq \left(\frac{8 \cdot \log A}{(1-\gamma)^3} + \frac{3}{1-\gamma} \right) \cdot \|u\|_2^2 \\
&\leq \frac{4 + 8 \cdot \log A}{(1-\gamma)^3} \cdot \|u\|_2^2, \qquad (78)
\end{aligned}
$$

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

$$
\begin{aligned}
\left| y^\top \frac{\partial^2 \mathbb{H}^{\pi_\theta}(s)}{\partial \theta^2} y \right| &= \left| \left(\frac{y}{\|y\|_2} \right)^\top \frac{\partial^2 \mathbb{H}^{\pi_\theta}(s)}{\partial \theta^2} \left(\frac{y}{\|y\|_2} \right) \right| \cdot \|y\|_2^2 \\
&\leq \max_{\|u\|_2=1} \left| \left\langle \frac{\partial^2 \mathbb{H}^{\pi_\theta}(s)}{\partial \theta^2} u, u \right\rangle \right| \cdot \|y\|_2^2 \\
&= \max_{\|u\|_2=1} \left| \left\langle \frac{\partial^2 \mathbb{H}^{\pi_{\theta_\alpha}}(s)}{\partial \theta_\alpha^2} \Big|_{\alpha=0} \frac{\partial \theta_\alpha}{\partial \alpha}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \cdot \|y\|_2^2 \\
&= \max_{\|u\|_2=1} \left| \left\langle \frac{\partial}{\partial \theta_\alpha} \left\{ \frac{\partial \mathbb{H}^{\pi_{\theta_\alpha}}(s)}{\partial \alpha} \right\} \Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \cdot \|y\|_2^2 \\
&= \max_{\|u\|_2=1} \left| \frac{\partial^2 \mathbb{H}^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2} \Big|_{\alpha=0} \right| \cdot \|y\|_2^2 \\
&\leq \frac{4 + 8 \cdot \log A}{(1-\gamma)^3} \cdot \|y\|_2^2. && \text{(by Eq. (78))} \qquad (79)
\end{aligned}
$$

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

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

$$\leq \frac{2 + 4 \cdot \log A}{(1 - \gamma)^3} \cdot \|\theta' - \theta\|_2^2. \qquad \text{(by Eq. (79))}$$

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

Lemma 15 (Non-uniform Łojasiewicz). Suppose $\mu(s) > 0$ for all states $s \in \mathcal{S}$ and $\pi_\theta(\cdot|s) = \text{softmax}(\theta(s, \cdot))$. Then,

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

Proof. According to the definition of soft value functions,

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

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

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

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

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

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

$$\bar{\pi}_\theta(a|s) = \frac{\exp\left\{ \tilde{Q}^{\pi_\theta}(s, a)/\tau \right\}}{\sum_{a'} \exp\left\{ \tilde{Q}^{\pi_\theta}(s, a')/\tau \right\}}, \quad \forall a. \qquad (80)$$

We have, $\forall s$,

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

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

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

Also note that,

$$
\begin{aligned}
\tilde{V}^{\pi_\theta}(s) &= \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) \right] \\
&= \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \bar{\pi}_\theta(a|s) + \tau \log \bar{\pi}_\theta(a|s) - \tau \log \pi_\theta(a|s) \right] \\
&= \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \bar{\pi}_\theta(a|s) \right] - \tau D_{\mathrm{KL}}(\pi_\theta(\cdot|s) \| \bar{\pi}_\theta(\cdot|s)) \\
&= \tau \log \sum_a \exp\left\{ \tilde{Q}^{\pi_\theta}(s,a)/\tau \right\} - \tau \cdot D_{\mathrm{KL}}(\pi_\theta(\cdot|s) \| \bar{\pi}_\theta(\cdot|s)).
\end{aligned}
$$

Combining the above,

$$
\begin{aligned}
\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) &= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \left[\sum_a \pi_\tau^*(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\tau^*(a|s) \right] - \tilde{V}^{\pi_\theta}(s) \right] \\
&\le \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \left[\tau \log \sum_a \exp\left\{ \tilde{Q}^{\pi_\theta}(s,a)/\tau \right\} - \tilde{V}^{\pi_\theta}(s) \right] \\
&= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_\theta(\cdot|s) \| \bar{\pi}_\theta(\cdot|s)) \quad (81) \\
&\le \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \frac{\tau}{2} \cdot \left\| \frac{\tilde{Q}^{\pi_\theta}(s,\cdot)}{\tau} - \theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot)/\tau - \theta(s,\cdot))^\top \mathbf{1}}{A} \cdot \mathbf{1} \right\|_\infty^2 \quad \text{(by Lemma 27)} \quad (82) \\
&= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \frac{1}{2\tau} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{A} \cdot \mathbf{1} \right\|_\infty^2,
\end{aligned}
$$

where $A = |\mathcal{A}|$ is the total number of actions. Taking square root of soft sub-optimality,

$$
\begin{aligned}
\left[\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \right]^{\frac{1}{2}} &\le \frac{1}{\sqrt{1-\gamma}} \cdot \left[\sum_s d_\rho^{\pi_\tau^*}(s) \cdot \frac{1}{2\tau} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{A} \cdot \mathbf{1} \right\|_\infty^2 \right]^{\frac{1}{2}} \\
&= \frac{1}{\sqrt{1-\gamma}} \cdot \left[\sum_s \left(\sqrt{d_\rho^{\pi_\tau^*}(s)} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{A} \cdot \mathbf{1} \right\|_\infty \right)^2 \right]^{\frac{1}{2}} \\
&\le \frac{1}{\sqrt{1-\gamma}} \cdot \sum_s \sqrt{d_\rho^{\pi_\tau^*}(s)} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{A} \cdot \mathbf{1} \right\|_\infty \quad \text{(by } \|x\|_2 \le \|x\|_1) \\
&\le \frac{1}{\sqrt{1-\gamma}} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{\frac{1}{2}} \sum_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot))^\top \mathbf{1}}{A} \cdot \mathbf{1} \right\|_\infty. \quad (83)
\end{aligned}
$$

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

$$
\begin{aligned}
\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 &= \left[\sum_{s,a} \left(\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,a)} \right)^2 \right]^{\frac{1}{2}} \\
&= \left[\sum_s \left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} \right\|_2^2 \right]^{\frac{1}{2}} \\
&\ge \frac{1}{\sqrt{S}} \sum_s \left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} \right\|_2, \quad \text{(by Cauchy-Schwarz, } \|x\|_1 = |\langle \mathbf{1}, |x| \rangle| \le \|\mathbf{1}\|_2 \cdot \|x\|_2)
\end{aligned}
$$

which is further lower bounded as

$$\left\|\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 \geq \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \left\|H(\pi_\theta(\cdot|s))\left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot)\right]\right\|_2 \qquad \text{(by Eq. (64), Lemma 10)}$$

$$= \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \left\|H(\pi_\theta(\cdot|s))\left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot)-\tau\theta(s,\cdot))^\top\mathbf{1}}{A}\cdot\mathbf{1}\right]\right\|_2$$
$$\text{(by Lemma 22)}$$

$$\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \min_a \pi_\theta(a|s) \cdot \left\|\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot)-\tau\theta(s,\cdot))^\top\mathbf{1}}{A}\cdot\mathbf{1}\right\|_2 \quad \text{(by Lemma 23)}$$

$$\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \min_a \pi_\theta(a|s) \cdot \left\|\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau\theta(s,\cdot) - \frac{(\tilde{Q}^{\pi_\theta}(s,\cdot)-\tau\theta(s,\cdot))^\top\mathbf{1}}{A}\cdot\mathbf{1}\right\|_\infty.$$

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

$$\left\|\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta}\right\|_2 \geq \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \min_a \pi_\theta(a|s) \cdot \|\zeta_\theta(s)\|_\infty$$

$$\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{\sqrt{1-\gamma}} \cdot \min_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \min_{s,a} \pi_\theta(a|s) \cdot \sqrt{2\tau} \cdot \left\|\frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{-\frac{1}{2}} \cdot \left[\frac{1}{\sqrt{1-\gamma}} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\|\frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}}\right\|_\infty^{\frac{1}{2}} \sum_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \|\zeta_\theta(s)\|_\infty\right]$$

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

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

where the last inequality is by $d_\mu^{\pi_\theta}(s) \geq (1-\gamma)\cdot\mu(s)$ (cf. Eq. (62)). $\qquad\qquad\square$

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

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 14. It follows then that $\tilde{V}^{\pi_{\theta_t}}(\rho)$ is upper bounded. Indeed,

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

$$= \frac{1}{1-\gamma}\sum_s d_\rho^{\pi_{\theta_t}}(s)\cdot\left[\sum_a \pi_{\theta_t}(a|s)\cdot(r(s,a) - \tau\log\pi_{\theta_t}(a|s))\right]$$

$$\leq \frac{1}{1-\gamma}\sum_s d_\rho^{\pi_{\theta_t}}(s)\cdot(1 + \tau\log A) \qquad\quad \text{(by } r(s,a)\leq 1 \text{ and } -\sum_a \pi_{\theta_t}(a|s)\cdot\log\pi_{\theta_t}(a|s)\leq\log A)$$

$$\leq \frac{1 + \tau\log A}{1-\gamma}. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (84)$$

According to the 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,

$$\mathcal{A}_0(s) = \{a : \pi_{\theta_\infty}(a|s) = 0\},$$
$$\mathcal{A}_+(s) = \{a : \pi_{\theta_\infty}(a|s) > 0\}.$$

Note that $\mathcal{A} = \mathcal{A}_0(s) \cup \mathcal{A}_+(s)$ since $\pi_\infty(a|s) \geq 0$, $\forall a \in \mathcal{A}$. 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 $t_0 \geq 1$, such that $\forall t \geq t_0$,

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

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

$$
\begin{aligned}
\frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a_0)} &= \frac{1}{1 - \gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a_0|s) \cdot \tilde{A}^{\pi_{\theta_t}}(s, a_0) \\
&= \frac{1}{1 - \gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a_0|s) \cdot \left[\tilde{Q}^{\pi_{\theta_t}}(s, a_0) - \tau \log \pi_{\theta_t}(a_0|s) - \tilde{V}^{\pi_{\theta_t}}(s) \right] \\
&\geq \frac{1}{1 - \gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a_0|s) \cdot \left[0 - \tau \log \pi_{\theta_t}(a_0|s) - \frac{1 + \tau \log A}{1 - \gamma} \right] \\
&\geq \frac{1}{1 - \gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \pi_{\theta_t}(a_0|s) \cdot \left[0 + \tau \cdot \frac{1 + \tau \log A}{\tau(1 - \gamma)} - \frac{1 + \tau \log A}{1 - \gamma} \right] = 0, \quad (85)
\end{aligned}
$$

where the first inequality is by

$$\tilde{Q}^{\pi_{\theta_t}}(s, a_0) = r(s, a_0) + \gamma \sum_{s'} \mathcal{P}(s'|s, a_0) \tilde{V}^{\pi_{\theta_t}}(s') \geq 0. \qquad \text{(by } r(s, a_0) \geq 0 \text{ and } \tilde{V}^{\pi_{\theta_t}}(s') \geq 0)$$

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

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

we have,

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

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

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

we have,

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

which implies,

$$\sum_{a_+ \in \mathcal{A}_+(s)} \theta_\infty(a_+|s) = \infty. \qquad (86)$$

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

$$
\begin{aligned}
\sum_a \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a)} &= \sum_{a_0 \in \mathcal{A}_0(s)} \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a_0)} + \sum_{a_+ \in \mathcal{A}_+(s)} \frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t(s, a_+)} \\
&= \frac{1}{1 - \gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \sum_a \pi_{\theta_t}(a|s) \cdot \tilde{A}^{\pi_{\theta_t}}(s, a) \\
&= \frac{1}{1 - \gamma} \cdot d_\mu^{\pi_{\theta_t}}(s) \cdot \left[\tilde{V}^{\pi_{\theta_t}}(s) - \tilde{V}^{\pi_{\theta_t}}(s) \right] = 0. \qquad (87)
\end{aligned}
$$

According to Eq. (85), $\forall t \geq t_0$,

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

According to Eq. (87), $\forall t \geq t_0$,

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

which means $\sum_{a_+ \in \mathcal{A}_+(s)} \theta_t(s, a_+)$ will decrease for all large enough $t \geq 1$. This contradicts with Eq. (86), 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 $\Delta(\mathcal{A})$. Furthermore, 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, $\forall s$,

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

We have $d_\mu^{\pi_{\theta_\infty}}(s) \geq (1 - \gamma) \cdot \mu(s) > 0$ for all states s (cf. Eq. (62)). Therefore we have, $\forall s$,

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

According to Lemma 22, $H(\pi_{\theta_\infty}(\cdot|s))$ 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 states s,

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

which is equivalent to

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

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

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

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 $t_0 \geq 1$, such that $\forall t \geq t_0$,

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

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

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

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

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

Proof. According to the soft sub-optimality lemma of Lemma 26,

$$\tilde{V}^{\pi^*_\tau}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) = \frac{1}{1-\gamma} \sum_s \left[d_\rho^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s) \| \pi^*_\tau(\cdot|s)) \right]$$

$$= \frac{1}{1-\gamma} \sum_s \frac{d_\rho^{\pi_{\theta_t}}(s)}{d_\mu^{\pi_{\theta_t}}(s)} \cdot \left[d_\mu^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s) \| \pi^*_\tau(\cdot|s)) \right]$$

$$\leq \frac{1}{(1-\gamma)^2} \sum_s \frac{1}{\mu(s)} \cdot \left[d_\mu^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s) \| \pi^*_\tau(\cdot|s)) \right]$$

$$\leq \frac{1}{(1-\gamma)^2} \cdot \left\| \frac{1}{\mu} \right\|_\infty \sum_s \left[d_\mu^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s) \| \pi^*_\tau(\cdot|s)) \right]$$

$$= \frac{1}{1-\gamma} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot \left[\tilde{V}^{\pi^*_\tau}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu) \right],$$

where the last equation is again by Lemma 26, and the first inequality is according to $d_\mu^{\pi_{\theta_t}}(s) \geq (1-\gamma) \cdot \mu(s)$ (cf. Eq. (62)). According to Lemmas 7 and 14, $V^{\pi_\theta}(\mu)$ is $8/(1-\gamma)^3$-smooth, and $\mathbb{H}(\mu, \pi_\theta)$ is $(4 + 8\log A)/(1-\gamma)^3$-smooth. Therefore, $\tilde{V}^{\pi_\theta}(\mu) = V^{\pi_\theta}(\mu) + \tau \cdot \mathbb{H}(\mu, \pi_\theta)$ is β-smooth with $\beta = (8 + \tau(4 + 8\log A))/(1-\gamma)^3$. Denote $\tilde{\delta}_t = \tilde{V}^{\pi^*_\tau}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu)$. And note $\eta = \frac{(1-\gamma)^3}{8 + \tau(4 + 8\log A)}$. We have,

$$\tilde{\delta}_{t+1} - \tilde{\delta}_t = \tilde{V}^{\pi_{\theta_t}}(\mu) - \tilde{V}^{\pi_{\theta_{t+1}}}(\mu)$$

$$\leq -\frac{(1-\gamma)^3}{16 + \tau(8 + 16\log A)} \cdot \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2^2 \qquad \text{(by Lemma 18)}$$

$$\leq -\frac{(1-\gamma)^3}{16 + \tau(8 + 16\log A)} \cdot \frac{2\tau}{S} \cdot \min_s \mu(s) \cdot \left[\min_{s,a} \pi_{\theta_t}(a|s) \right]^2 \cdot \left\| \frac{d_\mu^{\pi^*_\tau}}{d_\mu^{\pi_{\theta_t}}} \right\|_\infty^{-1} \cdot \left[\tilde{V}^{\pi^*_\tau}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu) \right]$$

$$\text{(by Lemma 15)}$$

$$\leq -\frac{(1-\gamma)^4}{(8/\tau + 4 + 8\log A) \cdot S} \cdot \min_s \mu(s) \cdot \left[\min_{s,a} \pi_{\theta_t}(a|s) \right]^2 \cdot \left\| \frac{d_\mu^{\pi^*_\tau}}{\mu} \right\|_\infty^{-1} \cdot \tilde{\delta}_t \qquad \text{(by } d_\mu^{\pi_{\theta_t}}(s) \geq (1-\gamma) \cdot \mu(s))$$

$$\leq -\frac{(1-\gamma)^4}{(8/\tau + 4 + 8\log A) \cdot S} \cdot \min_s \mu(s) \cdot \left[\inf_{t \geq 1} \min_{s,a} \pi_{\theta_t}(a|s) \right]^2 \cdot \left\| \frac{d_\mu^{\pi^*_\tau}}{\mu} \right\|_\infty^{-1} \cdot \tilde{\delta}_t,$$

According to Lemma 16, $c = \inf_{t \geq 1} \min_{s,a} \pi_{\theta_t}(a|s) > 0$ is independent with t. We have,

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

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

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

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

where the last inequality is according to Eq. (84). Therefore we have the final result,

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

$$\leq \frac{1}{\exp\{C \cdot t\}} \cdot \frac{1 + \tau \log A}{(1-\gamma)^2} \cdot \left\| \frac{1}{\mu} \right\|_\infty,$$

where

$$C = \frac{(1-\gamma)^4}{(8/\tau + 4 + 8\log A) \cdot S} \cdot \min_s \mu(s) \cdot c \cdot \left\| \frac{d_\mu^{\pi_\tau^*}}{\mu} \right\|_\infty^{-1} > 0,$$

is independent with t. $\qquad\square$

A.3. Proofs for Section 5 (Does Entropy Regularization Really Help?)

A.3.1. PROOFS FOR THE BANDIT CASE

Lemma 17 (Reversed Łojasiewicz). Take any $r \in [0,1]^K$. Let $\Delta = r(a^*) - \max_{a \neq a^*} r(a) > 0$ be the reward gap of r. Then,

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

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

$$
\begin{aligned}
(\pi^* - \pi_\theta)^\top r &= \sum_a \pi_\theta(a) \cdot r(a^*) - \sum_a \pi_\theta(a) \cdot r(a) \\
&= \sum_{a \neq a^*} \pi_\theta(a) \cdot r(a^*) - \sum_{a \neq a^*} \pi_\theta(a) \cdot r(a) \\
&= \sum_{a \neq a^*} \pi_\theta(a) \cdot \Delta(a) \\
&\geq \sum_{a \neq a^*} \pi_\theta(a) \cdot \Delta.
\end{aligned}
$$

On the other hand,

$$0 \leq r(a^*) - \pi_\theta^\top r = (\pi^* - \pi_\theta)^\top r = \sum_{a \neq a^*} \pi_\theta(a) \cdot \Delta(a) \leq \sum_{a \neq a^*} \pi_\theta(a) \cdot 1 = \sum_{a \neq a^*} \pi_\theta(a).$$

Therefore the ℓ_2 norm of gradient can be upper bounded as

$$
\begin{aligned}
\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 &= \left(\pi_\theta(a^*)^2 \cdot \left[r(a^*) - \pi_\theta^\top r \right]^2 + \sum_{a \neq a^*} \left[\pi_\theta(a)^2 \cdot (r(a) - \pi_\theta^\top r)^2 \right] \right)^{\frac{1}{2}} \\
&\leq \left(1^2 \cdot \left[\sum_{a \neq a^*} \pi_\theta(a) \right]^2 + \sum_{a \neq a^*} \left[\pi_\theta(a)^2 \cdot 1^2 \right] \right)^{\frac{1}{2}} \\
&\leq \left(\left[\sum_{a \neq a^*} \pi_\theta(a) \right]^2 + \left[\sum_{a \neq a^*} \pi_\theta(a) \right]^2 \right)^{\frac{1}{2}} \qquad\qquad \text{(by } \|x\|_2 \leq \|x\|_1\text{)} \\
&= \sqrt{2} \cdot \sum_{a \neq a^*} \pi_\theta(a).
\end{aligned}
$$

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). Take any $r \in [0, 1]^K$. For large enough $t \geq 1$, using Update 1 with learning rate $\eta_t \in (0, 1]$,

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

Proof. Denote $\delta_t = (\pi^* - \pi_{\theta_t})^\top r > 0$. Let $\theta_{t+1} = \theta_t + \eta_t \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t}$, and $\pi_{\theta_{t+1}} = \mathrm{softmax}(\theta_{t+1})$ be the next policy after one step gradient update. We have,

$$\delta_t - \delta_{t+1} = (\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle + \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \tag{88}$$

$$\leq \frac{5}{4} \cdot \|\theta_{t+1} - \theta_t\|_2^2 + \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \qquad \text{(by Lemma 2)}$$

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

$$\leq \frac{9}{2} \cdot \frac{1}{\Delta^2} \cdot \delta_t^2. \qquad \text{(by } \eta_t \in (0, 1] \text{ and by Lemma 17)}$$

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

$$\delta_{t+1} \geq \delta_t - \frac{9}{2} \cdot \frac{1}{\Delta^2} \cdot \delta_t^2$$

$$> \frac{10}{9} \cdot \delta_{t+1} - \frac{9}{2} \cdot \frac{1}{\Delta^2} \cdot \left(\frac{10}{9} \cdot \delta_{t+1} \right)^2 \qquad \text{(since } f(x) = x - ax^2 \text{ is increasing for all } x < \frac{1}{2a} \text{ and } a > 0)$$

$$= \frac{10}{9} \cdot \delta_{t+1} - \frac{50}{9} \cdot \frac{1}{\Delta^2} \cdot \delta_{t+1}^2, \tag{89}$$

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

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

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

$$\frac{1}{\delta_{T_1+t}} - \frac{1}{\delta_{T_1}} \leq \frac{5}{\Delta^2} \cdot (t - 1) \leq \frac{5}{\Delta^2} \cdot t.$$

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

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

By abusing notation $t := T_1 + t$ and $C \leq \frac{t}{\Delta^2}$, we have

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

for all large enough $t \geq 1$. $\qquad \square$

Theorem 8 (Lower bound). Take any MDP. For large enough $t \geq 1$, using softmax policy gradient Algorithm 1 with $\eta_t \in (0, 1]$,

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

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, and $a^*(s) = \arg\max_a \pi^*(a|s)$ 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 to Lemma 17. According to the reversed Łojasiewicz inequality of Lemma 28,

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

where $\delta_t = V^*(\mu) - V^{\pi_{\theta_t}}(\mu) > 0$. Let $\theta_{t+1} = \theta_t + \eta_t \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}$, and $\pi_{\theta_{t+1}}(\cdot|s) = \text{softmax}(\theta_{t+1}(s,\cdot))$, $\forall s \in \mathcal{S}$ be the next policy after one step gradient update. Using similar calculations as in Eq. (88),

$$
\begin{aligned}
\delta_t - \delta_{t+1} &= V^{\pi_{\theta_{t+1}}}(\mu) - V^{\pi_{\theta_t}}(\mu) - \left\langle \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}, \theta_{t+1} - \theta_t \right\rangle + \left\langle \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}, \theta_{t+1} - \theta_t \right\rangle \\
&\le \frac{4}{(1-\gamma)^3} \cdot \|\theta_{t+1} - \theta_t\|_2^2 + \left\langle \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}, \theta_{t+1} - \theta_t \right\rangle && \text{(by Lemma 7)} \\
&= \left(\frac{4\eta_t^2}{(1-\gamma)^3} + \eta_t \right) \cdot \left\| \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t} \right\|_2^2 && \text{(by } \theta_{t+1} = \theta_t + \eta_t \cdot \frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}) \\
&\le \frac{10}{(1-\gamma)^5} \cdot \frac{1}{(\Delta^*)^2} \cdot \delta_t^2. && \text{(by } \eta_t \in (0,1] \text{ and by Lemma 28)}
\end{aligned}
$$

According to Theorem 4, we have $\delta_t > 0$, $\delta_t \to 0$ as $t \to \infty$. Using similar arguments as in Eq. (89), we can show that for all large enough $t \ge 1$, $\delta_t \le \frac{11}{10} \cdot \delta_{t+1}$. Divide both sides of $\delta_t - \delta_{t+1} \le \frac{10}{(1-\gamma)^5} \cdot \frac{1}{(\Delta^*)^2} \cdot \delta_t^2$ by $\delta_t \cdot \delta_{t+1}$,

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

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

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

for all large enough $t \ge 1$. $\qquad\square$

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

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,

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

Consider the following example, $r = (0.6, 0.4, 0.2)^\top$, $\pi_\theta = (1 - 3\epsilon, 2\epsilon, \epsilon)^\top$ with small number $\epsilon > 0$.

$$(\pi^* - \pi_\theta)^\top r = r(a^*) - \pi_\theta^\top r = 0.6 - (0.6 - 0.8\epsilon) = 0.8 \cdot \epsilon.$$

According to the reversed Łojasiewicz inequality of Lemma 17,

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \le \frac{\sqrt{2}}{\Delta} \cdot (\pi^* - \pi_\theta)^\top r = \frac{\sqrt{2}}{2} \cdot (\pi^* - \pi_\theta)^\top r \le \frac{1.5}{2} \cdot (\pi^* - \pi_\theta)^\top r = 0.6 \cdot \epsilon.$$

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

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \le 0.6 \cdot \epsilon = \frac{1}{4} \cdot 3 \cdot 0.8 \cdot \epsilon < \pi_\theta(a^*) \cdot 3 \cdot 0.8 \cdot \epsilon = C(\theta) \cdot 3 \cdot 0.8 \cdot \epsilon.$$

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

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 < C(\theta) \cdot 3 \cdot 0.8 \cdot \epsilon = C(\theta) \cdot 3 \cdot (0.8 \cdot \epsilon)^\xi \cdot (0.8 \cdot \epsilon)^{1-\xi} < C(\theta) \cdot (0.8 \cdot \epsilon)^{1-\xi} = C(\theta) \cdot \left[(\pi^* - \pi_\theta)^\top r \right]^{1-\xi},$$

where the second inequality is by $(0.8 \cdot \epsilon)^\xi < 1/3$ for small $\epsilon > 0$ since $\xi > 0$. This is a contradiction with the assumption. Therefore the Łojasiewicz degree ξ cannot be larger than 0. $\qquad \square$

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

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

$$
\begin{aligned}
\delta_\theta &= \mathbb{E}_{a \sim \pi_\tau^*} \left[r(a) - \tau \log \pi_\tau^*(a) \right] - \mathbb{E}_{a \sim \pi_\theta} \left[r(a) - \tau \log \pi_\tau^*(a) \right] - \mathbb{E}_{a \sim \pi_\theta} \left[\tau \log \pi_\tau^*(a) - \tau \log \pi_\theta(a) \right] \\
&= \tau \log \sum_a \exp\{r(a)/\tau\} - \tau \log \sum_a \exp\{r(a)/\tau\} + \tau \cdot D_{\mathrm{KL}}(\pi_\theta \| \pi_\tau^*) && (\text{since } \pi_\tau^* = \mathrm{softmax}(r/\tau)) \\
&= \tau \cdot D_{\mathrm{KL}}(\pi_\theta \| \pi_\tau^*) \\
&\leq \frac{\tau}{2} \cdot \left\| \frac{r}{\tau} - \theta - \frac{(r/\tau - \theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty^2 && (\text{by Lemma 27}) \\
&= \frac{1}{2\tau} \cdot \left\| r - \tau\theta - \frac{(r - \tau\theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty^2. && (90)
\end{aligned}
$$

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

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

where the last two equations are by $H(\pi_\theta)\mathbf{1} = \mathbf{0}$ as shown in Lemma 22. Then we have,

$$
\begin{aligned}
\left\| \frac{d\{\pi_\theta^\top (r - \tau \log \pi_\theta)\}}{d\theta} \right\|_2 &= \left\| H(\pi_\theta) \left(r - \tau\theta - \frac{(r - \tau\theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right) \right\|_2 \\
&\geq \min_a \pi_\theta(a) \cdot \left\| r - \tau\theta - \frac{(r - \tau\theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_2 && (\text{by Lemma 23}) \\
&\geq \min_a \pi_\theta(a) \cdot \left\| r - \tau\theta - \frac{(r - \tau\theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty \\
&\geq \min_a \pi_\theta(a) \cdot \sqrt{2\tau} \cdot \sqrt{\delta_\theta} && (\text{by Eq. (90)}) \\
&= \sqrt{2\tau} \cdot \min_a \pi_\theta(a) \cdot \left(\mathbb{E}_{a \sim \pi_\tau^*} \left[r(a) - \tau \log \pi_\tau^*(a) \right] - \mathbb{E}_{a \sim \pi_\theta} \left[r(a) - \tau \log \pi_\theta(a) \right] \right)^{\frac{1}{2}},
\end{aligned}
$$

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 $C(\theta) = \sqrt{2\tau} \cdot \min_a \pi_\theta(a)$. $\qquad \square$

B. Miscellaneous Extra Supporting Results

Lemma 18 (Ascent lemma for smooth function). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a β-smooth function, $\theta \in \mathbb{R}^d$ and $\theta' = \theta + \frac{1}{\beta} \cdot \frac{\partial f(\theta)}{\partial \theta}$. We have,*

$$f(\theta) - f(\theta') \leq -\frac{1}{2\beta} \cdot \left\| \frac{\partial f(\theta)}{\partial \theta} \right\|_2^2.$$

Proof. According to the definition of smoothness, we have,

$$\left| f(\theta') - f(\theta) - \left\langle \frac{\partial f(\theta)}{\partial \theta}, \theta' - \theta \right\rangle \right| \leq \frac{\beta}{2} \cdot \|\theta' - \theta\|_2^2,$$

which implies,

$$
\begin{aligned}
f(\theta) - f(\theta') &\leq - \left\langle \frac{\partial f(\theta)}{\partial \theta}, \theta' - \theta \right\rangle + \frac{\beta}{2} \cdot \|\theta' - \theta\|_2^2 \\
&= -\frac{1}{\beta} \cdot \left\| \frac{\partial f(\theta)}{\partial \theta} \right\|_2^2 + \frac{\beta}{2} \cdot \frac{1}{\beta^2} \cdot \left\| \frac{\partial f(\theta)}{\partial \theta} \right\|_2^2 \qquad (\theta' = \theta + \frac{1}{\beta} \cdot \frac{\partial f(\theta)}{\partial \theta}) \\
&= -\frac{1}{2\beta} \cdot \left\| \frac{\partial f(\theta)}{\partial \theta} \right\|_2^2. \qquad\qquad \square
\end{aligned}
$$

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

$$
\begin{aligned}
V^{\pi'}(\rho) - V^{\pi}(\rho) &= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi'}(s) \sum_a \left(\pi'(a|s) - \pi(a|s) \right) \cdot Q^{\pi}(s, a) \\
&= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi'}(s) \sum_a \pi'(a|s) \cdot A^{\pi}(s, a).
\end{aligned}
$$

Proof. According to the definition of value function,

$$
\begin{aligned}
V^{\pi'}(s) - V^{\pi}(s) &= \sum_a \pi'(a|s) \cdot Q^{\pi'}(s, a) - \sum_a \pi(a|s) \cdot Q^{\pi}(s, a) \\
&= \sum_a \pi'(a|s) \cdot \left(Q^{\pi'}(s, a) - Q^{\pi}(s, a) \right) + \sum_a \left(\pi'(a|s) - \pi(a|s) \right) \cdot Q^{\pi}(s, a) \\
&= \sum_a \left(\pi'(a|s) - \pi(a|s) \right) \cdot Q^{\pi}(s, a) + \gamma \sum_a \pi'(a|s) \sum_{s'} \mathcal{P}(s'|s, a) \cdot \left[V^{\pi'}(s') - V^{\pi}(s') \right] \\
&= \frac{1}{1-\gamma} \sum_{s'} d_s^{\pi'}(s') \sum_{a'} \left(\pi'(a'|s') - \pi(a'|s') \right) \cdot Q^{\pi}(s', a') \\
&= \frac{1}{1-\gamma} \sum_{s'} d_s^{\pi'}(s') \sum_{a'} \pi'(a'|s') \cdot \left(Q^{\pi}(s', a') - V^{\pi}(s') \right) \\
&= \frac{1}{1-\gamma} \sum_{s'} d_s^{\pi'}(s') \sum_{a'} \pi'(a'|s') \cdot A^{\pi}(s', a'). \qquad\qquad \square
\end{aligned}
$$

Lemma 20 (Second performance difference lemma). *For any policies π and π',*

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

Proof. According to the definition of value function,

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

$\square$

Lemma 21 (Value sub-optimality lemma). *For any policy π,*

$$
V^*(\rho) - V^{\pi}(\rho) = \frac{1}{1-\gamma} \sum_s d_\rho^{\pi}(s) \sum_a \left(\pi^*(a|s) - \pi(a|s) \right) \cdot Q^*(s,a).
$$

Proof. According to the second performance difference lemma of Lemma 20, the result immediately holds. $\square$

Lemma 22 (Spectrum of H matrix). *Let $\pi \in \Delta(\mathcal{A})$. Denote $H(\pi) = diag(\pi) - \pi\pi^\top$. Let*

$$
\pi(1) \le \pi(2) \le \cdots \le \pi(K).
$$

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

$$
\lambda_1 \le \lambda_2 \le \cdots \le \lambda_K.
$$

Then we have,

$$
\lambda_1 = 0,
$$
$$
\pi(i-1) \le \lambda_i \le \pi(i), \ i = 2, 3, \ldots, K.
$$

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

$$
\pi(1) - \pi^\top \pi \le \lambda_1 \le \pi(1),
$$
$$
\pi(i-1) \le \lambda_i \le \pi(i), \ i = 2, 3, \ldots, K.
$$

We show $\lambda_1 = 0$. Note

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

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

$$
x^\top H(\pi) x = \mathbb{E}_{\pi}[x \odot x] - \left(\mathbb{E}_{\pi}[x] \right)^2 = \mathrm{Var}_{\pi}[x] \ge 0,
$$

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

Lemma 23. *Let $\pi \in \Delta(\mathcal{A})$. Denote $H(\pi) = diag(\pi) - \pi\pi^\top$. For any vector $x \in \mathbb{R}^K$,*

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

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

$$x = a_1 \cdot \frac{\mathbf{1}}{\sqrt{K}} + a_2 v_2 + \cdots + a_K v_K$$
$$= \frac{x^\top \mathbf{1}}{K} \cdot \mathbf{1} + a_2 v_2 + \cdots + a_K v_K.$$

Since $H(\pi)$ is symmetric, $\left\{ \frac{1}{\sqrt{K}}, v_2, \ldots, v_K \right\}$ are orthonormal. The last equation is because representation is unique, and

$$a_1 = x^\top \frac{\mathbf{1}}{\sqrt{K}} = \frac{x^\top \mathbf{1}}{\sqrt{K}}.$$

Denote

$$x' = x - \frac{x^\top \mathbf{1}}{K} \cdot \mathbf{1} = a_2 v_2 + \cdots + a_K v_K.$$

We have

$$\|x'\|_2^2 = a_2^2 + \cdots + a_K^2.$$

On the other hand,

$$(\mathbf{Id} - H(\pi))x' = a_2(1 - \lambda_2)v_2 + \cdots + a_K(1 - \lambda_K)v_K.$$

Therefore

$$\begin{aligned}
\|(\mathbf{Id} - H(\pi))x'\|_2 &= \left(a_2^2(1 - \lambda_2)^2 + \cdots + a_K^2(1 - \lambda_K)^2 \right)^{\frac{1}{2}} \\
&\leq \left((a_2^2 + \cdots + a_K^2) \cdot (1 - \lambda_2)^2 \right)^{\frac{1}{2}} \\
&= (1 - \lambda_2) \cdot \|x'\|_2 \\
&\leq \left(1 - \min_a \pi(a) \right) \cdot \|x'\|_2,
\end{aligned}$$

where the first inequality is by $0 \leq \pi(1) \leq \lambda_2 \leq \cdots \leq \lambda_K \leq \pi(K) \leq 1$, and the last inequality is according to $\lambda_2 \geq \pi(1) = \min_a \pi(a)$, and both are shown in Lemma 22. Similarly,

$$\begin{aligned}
\|H(\pi)x'\|_2 &= \left(a_2^2 \lambda_2^2 + \cdots + a_K^2 \lambda_K^2 \right)^{\frac{1}{2}} \\
&\geq \left((a_2^2 + \cdots + a_K^2) \cdot \lambda_2^2 \right)^{\frac{1}{2}} \\
&= \lambda_2 \cdot \|x'\|_2 \\
&\geq \min_a \pi(a) \cdot \|x'\|_2. \qquad \square
\end{aligned}$$

Lemma 24. *Let $\pi_\theta = \text{softmax}(\theta)$ and $\pi_{\theta'} = \text{softmax}(\theta')$. Then*

$$\|\pi_\theta - \pi_{\theta'}\|_\infty \leq 2 \cdot \|\theta - \theta'\|_\infty.$$

Proof. See Xiao et al. (2019, Lemma 5), our proof is for completeness. Since $\pi_\theta, \pi_{\theta'} \in [0, 1]^K$,

$$\begin{aligned}
\|\pi_\theta - \pi_{\theta'}\|_\infty &\leq \|\log \pi_\theta - \log \pi_{\theta'}\|_\infty \\
&= \left\| \theta - \theta' - \left(\log \sum_a \exp\{\theta(a)\} - \log \sum_a \exp\{\theta'(a)\} \right) \cdot \mathbf{1} \right\|_\infty \\
&\leq \|\theta - \theta'\|_\infty + \left| \log \sum_a \exp\{\theta(a)\} - \log \sum_a \exp\{\theta'(a)\} \right| \\
&\leq 2 \cdot \|\theta - \theta'\|_\infty,
\end{aligned}$$

where the last inequality is according to Nachum et al. (2017, Lemma 8). $\qquad \square$

Lemma 25 (Soft performance difference lemma). *For any policies π and π',*

$$\tilde{V}^{\pi'}(\rho) - \tilde{V}^{\pi}(\rho) = \frac{1}{1-\gamma} \sum_s d_\rho^\pi(s) \cdot \left[\sum_a (\pi'(a|s) - \pi(a|s)) \cdot \left[\tilde{Q}^{\pi'}(s,a) - \tau \log \pi'(a|s) \right] + \tau \cdot D_{\mathrm{KL}}(\pi(\cdot|s) \| \pi'(\cdot|s)) \right].$$

Proof. According to the definition of soft value function,

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

$$= \sum_a (\pi'(a|s) - \pi(a|s)) \cdot \left[\tilde{Q}^{\pi'}(s,a) - \tau \log \pi'(a|s) \right] + \sum_a \pi(a|s) \cdot \left[\tilde{Q}^{\pi'}(s,a) - \tau \log \pi'(a|s) - \tilde{Q}^{\pi}(s,a) + \tau \log \pi(a|s) \right]$$

$$= \sum_a (\pi'(a|s) - \pi(a|s)) \cdot \left[\tilde{Q}^{\pi'}(s,a) - \tau \log \pi'(a|s) \right] + \tau D_{\mathrm{KL}}(\pi(\cdot|s) \| \pi'(\cdot|s)) + \gamma \sum_a \pi(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \cdot \left[\tilde{V}^{\pi'}(s') - \tilde{V}^{\pi}(s') \right]$$

$$= \frac{1}{1-\gamma} \sum_{s'} d_s^\pi(s') \cdot \left[\sum_{a'} (\pi'(a'|s') - \pi(a'|s')) \cdot \left[\tilde{Q}^{\pi'}(s',a') - \tau \log \pi'(a'|s') \right] + \tau \cdot D_{\mathrm{KL}}(\pi(\cdot|s') \| \pi'(\cdot|s')) \right]. \qquad \square$$

Lemma 26 (Soft sub-optimality lemma). *For any policy π,*

$$\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi}(\rho) = \frac{1}{1-\gamma} \sum_s \left[d_\rho^\pi(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi(\cdot|s) \| \pi_\tau^*(\cdot|s)) \right].$$

Proof. According to [Nachum et al. (2017](#), Theorem 1), $\forall (s,a)$,

$$\tau \log \pi_\tau^*(a|s) = \tilde{Q}^{\pi_\tau^*}(s,a) - \tilde{V}^{\pi_\tau^*}(s). \qquad (91)$$

According to the soft performance difference lemma of Lemma 25,

$$\tilde{V}^{\pi_\tau^*}(s) - \tilde{V}^{\pi}(s) = \frac{1}{1-\gamma} \sum_{s'} d_s^\pi(s') \cdot \left[\sum_{a'} (\pi_\tau^*(a'|s') - \pi(a'|s')) \cdot \left[\tilde{Q}^{\pi_\tau^*}(s',a') - \tau \log \pi_\tau^*(a'|s') \right] + \tau \cdot D_{\mathrm{KL}}(\pi(\cdot|s') \| \pi_\tau^*(\cdot|s')) \right]$$

$$= \frac{1}{1-\gamma} \sum_{s'} d_s^\pi(s') \cdot \left[\sum_{a'} (\pi_\tau^*(a'|s') - \pi(a'|s')) \cdot \tilde{V}^{\pi_\tau^*}(s') + \tau \cdot D_{\mathrm{KL}}(\pi(\cdot|s') \| \pi_\tau^*(\cdot|s')) \right] \quad \text{(by Eq. (91))}$$

$$= \frac{1}{1-\gamma} \sum_{s'} d_s^\pi(s') \cdot \left[(1-1) \cdot \tilde{V}^{\pi_\tau^*}(s') + \tau \cdot D_{\mathrm{KL}}(\pi(\cdot|s') \| \pi_\tau^*(\cdot|s')) \right]$$

$$= \frac{1}{1-\gamma} \sum_{s'} \left[d_s^\pi(s') \cdot \tau \cdot D_{\mathrm{KL}}(\pi(\cdot|s') \| \pi_\tau^*(\cdot|s')) \right]. \qquad \square$$

Lemma 27 (KL-Logit inequality). *Let $\pi_\theta = \mathrm{softmax}(\theta)$ and $\pi_{\theta'} = \mathrm{softmax}(\theta')$. Then for any constant $c \in \mathbb{R}$,*

$$D_{\mathrm{KL}}(\pi_\theta \| \pi_{\theta'}) \le \frac{1}{2} \cdot \| \theta' - \theta - c \cdot \mathbf{1} \|_\infty^2.$$

In particular, let $c = \frac{(\theta' - \theta)^\top \mathbf{1}}{K}$, we have

$$D_{\mathrm{KL}}(\pi_\theta \| \pi_{\theta'}) \le \frac{1}{2} \cdot \left\| \theta' - \theta - \frac{(\theta' - \theta)^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_\infty^2.$$

Proof. According to the ℓ_1 norm strong convexity of negative entropy over probabilistic simplex, i.e., for any policies π, π',

$$\pi'^\top \log \pi' \ge \pi^\top \log \pi + (\pi' - \pi)^\top \log \pi + \frac{1}{2} \cdot \| \pi - \pi' \|_1^2,$$

we have (let $\pi = \pi_\theta$, and $\pi' = \pi_{\theta'}$),

$$
\begin{aligned}
D_{\mathrm{KL}}(\pi_\theta \| \pi_{\theta'}) &= \pi_\theta^\top \log \pi_\theta - \pi_{\theta'}^\top \log \pi_{\theta'} - (\pi_\theta - \pi_{\theta'})^\top \log \pi_{\theta'} \\
&\leq (\pi_\theta - \pi_{\theta'})^\top \log \pi_\theta - \frac{1}{2} \cdot \|\pi_\theta - \pi_{\theta'}\|_1^2 - (\pi_\theta - \pi_{\theta'})^\top \log \pi_{\theta'} \\
&= (\pi_\theta - \pi_{\theta'})^\top (\log \pi_\theta - \log \pi_{\theta'}) - \frac{1}{2} \cdot \|\pi_\theta - \pi_{\theta'}\|_1^2 \\
&= (\pi_\theta - \pi_{\theta'})^\top \left[\theta - \theta' - \left(\log \sum_a \exp\{\theta(a)\} - \log \sum_a \exp\{\theta'(a)\} \right) \cdot \mathbf{1} \right] - \frac{1}{2} \cdot \|\pi_\theta - \pi_{\theta'}\|_1^2 \\
&= (\pi_\theta - \pi_{\theta'})^\top (\theta - \theta') - \frac{1}{2} \cdot \|\pi_\theta - \pi_{\theta'}\|_1^2 \\
&= (\pi_\theta - \pi_{\theta'})^\top (\theta - \theta' - c \cdot \mathbf{1}) - \frac{1}{2} \cdot \|\pi_\theta - \pi_{\theta'}\|_1^2 \qquad \text{(since } (\pi_\theta - \pi_{\theta'})^\top c \cdot \mathbf{1} = 0 \text{ holds } \forall c \in \mathbb{R}) \\
&\leq \|\theta - \theta' - c \cdot \mathbf{1}\|_\infty \cdot \|\pi_\theta - \pi_{\theta'}\|_1 - \frac{1}{2} \cdot \|\pi_\theta - \pi_{\theta'}\|_1^2 \qquad \text{(by Hölder's inequality)} \\
&\leq \frac{1}{2} \cdot \|\theta - \theta' - c \cdot \mathbf{1}\|_\infty^2,
\end{aligned}
$$

where the last inequality is according to $ax - bx^2 \leq \frac{a^2}{4b}, \forall a, b > 0$. $\qquad \square$

Lemma 28 (Reversed Łojasiewicz). *Denote $\Delta^*(s) = Q^*(s, a^*(s)) - \max_{a \neq a^*(s)} Q^*(s, a) > 0$ as the optimal value gap of state s, where $a^*(s)$ is the action that the optimal policy selects under state s, and $\Delta^* = \min_{s \in \mathcal{S}} \Delta^*(s) > 0$ as the optimal value gap of the MDP. Then we have,*

$$
\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 \leq \frac{1}{1 - \gamma} \cdot \frac{\sqrt{2}}{\Delta^*} \cdot [V^*(\mu) - V^{\pi_\theta}(\mu)].
$$

Proof. Denote $\Delta^*(s, a) = Q^*(s, a^*(s)) - Q^*(s, a)$, and $\Delta^*(s) = \min_{a \neq a^*(s)} \Delta^*(s, a)$. We have,

$$
\begin{aligned}
V^*(\mu) - V^{\pi_\theta}(\mu) &= \frac{1}{1 - \gamma} \sum_s d_\mu^{\pi_\theta}(s) \sum_a (\pi^*(a|s) - \pi_\theta(a|s)) \cdot Q^*(s, a) \qquad \text{(by Lemma 21)} \\
&= \frac{1}{1 - \gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \left[\sum_a \pi_\theta(a|s) \cdot Q^*(s, a^*(s)) - \sum_a \pi_\theta(a|s) \cdot Q^*(s, a) \right] \\
&= \frac{1}{1 - \gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \left[\sum_{a \neq a^*(s)} \pi_\theta(a|s) \cdot Q^*(s, a^*(s)) - \sum_{a \neq a^*(s)} \pi_\theta(a|s) \cdot Q^*(s, a) \right] \\
&= \frac{1}{1 - \gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \left[\sum_{a \neq a^*(s)} \pi_\theta(a|s) \cdot \Delta^*(s, a) \right] \\
&\geq \frac{1}{1 - \gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \left[\sum_{a \neq a^*(s)} \pi_\theta(a|s) \right] \cdot \Delta^*(s).
\end{aligned}
$$

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,

$$
\begin{aligned}
|A^{\pi_\theta}(s, a^*(s))| &= \left| Q^{\pi_\theta}(s, a^*(s)) - \sum_a \pi_\theta(a|s) \cdot Q^{\pi_\theta}(s,a) \right| \\
&= \left| \sum_{a \neq a^*(s)} \pi_\theta(a|s) \cdot [Q^{\pi_\theta}(s, a^*(s)) - Q^{\pi_\theta}(s,a)] \right| \\
&\leq \sum_{a \neq a^*(s)} \pi_\theta(a|s) \cdot |Q^{\pi_\theta}(s, a^*(s)) - Q^{\pi_\theta}(s,a)| \qquad \text{(by the triangle inequality)} \\
&\leq \frac{1}{1-\gamma} \sum_{a \neq a^*(s)} \pi_\theta(a|s). \qquad \text{(because } Q^{\pi_\theta}(s,a) \in [0, 1/(1-\gamma)])
\end{aligned}
$$

Therefore the ℓ_2 norm of gradient can be upper bounded as

$$
\begin{aligned}
\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 &= \frac{1}{1-\gamma} \cdot \left[\sum_s d_\mu^{\pi_\theta}(s)^2 \sum_a \pi_\theta(a|s)^2 \cdot A^{\pi_\theta}(s,a)^2 \right]^{\frac{1}{2}} \\
&= \frac{1}{1-\gamma} \cdot \left[\sum_s d_\mu^{\pi_\theta}(s)^2 \cdot \left(\pi_\theta(a^*(s)|s)^2 \cdot A^{\pi_\theta}(s, a^*(s))^2 + \sum_{a \neq a^*(s)} \pi_\theta(a|s)^2 \cdot A^{\pi_\theta}(s,a)^2 \right) \right]^{\frac{1}{2}} \\
&\leq \frac{1}{1-\gamma} \cdot \left[\sum_s d_\mu^{\pi_\theta}(s)^2 \cdot \left(1 \cdot \frac{1}{(1-\gamma)^2} \cdot \left[\sum_{a \neq a^*(s)} \pi_\theta(a|s) \right]^2 + \sum_{a \neq a^*(s)} \pi_\theta(a|s)^2 \cdot \frac{1}{(1-\gamma)^2} \right) \right]^{\frac{1}{2}} \\
&\leq \frac{1}{(1-\gamma)^2} \cdot \left[\sum_s d_\mu^{\pi_\theta}(s)^2 \cdot 2 \cdot \left[\sum_{a \neq a^*(s)} \pi_\theta(a|s) \right]^2 \right]^{\frac{1}{2}} \qquad \text{(by } \|x\|_2 \leq \|x\|_1) \\
&\leq \frac{1}{(1-\gamma)^2} \cdot \sqrt{2} \cdot \sum_s d_\mu^{\pi_\theta}(s) \cdot \left[\sum_{a \neq a^*(s)} \pi_\theta(a|s) \right]. \qquad \text{(by } \|x\|_2 \leq \|x\|_1)
\end{aligned}
$$

Combining the results, we have

$$
\begin{aligned}
\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 &\leq \frac{1}{(1-\gamma)^2} \cdot \sqrt{2} \cdot \sum_s d_\mu^{\pi_\theta}(s) \cdot \left[\sum_{a \neq a^*(s)} \pi_\theta(a|s) \right] \\
&= \frac{1}{1-\gamma} \cdot \frac{\sqrt{2}}{\Delta^*} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \left[\sum_{a \neq a^*(s)} \pi_\theta(a|s) \right] \cdot \Delta^* \\
&\leq \frac{1}{1-\gamma} \cdot \frac{\sqrt{2}}{\Delta^*} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \left[\sum_{a \neq a^*(s)} \pi_\theta(a|s) \right] \cdot \Delta^*(s) \qquad \text{(by } \Delta^* \leq \Delta^*(s) \text{ holds for all } s) \\
&\leq \frac{1}{1-\gamma} \cdot \frac{\sqrt{2}}{\Delta^*} \cdot [V^*(\mu) - V^{\pi_\theta}(\mu)]. \qquad \square
\end{aligned}
$$

C. Sub-optimality Guarantees for Other Entropy-Based RL Methods

Some interesting insight worth mentioning in the proof of Lemma 15 is that the intermediate results provides sub-optimality guarantees for existing entropy regularized RL methods. In particular, Eqs. (80) and (81) provides policy improvement guarantee for Soft Actor-Critic (Haarnoja et al., 2018, SAC), and Eqs. (82) and (83) 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*

$$\pi_{\theta_{t+1}} = \arg\min_{\pi_\theta} D_{\mathrm{KL}} \left(\pi_\theta(\cdot|s) \middle\| \frac{\exp\{Q^{\pi_{\theta_t}}(s,\cdot)\}}{\sum_a \exp\{Q^{\pi_{\theta_t}}(s,a)\}} \right),$$

which is exactly the KL divergence in Eq. (81), with $\bar{\pi}_\theta(\cdot|s)$ defined in Eq. (80). The soft policy improvement inequality of Eq. (81) 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. (25) and (26);*

- *(ii) for any policy π that satisfies the consistency conditions, i.e., if $\forall s, a$,*

$$\pi(a|s) = \exp\left\{ (\tilde{Q}^\pi(s,a) - \tilde{V}^\pi(s))/\tau \right\},$$
$$\tilde{V}^\pi(s) = \tau \log \sum_a \exp\left\{ \tilde{Q}^\pi(s,a)/\tau \right\},$$

 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 15 provides insight, define the following "path inconsistency",

$$r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a)\tilde{V}^\pi(s') - \tilde{V}^\pi(s) - \tau \log \pi(a|s) = \tilde{Q}^\pi(s,a) - \tilde{V}^\pi(s) - \tau \log \pi(a|s), \tag{92}$$

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. (92) can be written in vector form as

$$\tilde{Q}^{\pi_\theta}(s,\cdot) - \tilde{V}^{\pi_\theta}(s) \cdot \mathbf{1} - \tau \log \pi_\theta(\cdot|s) = \tilde{Q}^{\pi_\theta}(s,\cdot) - \tilde{V}^{\pi_\theta}(s) \cdot \mathbf{1} - \tau\theta(s,\cdot) + \tau \log \sum_a \exp\{\theta(s,a)\} \cdot \mathbf{1}.$$

Denote $c_\theta(s) = \frac{\tilde{V}^{\pi_\theta}(s)}{\tau} - \log \sum_a \exp\{\theta(s,a)\}$, and using Lemma 27 in the proof of Lemma 15, in particular, Eq. (82),

$$D_{\mathrm{KL}}(\pi_\theta(\cdot|s)\|\bar{\pi}_\theta(\cdot|s)) \le \frac{1}{2} \cdot \left\| \frac{\tilde{Q}^{\pi_\theta}(s,\cdot)}{\tau} - \theta(s,\cdot) - c_\theta(s) \cdot \mathbf{1} \right\|_\infty^2$$
$$= \frac{1}{2\tau^2} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tilde{V}^{\pi_\theta}(s) \cdot \mathbf{1} - \tau \log \pi_\theta(\cdot|s) \right\|_\infty^2.$$

Using the above results in Eq. (83),

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

$$= \frac{1}{\sqrt{1-\gamma}} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{\frac{1}{2}} \sum_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \max_a \left| r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a)\tilde{V}^{\pi_\theta}(s') - \tau \log \pi_\theta(a|s) - \tilde{V}^{\pi_\theta}(s) \right|, \tag{93}$$

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. (93) 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 to $\delta_t = C'/t$. Subfigure (d) shows $\pi_{\theta_t}(a^*)$ as a function of t.
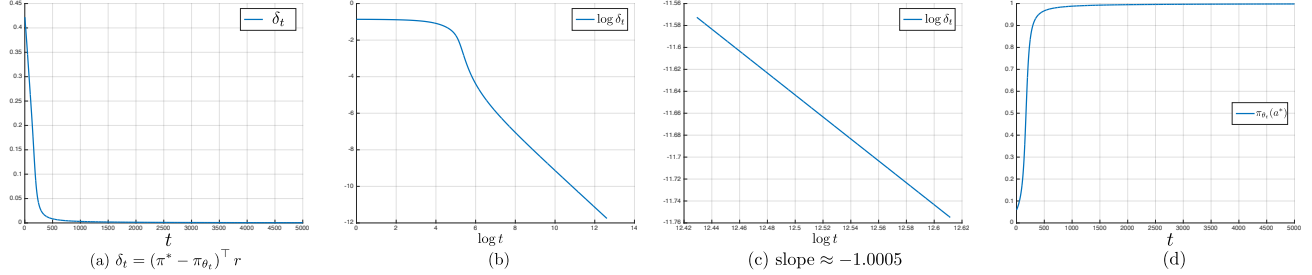

Figure 2. Softmax policy gradient, Update 1.

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 to $\tilde{\delta}_t = C_2'/\exp\{C_1' \cdot t\}$. Subfigure (c) shows ζ_t as defined in Lemma 11 as a function of t, which verifies Lemma 12. 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 13.
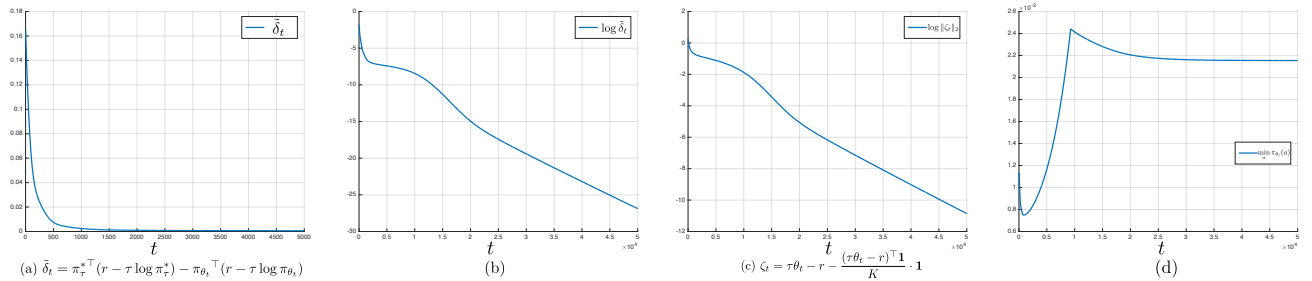

Figure 3. Entropy regularized softmax policy gradient, Update 2.

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.
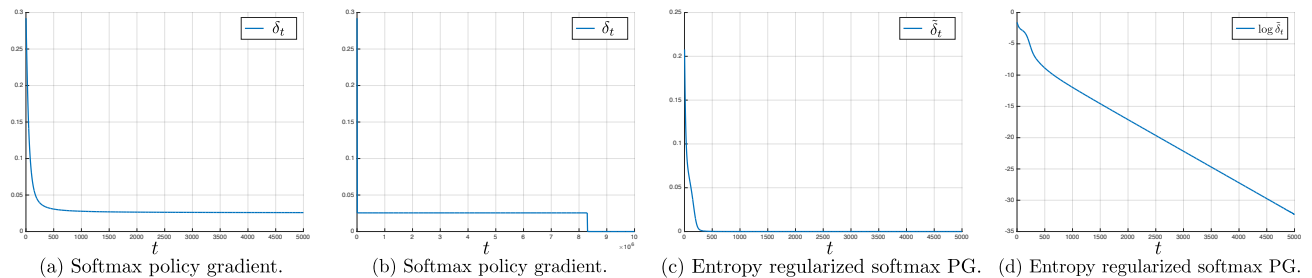

Figure 4. Bad initialization for softmax policy gradient.