Escaping the Gravitational Pull of Softmax

Jincheng Mei[1 4] **Chenjun Xiao**[1 4] **Bo Dai**[4] **Lihong Li**[3*]
Csaba Szepesvári[2 1] **Dale Schuurmans**[4 1]

[1]University of Alberta [2]DeepMind [3]Amazon [4]Google Research, Brain Team

{jmei2,chenjun,daes}@ualberta.ca {bodai,szepi}@google.com llh@amazon.com

Abstract

The softmax is the standard transformation used in machine learning to map real-valued vectors to categorical distributions. Unfortunately, this transform poses serious drawbacks for gradient descent (ascent) optimization. We reveal this difficulty by establishing two negative results: (1) optimizing any expectation with respect to the softmax must exhibit sensitivity to parameter initialization ("softmax gravity well"), and (2) optimizing log-probabilities under the softmax must exhibit slow convergence ("softmax damping"). Both findings are based on an analysis of convergence rates using the Non-uniform Łojasiewicz (NŁ) inequalities. To circumvent these shortcomings we investigate an alternative transformation, the *escort* mapping, that demonstrates better optimization properties. The disadvantages of the softmax and the effectiveness of the escort transformation are further explained using the concept of NŁ coefficient. In addition to proving bounds on convergence rates to firmly establish these results, we also provide experimental evidence for the superiority of the escort transformation.

1 Introduction

The probability transformation plays an essential role in machine learning, used whenever the output of a learned model needs to be mapped to a probability distribution. For example, in reinforcement learning (RL), a probability transformation is used to parameterize policy representations that provide a conditional distribution over a finite set of actions given an input state or observation [18]. In supervised learning (SL), particularly classification, a probability transformation is used to parameterize classifiers that provide a conditional distribution over a finite set of classes given an input observation [7]. Attention models [21] also use probability transformations to provide differentiable forms of memory addressing.

Among the myriad ways one might map vectors to probability distributions, the *softmax* transform is the most common. For $\theta \in \mathbb{R}^K$, the transformation $\pi_\theta = \mathrm{softmax}(\theta)$ is defined by $\pi_\theta(a) = \exp\{\theta(a)\} / \sum_{a'} \exp\{\theta(a')\}$ for all $a \in \{1, ..., K\}$, which ensures $\pi_\theta(a) > 0$ and $\sum_a \pi_\theta(a) = 1$ [4]. The softmax transform can also be extended to continuous output spaces through the concept of a Gibbs function [13], but for concreteness we restrict attention to finite output sets.

Despite the ubiquity of the softmax in machine learning, it is not clear why it should be the default choice of probability transformation. Some alternative transformations have been investigated in the literature [6, 12], but a comprehensive understanding of why one choice might be advantageous over another remains incomplete. It is natural to ask what options might be available and what properties are desirable. In fact, we find that the softmax is a particularly *undesirable* choice from the perspective of gradient descent (ascent) optimization. Moreover, better alternatives are readily available at no

[*] Work done when Lihong Li was with Google. Correspondence to: Jincheng Mei.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

computational overhead. This paper seeks to fill the gap in understanding key properties of probability transformations in general and how they compare to the softmax.

We start by considering *reinforcement learning* and investigate gradient ascent optimization of expected reward using the softmax transform, an algorithm we refer to as softmax policy gradient (SPG) [1, 14]. In this setting, we identify an inherent disadvantage of SPG, the "softmax gravity well (SGW)", whereby gradient ascent trajectories are drawn toward suboptimal corners of the probability simplex and subsequently slowed in their progress toward the optimal vertex. We establish these facts both through theoretical analysis and empirical observation, revealing that the behavior of SPG depends strongly on initialization. Then we propose the use of the *escort* transform as an alternative to softmax for expected reward optimization. We analyze the resulting gradient ascent algorithm, escort policy gradient (EPG), and prove that it enjoys *strictly* better convergence behavior than SPG, significantly mitigating sensitivity to initialization. These findings are also verified experimentally.

Next we consider *supervised learning* and investigate gradient descent optimization of cross entropy loss using the softmax transform, an algorithm we refer to as softmax cross entropy (SCE). Here, even though the optimization landscape at the output layer is convex, we identify a detrimental phenomenon we refer to as "softmax damping". In particular, given deterministic ("one-hot") true label distributions, we show that SCE achieves a slower than linear rate of convergence. Then we propose the use of the escort transform as an alternative to softmax for cross entropy minimization. We analyze the resulting gradient descent algorithm, escort cross entropy (ECE), and show that it is guaranteed to enjoy *strictly* faster convergence than SCE. In particular, a special choice of the escort transform fully eliminates the softmax damping phenomenon, preserving the linear convergence rate for cross entropy minimization.

Finally we propose a unifying concept, the Non-uniform Łojasiewicz (NŁ) coefficient, to explain both the softmax gravity well and softmax damping, even when these might otherwise appear to be disconnected phenomena. We show that by increasing the NŁ coefficient, EPG achieves strictly better initialization dependence than SPG. Moreover, by making the NŁ coefficient non-vanishing, ECE enjoys strictly faster convergence than SCE.

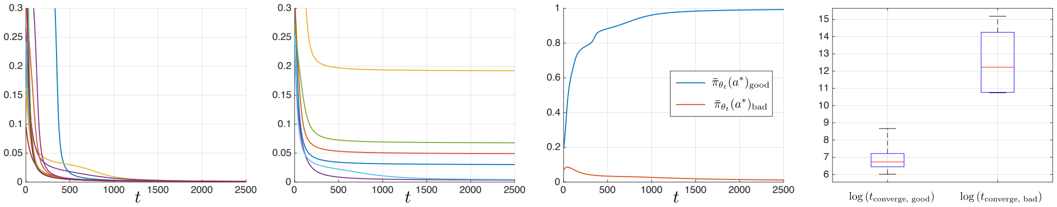
2 Illustrating the Softmax Gravity Wells with Softmax Policy Gradient

We begin by considering the domain of reinforcement learning (RL), where the goal is to learn a policy that maximizes expected return. A core method in RL is *policy gradient* [19], where a parameterized policy is directly optimized to maximize long-term expected reward. It is conventional in this area to represent parametric policies using a softmax transform to produce conditional action distributions, hence policy gradient in practice is almost always the *softmax* policy gradient (SPG).

Despite the fact that SPG has been a dominant RL method for decades, only recently has it been proved to be globally convergent for general MDPs [1]. This result is far from obvious, since the optimization objective is not concave, nevertheless it was shown that SPG converges to the optimal policy under general conditions [1]. More recently, this result was strengthened to establish a $\Theta(1/t)$ bound on the *rate of convergence* [14], with constants that depend on the problem and initialization.

Although these theoretical results are general and impressive, they seem at odds with the behavior of policy gradient methods, which are notoriously difficult to tune in practice [17]. To reconcile theory with empirical observation, we first demonstrate that the "constants" in these results are in fact important, and understanding their role explains much of the real-world performance of SPG.

Illustration To illustrate the point concretely, consider a simple experiment on a single-state Markov Decision Process (MDP) (i.e., a multi-armed bandit) with $K = 6$ actions. In this case, the SPG of a policy π_θ for a given reward vector $r \in [0,1]^K$ reduces to the update $\theta_{t+1}(a) = \theta_t(a) + \eta \cdot \pi_{\theta_t}(a) \cdot \big[r(a) - \pi_{\theta_t}^\top r \big]$, $\forall a \in [K] := \{1, ..., K\}$, and $\pi_{\theta_{t+1}} = \mathrm{softmax}(\theta_{t+1})$. Fig. 1 shows the result of multiple runs using SPG with full gradients. Depending on whether the last iteration satisfies $\pi_{\theta_T}(a^*) \ge 0.99$, we group the 20 runs as "good" and "bad" initializations. As shown in Fig. 1(a) and (b), for good initializations, the sub-optimality $(\pi^* - \pi_{\theta_t})^\top r$ quickly approaches 0, whereas for bad initializations, the iterates get stuck near local optima. Subfigure (c) shows average probability of optimal actions, which shows that the trajectories from bad initializations stay near local optima, since the optimal action probabilities stay close to 0. However, we know from the theory that from any initializations SPG must *eventually* converge to the optimal policy π^*, and that is indeed the case here: Subfigure (d) shows the long-run time to convergence (boxes are 25 to 75th



(a) "good" initializations: 14/20 runs.

(b) "bad" initializations: 6/20 runs.

(c) Average probability of optimal action.

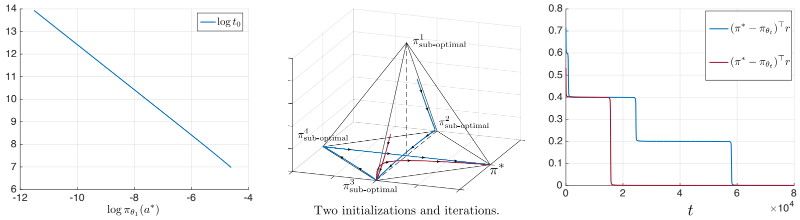
(d) Average convergence time.

Figure 1: SPG behavior on single-state MDPs with $K = 6$ arms, fully parameterized policy (no approximation error), rewards randomly generated (uniform within $[0, 1]$ for each $r(a)$) and policy randomly initialized on each run, 20 runs. Full gradient SPG updates with stepsize $\eta = 0.4$ [14] for $T = 3 \times 10^4$ steps. An initialization is "good" if $\pi_{\theta_T}(a^*) \geq 0.99$ at the last iterate.

percentiles) for good versus bad initializations, where the y-axis is $\log T$ such that $\pi_{\theta_T}(a^*) \geq 0.99$, showing bad runs take *many orders of magnitude* longer.

Although these findings seem not to comport with theory, they can in fact be explained by delving deeper into the detailed nature of the $\Theta(1/t)$ rates proved in [14].

Escape time To control the effect of initialization, consider a specialization of the previous problem where we let $r = (b + \Delta, b, \dots, b)^\top \in [0, 1]^K$ for some b, such that $\Delta > 0$ is the reward gap. For a given initialization, we say that SPG "escapes" at time t_0 if for all $t \geq t_0$ it holds that $(\pi^* - \pi_{\theta_t})^\top r < 0.9 \cdot \Delta$, i.e., after t_0 the sub-optimality stays "small". Fig. 2(a) shows that as the initial probability of the optimal action $\pi_{\theta_1}(a^*)$ decreases, the "escape time" t_0 increases proportionally. In particular, the slope in subfigure (a) approaches -1 as $\pi_{\theta_1}(a^*)$ decreases, indicating that $\log t_0 = -\log \pi_{\theta_1}(a^*) + C$, or equivalently $t_0 = C' / \pi_{\theta_1}(a^*)$. Two trajectories for SPG on a single-state MDP with $K = 5$ is shown in Fig. 2(b) and (c). This example reveals that every suboptimal vertex $i \in \{2, 3, 4\}$ has the potential to attract the iterates, while also slowing progress to render the sub-optimality plateaus in subfigure (c). Therefore, SPG spends some "escape time" around each suboptimal corner.



(a) Escape time t_0.

(b) Softmax gravity wells.

(c) Multiple plateaus.

Figure 2: Dependence on initialization and softmax gravity wells.

We can see as SPG follows a trajectory defined by exact gradients, it effectively encounters "Softmax Gravity Wells (SGWs)" at the vertices (deterministic policies), each of which attracts the trajectory and significantly slows down progress in their vicinity. To see why the attraction to suboptimal vertices is possible, consider the SPG in detail: for a single-state MDP, $\forall a \in [K]$, we have

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

Note that it is possible for an optimal action, say a_1, to be less attractive than a suboptimal action a_2, even when $r(a_1) > r(a_2)$, since it is possible to have both $r(a_1) - \pi_{\theta_t}^\top r > r(a_2) - \pi_{\theta_t}^\top r > 0$ and $\pi_{\theta_t}(a_2) > \pi_{\theta_t}(a_1)$, and yet still have $\pi_{\theta_t}(a_2) \cdot \left[r(a_2) - \pi_{\theta_t}^\top r \right] > \pi_{\theta_t}(a_1) \cdot \left[r(a_1) - \pi_{\theta_t}^\top r \right]$. This configuration causes the probability on the suboptimal action to stay above the optimal action probability, $\pi_{\theta_{t+1}}(a_2) > \pi_{\theta_{t+1}}(a_1)$. Even though the examples and analysis above might seem specific, they provide the foundation for a useful and informative lower bound.

Theorem 1 (Escape time lower bound)**.** *Even in a single-state MDP, for any learning rate $\eta_t \in (0, 1]$, there exists an initialization of the policy π_{θ_1} and a positive constant C, such that SPG with full gradients cannot escape a suboptimal corner before time $t_0 := \frac{C}{\Delta \cdot \pi_{\theta_1}(a^*)}$, i.e., it will hold that*

$$(\pi^* - \pi_{\theta_t})^\top r \geq 0.9 \cdot \Delta, \tag{2}$$

for all $t \leq t_0$, where $\Delta \coloneqq r(a^*) - \max_{a \neq a^*} r(a) > 0$ is the reward gap of $r \in [0,1]^K$.

Theorem 1 shows that for SPG with bounded learning rates (needed for monotonic improvement [1, 14]) the time to escape suboptimal vertices is lower bounded inversely to optimal action probability $\pi_\theta(a^*)$, which is necessarily small near suboptimal vertices, leading to long suboptimal plateaus.

Failure of SPG heuristics Given the insight from Theorem 1, one might wonder if simple heuristics can compensate for the slow progress of SPG, possibly by using large learning rates or normalizing the policy gradient. We show in the appendix that these heuristics unfortunately do not work well even in simple bandit problems.

Existing observations of plateaus SPG plateaus have been observed in the literature. Previous work [14] did observe this effect empirically, but did not take a deeper look into the underlying causes. With function approximation, feature interference has also been considered to be a source of plateaus [16]. In the multi-agent setting, it has been observed that the non-stationary nature of the environment can also cause difficulties for SPG to adapt [8]. However, the analysis in this paper shows that SPG still suffers from plateaus even in the simplest setting (exact gradients, no approximation, stationary environments). In Section 4 we provide additional mathematical insight to explain why the softmax transformation itself is the root cause, which also justifies the name SGW.

3 Escort Transform for Policy Gradient

As explained, a difficulty encountered by SPG comes from the $\pi_\theta(a)$ factor that appears in the gradient, Eq. (1). This creates a dependence on the current policy that potentially discounts the signal from high-reward actions. Unfortunately, the problem is unavoidable if using SPG with bounded learning rates to perform updates (Theorem 1). Therefore, we study the following alternative transform, which we refer to as the "escort transform" [3, 20].

Definition 1 (Escort transform)**.** *Given $\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, define $\pi_\theta = f_p(\theta)$ for $p \geq 1$ by*

$$\pi_\theta(a|s) = \frac{|\theta(s,a)|^p}{\sum_{a' \in \mathcal{A}} |\theta(s,a')|^p}, \quad \text{for all } (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{3}$$

If there is only one state, the escort transform is defined as $\pi_\theta(a) = |\theta(a)|^p / \sum_{a'} |\theta(a')|^p$, $\forall a \in [K]$.

To explain why this alternative transform might help alleviate the problems encountered by the softmax, consider the gradient of expected reward using the escort transform, i.e., the Escort Policy Gradient (EPG), for a single-state MDP, $\forall a \in [K]$ (detailed calculations are shown in the appendix):

$$\begin{aligned} \frac{d\pi_\theta^\top r}{d\theta(a)} &= p \cdot \operatorname{sgn}\{\theta(a)\} \cdot \frac{|\theta(a)|^{p-1}}{\sum_{a'} |\theta(a')|^p} \cdot \left[r(a) - \pi_\theta^\top r \right] & (4) \\ &= \frac{p}{\|\theta\|_p} \cdot \operatorname{sgn}\{\theta(a)\} \cdot \pi_\theta(a)^{1-1/p} \cdot \left[r(a) - \pi_\theta^\top r \right]. & (5) \end{aligned}$$

Note the key difference between SPG and EPG, in which the $\pi_\theta(a)$ term in Eq. (1) now becomes $\pi_\theta(a)^{1-1/p}$ in Eq. (5). Thus, for any $p \geq 1$, we have $1 - 1/p \in [0,1)$, which implies $\pi_\theta(a)^{1-1/p} > \pi_\theta(a)$ since $\pi_\theta(a) \in [0,1]$. This change will have important implications in convergence rate.

Remark 1. *$\pi_\theta(a)^{1-1/p} \to \pi_\theta(a)$ as $p \to \infty$, which suggests that large values of p lead to similar iteration behavior as SPG, whereas small values of p weaken the dependence on $\pi_\theta(a)$. In particular, if $p = 1$ then $\pi_\theta(a)^{1-1/p} = 1$, which entirely eliminates the dependence on current policy π_θ.*

As is the case for the softmax transform, the expected reward objective remains non-concave over parameter θ when using the alternative escort transform.

Proposition 1. *$\theta \mapsto \pi_\theta^\top r$ is a non-concave function over $\mathbb{R}^K$ using the map $\pi_\theta \coloneqq f_p(\theta)$.*

Despite the non-concavity, we manage to obtain surprisingly strong convergence results for EPG, with proofs provided in the appendix. In particular, thanks to what we call non-uniform smoothness and the Non-uniform Łojasiewicz (NŁ) inequality enjoyed by the objective, EPG is shown to enjoy an upper bound on the sub-optimality for single-state MDPs that has a strictly better initialization dependence than SPG.

Theorem 2. *For a single-state MDP, following the escort policy gradient with any initialization such that $|\theta_1(a)| > 0$, $\forall a$, we obtain the following upper bounds on the sub-optimality for all $t \geq 1$:* [2]

$$\textit{(gradient ascent)} \quad \text{for } p \geq 2,\ p = 1, \text{ with } \eta_t = \frac{2 \cdot \|\theta_t\|_p^2}{9 \cdot p^2 \cdot K^{1/p}}, \quad (\pi^* - \pi_{\theta_t})^\top r \leq \frac{9 \cdot K^{1/p}}{c^{2-2/p}} \cdot \frac{1}{t},$$

$$\textit{(gradient flow)} \quad \text{for } p \geq 1, \text{ with } \eta_t = \frac{\|\theta_t\|_p^2}{p^2}, \qquad (\pi^* - \pi_{\theta_t})^\top r \leq \frac{1}{c^{2-2/p} \cdot (t-1) + 1},$$

where $c := \inf_{t \geq 1} \pi_{\theta_t}(a^*) > 0$ depends on the problem and initialization, but is time-independent.

As $p \to \infty$, Theorem 2 implies an $O(1/(c^2\, t))$ convergence rate, recovering the same rate for SPG [14], as expected (Remark 1). For $p < \infty$, EPG achieves the same $O(1/t)$ rate as SPG, but enjoys a strictly better $c^{2-2/p} > c^2$ dependence. In particular, for $p = 1$, there is no dependence on c, which is also consistent with Remark 1. On the other hand, $K^{1/p} \to K$ increases as p decreases, which means it is not always good to choose small p values due to trade-off between K and c.

Similar results can in fact be obtained for EPG in general finite MDPs, denoted as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, \mathcal{P}, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function, $\Delta(\mathcal{X})$ denotes the set of probability distributions over any finite set $\mathcal{X}$, and $\gamma \in [0, 1)$ is the discount factor. Let $V^\pi(\rho) := \mathbb{E}_{s_0 \sim \rho(\cdot), a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)} \sum_{t=0}^\infty \gamma^t r(s_t, a_t)$ denote the expected return (value function) achieved by policy π, where $\rho \in \Delta(\mathcal{S})$ is an initial state distribution. The goal is to maximize the value function, i.e., to find a policy π^* that attains the value $V^*(\rho) := \max_{\pi : \mathcal{S} \to \Delta(\mathcal{A})} V^\pi(\rho)$.

Theorem 3. *Following the escort policy gradient with any initialization such that $|\theta_1(s, a)| > 0$, $\forall (s, a)$, and $\eta_t(s) = \frac{(1-\gamma)^3 \cdot \|\theta_t(s,\cdot)\|_p^2}{10 \cdot p^2 \cdot A^{2/p}}$ to get $\{\theta_t\}_{t \geq 1}$, for all $t \geq 1$, the following sub-optimality upper bounds hold for π_{θ_t},*

$$\text{for } p \geq 2 \text{ and } p = 1, \qquad V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \leq \frac{20 \cdot A^{2/p} \cdot S}{c^{2-2/p} \cdot (1-\gamma)^6 \cdot t} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^2 \cdot \left\| \frac{1}{\mu} \right\|_\infty,$$

where $c := \inf_{s \in \mathcal{S}} \inf_{t \geq 1} \pi_{\theta_t}(a^*(s)|s) > 0$ is problem- and initialization-dependent constant, $A := |\mathcal{A}|$ and $S := |\mathcal{S}|$ are the total number of actions and states, respectively, and $\mu \in \Delta(\mathcal{S})$ is an initial state distribution which provides initial states for the policy gradient method.

Remark 2. *Using $p = 1$ in Theorem 3, the iteration complexity of EPG depends on polynomial functions of S and A, which significantly improves the corresponding results for SPG [14, Theorem 4], where the worst case dependence can be exponential in S and A.*

Finally, as for SPG, adding entropy regularization leads to linear convergence rates for EPG. Note that SPG with entropy regularization enjoys a linear convergence rate $O(1/\exp\{c^2 t\})$ with dependence on $c = \inf_{t \geq 1} \min_{(s,a)} \pi_{\theta_t}(a|s)$ [14]. Our results show that EPG with entropy regularization has strictly better dependence than SPG.

Theorem 4. *For an entropy regularized MDP with finite states and actions, following the escort policy gradient with any initialization such that $|\theta_1(s, a)| > 0$, $\forall (s, a)$, and $\eta_t = (1-\gamma)^3/(10 \cdot p^2 \cdot A^{1/p} + c_\tau)$ to get $\{\theta_t\}_{t \geq 1}$, for all $t \geq 1$, the following sub-optimality upper bounds hold for π_{θ_t}:*

$$\text{for } p \geq 2, \qquad \tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) \leq \frac{\|1/\mu\|_\infty}{\exp\{C_\tau \cdot c'^2 \cdot t\}} \cdot \frac{1 + \tau \log A}{(1-\gamma)^2}, \tag{6}$$

where $c' > c := \inf_{(s,a)} \inf_{t \geq 1} \pi_{\theta_t}(a|s) > 0$, τ is the temperature for entropy regularization, π_τ^* is the softmax optimal policy, and c_τ, C_τ are problem-dependent constants.

Relationship to Mirror Descent (MD) As an additional observation, note that simply removing $\pi_\theta(a)$ in Eq. (1) yields an update $\theta_{t+1} = \theta_t(a) + \eta_t \cdot r(a)$ and $\pi_{\theta_{t+1}} = \mathrm{softmax}(\theta_{t+1})$, which can be combined to yield an update $\pi_{\theta_{t+1}}(a) \propto \pi_{\theta_t}(a) \cdot \exp\{\eta_t\, r(a)\}$ that is equivalent to Mirror Descent

[2] Here, gradient ascent, as expected, refers to $\theta_{t+1} = \theta_t + \eta_t \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t}$ and gradient flow refers to the continuous version when $\frac{d\theta_t}{dt} = \eta_t \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t}$.

(MD) with KL regularization. Given this similarity between SPG, EPG and MD, one might hope that EPG could be reduced to a particular version of MD. However, unlike SPG and MD, the EPG gradient does not specify a conservative vector field and cannot be recovered by MD using any regularization.

Remark 3 (EPG cannot be reduced to MD)**.** *Recall that for a (convex) potential $\Phi : \Delta \to \mathbb{R}$ and its Bregman divergence $D_\Phi : \Delta \times \Delta \to \mathbb{R}$, the MD update is $\pi_{t+1} = \arg\max_{\pi \in \Delta} \pi^\top r - (1/\eta_t)\, D_\Phi(\pi \| \pi_t)$. In particular, using $\Phi(\pi) = \pi^\top \log \pi$ as the potential and $D_\Phi(\pi\|\pi') = D_{\mathrm{KL}}(\pi\|\pi')$ as the divergence one obtains $\pi_{\theta_{t+1}}(a) \propto \pi_{\theta_t}(a) \cdot \exp\{\eta_t\, r(a)\}$. Equivalently, this update can be expressed $\pi_{\theta_{t+1}} = \arg\max_{\pi \in \Delta} \pi^\top \theta_{t+1} - \Phi(\pi)$ where $\theta_{t+1} = \theta_t(a) + \eta_t \cdot r(a)$.*

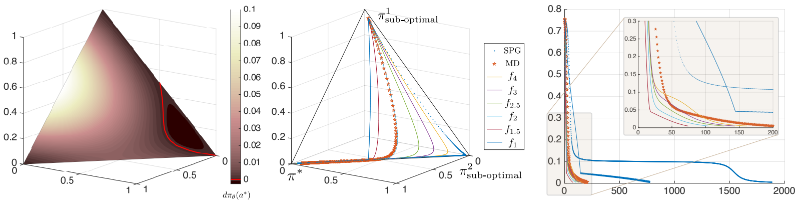
Now suppose EPG is MD, i.e., there is some Φ such that $f_p(\theta_{t+1}) = \arg\max_{\pi \in \Delta} \pi^\top \theta_{t+1} - \Phi(\pi)$. Then we would have to have $f_p(\theta_{t+1}) = \nabla \Phi^*(\theta_{t+1})$ where Φ^* is the Fenchel conjugate of Φ. Taking the derivative w.r.t. θ yields

$$\left(\frac{d\pi_\theta}{d\theta}\right)^\top = \left(\frac{df_p(\theta)}{d\theta}\right)^\top = p \cdot diag(1/\theta)\left(diag(\pi_\theta) - \pi_\theta \pi_\theta^\top\right) \overset{(?)}{=} \frac{d^2 \Phi^*(\theta)}{d\theta^2}. \tag{7}$$

By Schwarz's theorem, $\frac{d^2\Phi^*(\theta)}{d\theta^2}$ must be symmetric, however $diag(1/\theta)\left(diag(\pi_\theta) - \pi_\theta\pi_\theta^\top\right)$ is not symmetric. Therefore, there cannot be a regularizer Φ that makes EPG equivalent to MD.

Remark 3 implies that standard techniques for analyzing mirror descent (e.g., Bregman divergence and convex duality) cannot be directly applied to EPG, necessitating our analysis based on the non-uniform smoothness and NŁ inequalities for Theorems 2 to 4.

Experimental Verification To support these findings and reveal some of the practical implications of EPG versus SPG, we conducted a simple experiment on a single-state MDP with $K = 3$ and $r = (0.2, 0.9, 1.0)^\top$. Fig. 3(a) depicts the $\frac{d\pi_\theta(a^*)}{dt}$ values for SPG, where the dark regions around the corners show areas of slow progress. In particular, the region around the lower-right suboptimal corner exhibits $\frac{d\pi_\theta(a^*)}{dt} < 0$, and $\pi_\theta(a^*)$ will actually *decrease* under SPG updating in this region, prolonging the escape time according to Theorem 1. In short, the dark regions correspond to SGWs for SPG. Subfigure (b) further shows how SPG is attracted toward the suboptimal corner, visually consistent with subfigure (a). By contrast, the solid lines indicate EPG methods with different p values. As noted in Remark 1, smaller p values have better resistance against attraction to SPG gravity wells, while larger p values behave more similarly to SPG. We also observe that MD (with KL regularization) has similar performance to EPG with $p = 2$ in this case. Finally, Subfigure (c) plots the suboptimaliy gap before $(\pi^* - \pi_{\theta_t})^\top r \leq 0.005$ is achieved. It is clear that SPG does get stuck on a suboptimal plateau while EPG methods do not suffer from this disadvantage. We note that EPG curves for $p \geq 2$ behave nicer than $p = 1$ since the escort is differentiable when $p \geq 2$.



(a) Visualization for SGW.

(b) Iteration behaviors of EPG, SPG, and MD.

(c) Iteration numbers for $(\pi^* - \pi_{\theta_t})^\top r \leq 0.005$.

Figure 3: *Empirical visualization for EPG and SPG.*

4 Non-uniform Łojasiewicz Coefficient: An Underlying Explanation

Remark 1 provides an intuition for why EPG has better initialization dependence than SPG. This intuition can be formalized using the notion of Non-uniform Łojasiewicz (NŁ) coefficient, which plays an important role here since both SPG [14] and EPG analyses are based on NŁ inequalities.

Definition 2 (Non-uniform Łojasiewicz (NŁ) coefficient)**.** *A function $f : \mathcal{X} \to \mathbb{R}$ has NŁ coefficient $C(x) > 0$ if it satisfies NŁ inequality with coefficient $C(x)$, i.e., there exists $\xi \in [0, 1]$ such that for all $x \in \mathcal{X}$, $\|\nabla f(x)\|_2 \geq C(x) \cdot |f(x) - f(x^*)|^{1-\xi}$.*

In Definition 2, ξ is called NŁ degree, which impacts the convergence rates of SPG methods [14, Definition 1]. From a result in [14], if $\pi_\theta = \mathrm{softmax}(\theta)$, then $\pi_\theta^\top r$ has NŁ coefficient $\pi_\theta(a^*)$; that is

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 = \left\| \left(\mathrm{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top \right) r \right\|_2 \geq \pi_\theta(a^*) \cdot \left(\pi^* - \pi_\theta \right)^\top r. \tag{8}$$

Moreover, this coefficient is not improvable and it appears in the SPG convergence rate $O(1/(c^2\, t))$, where $c := \inf_{t \geq 1} \pi_{\theta_t}(a^*)$. Now consider EPG. If $\pi_\theta = f_p(\theta)$, then we have

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 = \left\| p \cdot \mathrm{diag}(1/\theta) \left(\mathrm{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top \right) r \right\|_2 \geq p \cdot \frac{\pi_\theta(a^*)}{|\theta(a^*)|} \cdot \left(\pi^* - \pi_\theta \right)^\top r. \tag{9}$$

This implies that $\frac{\pi_\theta(a^*)}{|\theta(a^*)|} = \frac{1}{\|\theta\|_p} \cdot \pi_\theta(a^*)^{1-1/p}$, where $\pi_\theta(a^*)^{1-1/p} > \pi_\theta(a^*)$ provides strictly larger (partial) NŁ coefficient, hence in Theorem 2 EPG obtains a strictly better result than SPG.

The improvement of NŁ coefficient explains a better dependence of EPG on initialization. It is then natural to ask whether the escort transform can also benefit other scenarios, which is answered affirmatively in the next section.

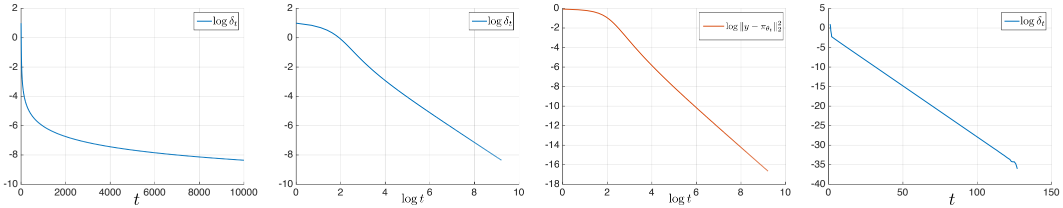
5 Escort Transform for Cross Entropy

We now turn to classification, where the goal is to learn a classifier that minimizes the cross-entropy loss. As in RL, the softmax transform is the default choice for parameterizing a probabilistic classifier. Different from RL where the objective is linear, the objective here involves log probabilities:

$$\min_{\theta: \mathcal{A} \to \mathbb{R}} - \log \pi_\theta(a_y) = \mathcal{H}(y) + \min_{\theta: \mathcal{A} \to \mathbb{R}} D_{\mathrm{KL}}(y \| \pi_\theta), \tag{10}$$

where $\pi_\theta = \mathrm{softmax}(\theta)$, $y \in \{0, 1\}^K$ is a one-hot vector encoding the class label, and a_y is the true label class so that $y(a_y) = 1$. Note that the entropy $\mathcal{H}(y) = 0$ here. The objective in Eq. (10) is smooth and convex in θ, which implies that gradient descent will achieve an $O(1/t)$ rate [15]. Furthermore, for θ that satisfies $\min_a \pi_\theta(a) \geq \pi_{\min}$ with some constant $\pi_{\min} > 0$ (π_θ is bounded away from the simplex boundary), the objective is strongly convex, resulting in an even better, linear rate $O(e^{-t})$. Despite these nice properties, we still find that the softmax transform proves problematic for gradient descent optimization. We refer to this new disadvantage as "softmax damping".

Illustration Consider running gradient descent in a simple experiment where $K = 10$ and y a one-hot vector. Let $\delta_t := -\log \pi_{\theta_t}(a_y)$. If one hopes for a linear convergence rate, i.e., $\delta_t = O(e^{-t})$, then $\log \delta_t = -O(t)$ is expected. But Fig. 4(a) shows a different picture with a flattening slope. Subfigure (b) plots $\log \delta_t$ as a function of $\log t$, which shows a straight line for sufficiently large t with a slope approaching -1. This figure verifies the convergence rate is indeed $\delta_t = O(1/t)$, instead of the linear $O(e^{-t})$ rate. Subfigure (c) shows the ℓ_2 measure $\|y - \pi_{\theta_t}\|_2^2$ also has a similar rate, indicating that this is an inherent optimization phenomenon and is independent of the measurement.



(a) Softmax damping: degenerating rates. (b) Softmax cross entropy. (c) ℓ_2 measure for SCE. (d) Escort cross entropy.

Figure 4: Softmax damping phenomenon and escort cross entropy.

NŁ Coefficient Explanation The NŁ coefficient can be used to explain why this rate degeneration occurs for softmax cross entropy (SCE). Note that for $\pi_\theta = \mathrm{softmax}(\theta)$ we obtain

$$\left\| \frac{d\{D_{\mathrm{KL}}(y \| \pi_\theta)\}}{d\theta} \right\|_2^2 = \|y - \pi_\theta\|_2^2 \geq \min_a \pi_\theta(a) \cdot D_{\mathrm{KL}}(y \| \pi_\theta). \tag{11}$$

Once again the $\min_a \pi_\theta(a)$ term cannot be eliminated for the softmax transform, but here it has a different consequence than before. To see the NŁ coefficient of SCE cannot be improved, consider the example where $y = (0, 1)^\top$ and $\pi = (\epsilon, 1-\epsilon)^\top$, where $\epsilon > 0$ is small. Note that $D_{\mathrm{KL}}(y\|\pi) = -\log(1-\epsilon) \geq \epsilon$ and $\|y - \pi\|_2^2 = 2\,\epsilon^2$, which means for any constant $C > 0$, we have $C \cdot D_{\mathrm{KL}}(y\|\pi) \geq C \cdot \epsilon > 2\,\epsilon^2 = \|y - \pi\|_2^2$. Therefore, for any Łojasiewicz-type inequality, C necessarily depends on $\min_a \pi_\theta(a)$. Now for any convergent sequence π_{θ_t}, i.e., such that $D_{\mathrm{KL}}(y\|\pi_{\theta_t}) \to 0$, we necessarily have $\min_a \pi_{\theta_t}(a) \to 0$, which makes the gradient information insufficient to sustain a linear convergence rate. That is, the fast convergence rate is "damped" in this case. The difference between this phenomenon and the previous "softmax gravity well" is that here the vanishing NŁ coefficients change the rates rather than the constant in the bound on the sub-optimality gap.

Escort Cross Entropy As in Section 3 for policy gradient, we propose to also use the escort transform for cross entropy minimization. A simple calculation for $\pi_\theta = f_p(\theta)$ shows

$$\left\| \frac{d\{D_{\mathrm{KL}}(y\|\pi_\theta)\}}{d\theta} \right\|_2^2 = \|p \cdot \mathrm{diag}(1/\theta)(y - \pi_\theta)\|_2^2 \geq \frac{p^2}{\|\theta\|_p^2} \cdot \min_a \pi_\theta(a)^{1-2/p} \cdot D_{\mathrm{KL}}(y\|\pi_\theta). \quad (12)$$

Note that the term $\min_a \pi_\theta(a)^{1-2/p} > \min_a \pi_\theta(a)$ is strictly better than the softmax cross entropy when $p \geq 2$. In particular, for $p = 2$, the escort cross entropy (ECE) has (partial) NŁ coefficient $\min_a \pi_\theta(a)^{1-2/p} = 1$, which fully eliminates the dependence on the current policy π_θ. This leads to our last main result, which restores the linear convergence rate.

Theorem 5. *Using the escort transform with $p = 2$ and gradient descent on the cross entropy objective with learning rate $\eta_t = \frac{\|\theta_t\|_p^2}{4\cdot(3+c_1^2)}$, we obtain for all $t \geq 1$,*

$$-\log \pi_{\theta_t}(a_y) = D_{\mathrm{KL}}(y\|\pi_{\theta_t}) \leq D_{\mathrm{KL}}(y\|\pi_{\theta_1}) \cdot \exp\left\{ -\frac{(t-1)}{2\cdot(3+c_1^2)} \right\}, \quad (13)$$

where $1/c_1^2 = \pi_{\theta_1}(a_y) \in (0, 1]$ only depends on initialization.

For reference, we run gradient descent on the cross entropy objective in the same experiment above, but with the escort transform. As shown in Fig. 4(d), $\log \delta_t$ now becomes linear in t, or equivalently $-\log \pi_{\theta_t}(a_y) = C \cdot e^{-c\cdot t}$, verifying the theoretical finding of Theorem 5.

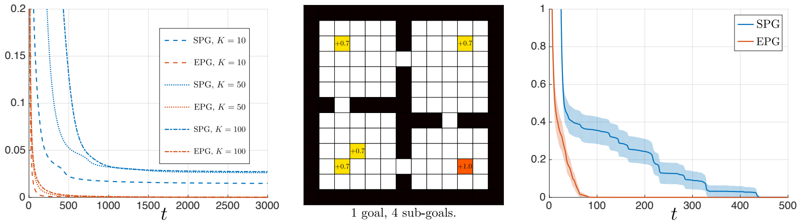
6 Experimental Results

We conduct several experiments to verify the effectiveness of the proposed escort transform in policy gradient and cross entropy minimization.

First, we conduct experiments on one-state MDPs with $K = 10$, 50, and 100, and 20 runs for each K value and for each algorithm. In each run, the reward $r \in [0, 1]^K$ and π_{θ_1} are randomly generated. The total iteration number $T = 5 \times 10^4$. As shown in Fig. 5(a), EPG with $p = 2$ quickly converges to optimal policies consistently across all the K values, significantly outperforming SPG.

Second, we compare the algorithms on Four-room environment for 20 runs. There is one goal with reward 1.0 and 4 sub-goals ("sub-goals" mean goals with lower rewards) with reward 0.7 as shown in Fig. 5(b). At a (sub-)goal state, the agent can step away then step back to receive rewards. The policy is parameterized by one hidden layer ReLU neural network with 64 hidden nodes. In each run, the starting position is randomly generated. We use value iteration to calculate V^* and we calculate the true gradient by closed form. As shown in Fig. 5(c), SPG is easily stuck in plateaus due to the presence of the sub-goals, while EPG with $p = 2$ quickly achieves the optimal goal.

Next, we do experiments on MNIST dataset. For each (x, y), where $x \in \mathbb{R}^{784}$ is image data and $y \in \{0, 1\}^{10}$ is the true label, the training objective is $1 - \pi_\theta(a_y|x)$, where $y(a_y) = 1$. We use policy gradient methods, since the mis-classification probability minimization problem is a special case of expected reward maximization. We use one hidden layer neural network with 512 hidden nodes and ReLU activation to parameterize θ. The dataset is split into training set with 55000, validation set with 5000, and testing set with 10000 data points. As shown in Fig. 6(a) and (b), for both training objective and test error, SPG has plateaus due to SGWs, which is consistent with the observation in [5]. At the same time, EPG with $p = 4$ does not have this disadvantage: it converges quickly and achieves better results than SPG. Experiments for other p values are shown in the appendix.



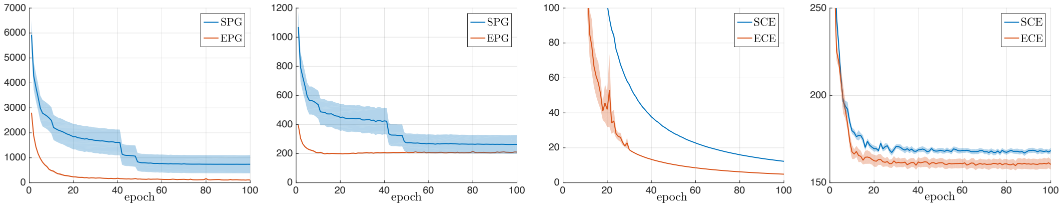
(a) $(\pi^* - \pi_{\theta_t})^\top r$

(b) Four-room environment.

(c) $V^*(\rho) - V^{\pi_t}(\rho)$

Figure 5: *Results on one-state MDPs and Four-room.*

We use mini-batch stochastic gradient descent in this experiment, and the results show that with stochastic gradients and neural network function approximations, (1) SPG still plateaus even when starting from nearly uniform initializations; (2) EPG outperforms SPG in terms of not suffering from plateaus even with estimated gradients.



(a) Training objective.

(b) Test error.

(c) Training objective.

(d) Test error.

Figure 6: *Results on MNIST.*

Finally, for SL, we compare ECE and SCE on MNIST. For each training data (x, y), the training objective is $-\log \pi_\theta(a_y|x)$, where $y(a_y) = 1$. The neural network and dataset settings are the same as above. As shown in Fig. 6(c) and (d), ECE with $p = 2$ is faster than SCE to achieve the same training objective, which benefits generalization, providing smaller test error than SCE.

7 Conclusion and Future Work

We discovered two phenomena that arise from the use of the standard softmax probability transformation in reinforcement learning and supervised learning, and proposed the escort transform to alleviate or eliminate these disadvantages. Our findings of the softmax gravity well and softmax damping phenomena challenge the common practice of using the softmax transformation in machine learning. However, there are other factors to consider when assessing such transformations in machine learning, such as the "temperature" of the softmax, or how the different transforms might impact generalization in the learned models. An important direction for future work is to investigate whether similar phenomena occur in other scenarios where the softmax is commonly utilized, such as attention models and exponential exploration.

Our underlying explanation using the concept of Non-uniform Łojasiewicz (NŁ) coefficient also provides an alternative theory to systematically study probability transforms, which goes beyond the classic convex "matching loss" theory [2, 11] and guarantees better optimization results. We expect further use of the NŁ coefficient to be beneficial in other problems in machine learning.

Broader Impact

This research pursues a fundamental and mostly theoretical goal of understanding how a basic component in machine learning, the softmax transformation, impacts the convergence properties of subsequent optimization methods. The implications of this research are very high level and broad, since we investigate a widely used component. It is difficult to identify specific impacts, since this research does not target any specific application area that would directly impact people or society. If forced to make society level claims, we could attempt to claim that modifying architectures in ways that that improve optimization efficiency would have an effect on the overall energy footprint consumed by machine learning technologies, given how much computation is currently being expended on training softmax classifiers and policies.

Acknowledgments and Disclosure of Funding

Jincheng Mei and Lihong Li would like to thank Christoph Dann for providing feedback on a draft of this manuscript. Csaba Szepesvári gratefully acknowledges funding from the Canada CIFAR AI Chairs Program, Amii and NSERC. Dale Schuurmans acknowledges funding from Amii and NSERC.

References

[1] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in Markov decision processes. arXiv preprint arXiv:1908.00261, 2019.

[2] Peter Auer, Mark Herbster, and Manfred KK Warmuth. Exponentially many local minima for single neurons. In *Advances in neural information processing systems*, pages 316–322, 1996.

[3] Christian Beck and Friedrich Schögl. *Thermodynamics of chaotic systems: an introduction*. Number 4. Cambridge University Press, 1995.

[4] John S Bridle. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In *Advances in neural information processing systems*, 1989.

[5] Minmin Chen, Ramki Gummadi, Chris Harris, and Dale Schuurmans. Surrogate objectives for batch policy optimization in one-step decision making. In *Advances in Neural Information Processing Systems*, pages 8825–8835, 2019.

[6] Alexandre de Brébisson and Pascal Vincent. An exploration of softmax alternatives belonging to the spherical loss family. In *International Conference on Learning Representations*, 2015.

[7] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Springer series in statistics New York, 2001.

[8] Daniel Hennes, Dustin Morrill, Shayegan Omidshafiei, Remi Munos, Julien Perolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, Paavo Parmas, Edgar Duenez-Guzman, et al. Neural replicator dynamics. *arXiv preprint arXiv:1906.00190*, 2019.

[9] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 267–274, 2002.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Jyrki Kivinen and Manfred KK Warmuth. Relative loss bounds for multidimensional regression problems. In *Advances in neural information processing systems*, pages 287–293, 1998.

[12] Anirban Laha, Saneem A. Chemmengath, Priyanka Agrawal, Mitesh M. Khapra, Karthik Sankaranarayanan, and Harish G. Ramaswamy. On controllable sparse alternatives to softmax. In *Advances in Neural Information Processing Systems*, page 6423–6433, 2018.

[13] Yann LeCun, Sumit Chopra, Raia Hadsell, Marc Aurlio Ranzato, and Fu Jie Huang. A tutorial on energy based learning. In *Predicting Structured Data*. MIT Press, 2006.

[14] Jincheng Mei, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. arXiv preprint arXiv:2005.06392, 2020.

[15] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[16] Tom Schaul, Diana Borsa, Joseph Modayil, and Razvan Pascanu. Ray interference: a source of plateaus in deep reinforcement learning. *arXiv preprint arXiv:1904.11455*, 2019.

[17] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897, 2015.

[18] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT Press, second edition, 2018.

[19] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

[20] Constantino Tsallis, Renio S. Mendes, and Anel R. Plastino. The role of constraints within generalized nonextensive statistics. *Physica A: Statistical Mechanics and its Applications*, 261(3-4):534–554, 1998.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

Appendix

The appendix is organized as follows.

A Policy Gradient Method

The policy gradient method we analyze in the main paper is the following Algorithm 1.

Algorithm 1 Policy Gradient Method

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

Note that $\mu \in \Delta(\mathcal{S})$ is an initial state distribution which provides the initial states for policy gradient. Using different parameterizations, we will have different policy gradients $\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta}$ in Algorithm 1.

B Proofs for Section 2 (Softmax Gravity Well)

Lemma 1 (Smoothness [14]). *$\forall r \in [0, 1]^K$, $\theta \mapsto \pi_\theta^\top r$ is $5/2$-smooth, i.e., for all $\pi_\theta := \mathrm{softmax}(\theta)$ and $\pi_{\theta'} := \mathrm{softmax}(\theta')$, we have,*

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

Proof. See the proof in [14, Lemma 2]. □

Theorem 1 (Escape time lower bound). Even in a single-state MDP, for any learning rate $\eta_t \in (0, 1]$, there exists an initialization of the policy π_{θ_1} and a positive constant C, such that SPG with full gradients cannot escape a suboptimal corner before time $t_0 \coloneqq \frac{C}{\Delta \cdot \pi_{\theta_1}(a^*)}$, i.e., it will hold that

$$(\pi^* - \pi_{\theta_t})^\top r \ge 0.9 \cdot \Delta, \tag{14}$$

for all $t \le t_0$, where $\Delta \coloneqq r(a^*) - \max_{a \ne a^*} r(a) > 0$ is the reward gap of $r \in [0,1]^K$.

Proof. Consider the reward vector $r = (b + \Delta, b, \ldots, b)^\top \in [0,1]^K$ for some b, where $\Delta > 0$ is the reward gap. Then we have,

$$\pi_\theta^\top r = \pi_\theta(1) \cdot (b + \Delta) + (1 - \pi_\theta(1)) \cdot b.$$

Note that $a^* = 1$. We have,

$$\begin{aligned} r(a^*) - \pi_\theta^\top r &= b + \Delta - \pi_\theta^\top r \\ &= (1 - \pi_\theta(1)) \cdot \Delta. \end{aligned} \tag{15}$$

And $\forall a \ne 1$, we have,

$$\begin{aligned} r(a) - \pi_\theta^\top r &= b - \pi_\theta^\top r \\ &= -\pi_\theta(1) \cdot \Delta. \end{aligned} \tag{16}$$

Therefore, the ℓ_2 norm of softmax policy gradient can be upper bounded as

$$\begin{aligned} \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 &= \left[\pi_\theta(a^*)^2 \cdot \big(r(a^*) - \pi_\theta^\top r \big)^2 + \sum_{a=2}^K \pi_\theta(a)^2 \cdot \big(r(a) - \pi_\theta^\top r \big)^2 \right]^{\frac{1}{2}} \\ &= \left[\pi_\theta(1)^2 \cdot (1 - \pi_\theta(1))^2 \cdot \Delta^2 + \pi_\theta(1)^2 \cdot \Delta^2 \cdot \sum_{a=2}^K \pi_\theta(a)^2 \right]^{\frac{1}{2}} && \text{(by Eqs. (15) and (16))} \\ &= \pi_\theta(1) \cdot \Delta \cdot \left[(1 - \pi_\theta(1))^2 + \sum_{a=2}^K \pi_\theta(a)^2 \right]^{\frac{1}{2}} \\ &\le \pi_\theta(1) \cdot \Delta \cdot \left[(1 - \pi_\theta(1))^2 + \left(\sum_{a=2}^K \pi_\theta(a) \right)^2 \right]^{\frac{1}{2}} && (\|x\|_2 \le \|x\|_1) \\ &= \sqrt{2} \cdot \pi_\theta(1) \cdot (1 - \pi_\theta(1)) \cdot \Delta. \end{aligned} \tag{17}$$

Let $\theta_{t+1} \leftarrow \theta_t + \eta_t \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t}$, and $\pi_{\theta_{t+1}} = \mathrm{softmax}(\theta_{t+1})$ be the next policy after one step gradient update. Define the following two kinds of iterations:

$$\begin{aligned} t_{\text{good}} &\coloneqq \left\{ t \ge 1 : \pi_{\theta_{t+1}}(1) > \pi_{\theta_t}(1) \right\}, \\ t_{\text{bad}} &\coloneqq \left\{ t \ge 1 : \pi_{\theta_{t+1}}(1) \le \pi_{\theta_t}(1) \right\}. \end{aligned}$$

For all $t \in t_{\text{bad}}$, we have,

$$\frac{1}{\pi_{\theta_t}(1)} - \frac{1}{\pi_{\theta_{t+1}}(1)} = \frac{1}{\pi_{\theta_{t+1}}(1) \cdot \pi_{\theta_t}(1)} \cdot \big(\pi_{\theta_{t+1}}(1) - \pi_{\theta_t}(1) \big) \le 0. \tag{18}$$

For all $t \in t_{\text{good}}$, we have,

$$\begin{aligned}
\pi_{\theta_{t+1}}(1) - \pi_{\theta_t}(1) &= \left[1 - \frac{1}{\Delta} \cdot \left(r(a^*) - \pi_{\theta_{t+1}}^\top r\right)\right] - \left[1 - \frac{1}{\Delta} \cdot \left(r(a^*) - \pi_{\theta_t}^\top r\right)\right] \qquad \text{(by Eq. (15))} \\
&= \frac{1}{\Delta} \cdot \left[(\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \Big\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \Big\rangle + \Big\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \Big\rangle\right] \\
&\le \frac{1}{\Delta} \cdot \left[\frac{5}{4} \cdot \|\theta_{t+1} - \theta_t\|_2^2 + \Big\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \Big\rangle\right] \qquad \text{(by Lemma 1)} \\
&= \frac{1}{\Delta} \cdot \left(\frac{5\eta_t^2}{4} + \eta_t\right) \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \qquad \left(\theta_{t+1} = \theta_t + \eta_t \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right) \\
&\le \frac{1}{\Delta} \cdot \left(\frac{5\eta_t^2}{4} + \eta_t\right) \cdot 2 \cdot \pi_{\theta_t}(1)^2 \cdot (1 - \pi_{\theta_t}(1))^2 \cdot \Delta^2 \qquad \text{(by Eq. (17))} \\
&\le \frac{9}{2} \cdot \pi_{\theta_t}(1)^2 \cdot (1 - \pi_{\theta_t}(1))^2 \cdot \Delta \qquad (\eta_t \in (0, 1]) \\
&\le \frac{9}{2} \cdot \pi_{\theta_t}(1)^2 \cdot \Delta. \qquad (\pi_{\theta_t}(1) \in [0, 1])
\end{aligned} \tag{19}$$

Dividing both sides of Eq. (19) with $\pi_{\theta_{t+1}}(1) \cdot \pi_{\theta_t}(1)$, we have,

$$\begin{aligned}
\frac{1}{\pi_{\theta_t}(1)} - \frac{1}{\pi_{\theta_{t+1}}(1)} &\le \frac{9}{2} \cdot \frac{\pi_{\theta_t}(1)}{\pi_{\theta_{t+1}}(1)} \cdot \Delta \\
&\le \frac{9}{2} \cdot \Delta. \qquad \left(\pi_{\theta_{t+1}}(1) \ge \pi_{\theta_t}(1) > 0\right)
\end{aligned} \tag{20}$$

Therefore, we have,

$$\begin{aligned}
\frac{1}{\pi_{\theta_1}(1)} - \frac{1}{\pi_{\theta_t}(1)} &= \sum_{s=1}^{t-1} \left[\frac{1}{\pi_{\theta_s}(1)} - \frac{1}{\pi_{\theta_{s+1}}(1)}\right] \\
&= \sum_{s=1,\ s \in t_{\text{good}}}^{t-1} \left[\frac{1}{\pi_{\theta_s}(1)} - \frac{1}{\pi_{\theta_{s+1}}(1)}\right] + \sum_{s=1,\ s \in t_{\text{bad}}}^{t-1} \left[\frac{1}{\pi_{\theta_s}(1)} - \frac{1}{\pi_{\theta_{s+1}}(1)}\right] \\
&\le \sum_{s=1,\ s \in t_{\text{good}}}^{t-1} \left[\frac{1}{\pi_{\theta_s}(1)} - \frac{1}{\pi_{\theta_{s+1}}(1)}\right] \qquad \text{(by Eq. (18))} \\
&\le \sum_{s=1,\ s \in t_{\text{good}}}^{t-1} \left[\frac{9}{2} \cdot \Delta\right] \qquad \text{(by Eq. (20))} \\
&\le \frac{9}{2} \cdot \Delta \cdot t.
\end{aligned} \tag{21}$$

Let $\pi_{\theta_1}(1) \le \frac{1}{c}$, for some constant $c \ge 11$. If $t \le \frac{2}{9c} \cdot \frac{1}{\Delta} \cdot \frac{1}{\pi_{\theta_1}(1)}$, then we have,

$$\begin{aligned}
\frac{1}{\pi_{\theta_t}(1)} &\ge \frac{1}{\pi_{\theta_1}(1)} - \frac{9}{2} \cdot \Delta \cdot t \qquad \text{(by Eq. (21))} \\
&\ge \frac{1}{\pi_{\theta_1}(1)} \cdot \left(1 - \frac{1}{c}\right) \\
&\ge c \cdot \left(1 - \frac{1}{c}\right) = c - 1 \ge 10,
\end{aligned}$$

which implies $\pi_{\theta_t}(1) \le \frac{1}{10}$. Therefore, for all $t \le \frac{2}{9c} \cdot \frac{1}{\Delta} \cdot \frac{1}{\pi_{\theta_1}(1)}$, we have,

$$\begin{aligned}
(\pi^* - \pi_{\theta_t})^\top r &= (1 - \pi_{\theta_t}(1)) \cdot \Delta \qquad \text{(by Eq. (15))} \\
&\ge 0.9 \cdot \Delta.
\end{aligned}$$

□

C Proofs for Section 3 (Escort Policy Gradient)

C.1 Escort Policy Gradient Closed Form in Bandits

For completeness, we show the detailed calculations for the escort policy gradient in bandits, i.e., Eqs. (4) and (5), which are duplicated here for convenience,

$$\begin{aligned}\frac{d\pi_\theta^\top r}{d\theta(a)} &= p \cdot \operatorname{sgn}\{\theta(a)\} \cdot \frac{|\theta(a)|^{p-1}}{\sum_{a'} |\theta(a')|^p} \cdot \left[r(a) - \pi_\theta^\top r \right] \\ &= \frac{p}{\|\theta\|_p} \cdot \operatorname{sgn}\{\theta(a)\} \cdot \pi_\theta(a)^{1-1/p} \cdot \left[r(a) - \pi_\theta^\top r \right].\end{aligned}$$

According to the chain rule, we have,

$$\frac{d\pi_\theta^\top r}{d\theta} = \left(\frac{d\pi_\theta}{d\theta} \right)^\top \left(\frac{d\pi_\theta^\top r}{d\pi_\theta} \right) = \left(\frac{d\pi_\theta}{d\theta} \right)^\top r. \tag{22}$$

We calculate the Jacobian of the escort transform $\pi_\theta = f_p(\theta)$. We have, for all $i, j \in [K]$,

$$\begin{aligned}\frac{d\pi_\theta(i)}{d\theta(j)} &= \frac{d}{d\theta(j)} \left\{ \frac{|\theta(i)|^p}{\sum_{a'} |\theta(a')|^p} \right\} \\ &= \frac{\delta_{ij} \cdot p \cdot |\theta(i)|^{p-1} \cdot \operatorname{sgn}\{\theta(i)\} \cdot \left(\sum_{a'} |\theta(a')|^p\right) - |\theta(i)|^p \cdot p \cdot |\theta(j)|^{p-1} \cdot \operatorname{sgn}\{\theta(j)\}}{\left(\sum_{a'} |\theta(a')|^p\right)^2} \\ &= \delta_{ij} \cdot p \cdot \operatorname{sgn}\{\theta(i)\} \cdot \frac{|\theta(i)|^{p-1}}{\sum_{a'} |\theta(a')|^p} - p \cdot \operatorname{sgn}\{\theta(j)\} \cdot \frac{|\theta(j)|^{p-1}}{\sum_{a'} |\theta(a')|^p} \cdot \pi_\theta(i),\end{aligned}$$

where

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases} \tag{23}$$

Then we have the Jacobian as,

$$\left(\frac{d\pi_\theta}{d\theta} \right)^\top = p \cdot \frac{\operatorname{diag}(\operatorname{sgn}\{\theta\}) \operatorname{diag}(|\theta|^{p-1})}{\sum_{a'} |\theta(a')|^p} \left[\mathbf{Id} - \mathbf{1}\pi_\theta^\top \right]. \tag{24}$$

Combing Eqs. (22) and (24), we have,

$$\frac{d\pi_\theta^\top r}{d\theta} = p \cdot \frac{\operatorname{diag}(\operatorname{sgn}\{\theta\}) \operatorname{diag}(|\theta|^{p-1})}{\sum_{a'} |\theta(a')|^p} \left[r - \mathbf{1} \cdot \left(\pi_\theta^\top r \right) \right],$$

which implies Eq. (4). Using $\pi_\theta(a) = \frac{|\theta(a)|^p}{\sum_{a'} |\theta(a')|^p}$, we have, if $\theta(a) \neq 0$,

$$\begin{aligned}\frac{d\pi_\theta^\top r}{d\theta(a)} &= p \cdot \operatorname{sgn}\{\theta(a)\} \cdot \frac{|\theta(a)|^{p-1}}{\sum_{a'} |\theta(a')|^p} \cdot \left[r(a) - \pi_\theta^\top r \right] \\ &= \frac{p}{|\theta(a)|} \cdot \operatorname{sgn}\{\theta(a)\} \cdot \frac{|\theta(a)|^p}{\sum_{a'} |\theta(a')|^p} \cdot \left[r(a) - \pi_\theta^\top r \right] \\ &= \frac{p}{\|\theta\|_p \cdot \pi_\theta(a)^{1/p}} \cdot \operatorname{sgn}\{\theta(a)\} \cdot \pi_\theta(a) \cdot \left[r(a) - \pi_\theta^\top r \right] \\ &= \frac{p}{\|\theta\|_p} \cdot \operatorname{sgn}\{\theta(a)\} \cdot \pi_\theta(a)^{1-1/p} \cdot \left[r(a) - \pi_\theta^\top r \right],\end{aligned}$$

which is Eq. (5). On the other hand, if $\theta(a) = 0$, then $\operatorname{sgn}\{\theta(a)\} = \operatorname{sgn}\{0\} = 0$ makes Eq. (4) trivially equal to Eq. (5).

C.2 One-state MDPs

Proposition 1. $\theta \mapsto \pi_\theta^\top r$ is a non-concave function over $\mathbb{R}^K$ using the map $\pi_\theta := f_p(\theta)$, $p \geq 1$.

Proof. Consider the following example with $K = 3$: $r = (1, 9/10, 1/10)^\top$, $\theta_1 = (1, 1, 1)^\top$, and $\theta_2 = (1, 1, 3)^\top$. Then we have,

$$\begin{aligned}\pi_{\theta_1} &= (1/3, 1/3, 1/3)^\top, \\ \pi_{\theta_2} &= \frac{1}{2 + 3^p} \cdot (1, 1, 3^p)^\top.\end{aligned}$$

Denote $\bar{\theta} \coloneqq \frac{1}{2} \cdot (\theta_1 + \theta_2) = (1, 1, 2)^\top$. We have,

$$\begin{aligned}\frac{1}{2} \cdot \left(\pi_{\theta_1}^\top r + \pi_{\theta_2}^\top r\right) &= \frac{1}{2} \cdot \left(\frac{2}{3} + \frac{19 + 3^p}{10 \cdot (2 + 3^p)}\right) \\ &= \frac{97 + 23 \cdot 3^p}{60 \cdot (2 + 3^p)} \\ &= \frac{194 + 97 \cdot 2^p + 46 \cdot 3^p + 23 \cdot 2^p \cdot 3^p}{60 \cdot (2 + 3^p) \cdot (2 + 2^p)} \\ &= \frac{19 + 2^p}{10 \cdot (2 + 2^p)} + \frac{17 \cdot (-2 + 5 \cdot 2^p - 4 \cdot 3^p + 2^p \cdot 3^p)}{60 \cdot (2 + 3^p) \cdot (2 + 2^p)} \\ &\ge \frac{19 + 2^p}{10 \cdot (2 + 2^p)} \qquad \text{(see below)} \\ &= \pi_{\bar{\theta}}^\top r, \end{aligned} \tag{25}$$

where the last inequality is because of the function $g(x) : x \mapsto -2 + 5 \cdot 2^x - 4 \cdot 3^x + 2^x \cdot 3^x$ is non-negative for all $x \ge 1$. In fact, for any $x \ge 2$, we have,

$$\begin{aligned}g(x) &\coloneqq -2 + 5 \cdot 2^x - 4 \cdot 3^x + 2^x \cdot 3^x \\ &\ge -2 + 5 \cdot 2^x - 4 \cdot 3^x + 4 \cdot 3^x \qquad (x \ge 2) \\ &= -2 + 5 \cdot 2^x \ge 0.\end{aligned}$$

On the other hand, for any $x \in [1, 2)$, we have,

$$\begin{aligned}g(x) &\coloneqq -2 + 5 \cdot 2^x - 4 \cdot 3^x + 2^x \cdot 3^x \\ &\ge -2 + 5 \cdot 2^x - 4 \cdot 3^x + 2 \cdot 3^x \qquad (x \ge 1) \\ &= -2 + 5 \cdot 2^x - 2 \cdot 3^x \ge 0,\end{aligned}$$

which is easy to verify. According to Eq. (25), $\mathbb{E}_{a \sim \pi_\theta(\cdot)}\left[r(a)\right]$ is a non-concave function of θ. □

Lemma 2 (Non-uniform smoothness)**.** *Suppose $r \in [0, 1]^K$. Let $\pi_\theta \coloneqq f_p(\theta)$, and $\pi_{\theta'} \coloneqq f_p(\theta')$. Denote $\theta_\xi \coloneqq \theta + \xi \cdot (\theta' - \theta)$ with some $\xi \in [0, 1]$. Then, we have,*

- *for $p \ge 2$, $\pi_\theta^\top r$ is $\frac{3 \cdot p^2 \cdot K^{1/p}}{\|\theta_\xi\|_p^2}$-smooth, i.e.,*

$$\left| (\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d \pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle \right| \le \frac{3 \cdot p^2 \cdot K^{1/p}}{2 \cdot \|\theta_\xi\|_p^2} \cdot \|\theta' - \theta\|_2^2.$$

- *for $p = 1$, $\pi_\theta^\top r$ is $\frac{2 \cdot K}{\|\theta_\xi\|_1^2}$-smooth, i.e.,*

$$\left| (\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d \pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle \right| \le \frac{K}{\|\theta_\xi\|_1^2} \cdot \|\theta' - \theta\|_2^2.$$

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

$$\begin{aligned}S(r, \theta) &= \frac{d}{d\theta} \left\{ \frac{d \pi_\theta^\top r}{d\theta} \right\} \\ &= p \cdot \frac{d}{d\theta} \left\{ \text{diag}\Big(\frac{\pi_\theta}{\theta}\Big) (r - \pi_\theta^\top r \cdot \mathbf{1}) \right\}.\end{aligned}$$

Note that $S(r,\theta) \in \mathbb{R}^{K\times K}$, whose element at position $(i,j) \in [K]^2$ is

$$\begin{aligned}
S_{i,j} &= p \cdot \frac{d\{\frac{\pi_\theta(i)}{\theta(i)} \cdot (r(i) - \pi_\theta^\top r)\}}{d\theta(j)} \\
&= p \cdot \frac{d\{\frac{\pi_\theta(i)}{\theta(i)}\}}{d\theta(j)} \cdot (r(i) - \pi_\theta^\top r) + p \cdot \frac{\pi_\theta(i)}{\theta(i)} \cdot \frac{d\{r(i) - \pi_\theta^\top r\}}{d\theta(j)} \\
&= p \cdot \frac{\frac{p}{\theta(j)} \cdot [\delta_{ij} \cdot \pi_\theta(i) - \pi_\theta(i) \cdot \pi_\theta(j)] \cdot \theta(i) - \pi_\theta(i) \cdot \delta_{ij}}{\theta(i)^2} \cdot (r(i) - \pi_\theta^\top r) \\
&\quad - \frac{\pi_\theta(i)}{\theta(i)} \cdot p^2 \cdot \frac{\pi_\theta(j)}{\theta(j)} \cdot (r(j) - \pi_\theta^\top r) \\
&= p \cdot (p-1) \cdot \delta_{ij} \cdot \frac{\pi_\theta(i)}{\theta(i)^2} \cdot (r(i) - \pi_\theta^\top r) \\
&\quad - p^2 \cdot \frac{\pi_\theta(i)}{\theta(i)} \cdot \frac{\pi_\theta(j)}{\theta(j)} \cdot (r(i) - \pi_\theta^\top r) - p^2 \cdot \frac{\pi_\theta(i)}{\theta(i)} \cdot \frac{\pi_\theta(j)}{\theta(j)} \cdot (r(j) - \pi_\theta^\top r),
\end{aligned}$$

where δ_{ij} is defined in Eq. (23). We calculate the spectral radius of $S(r,\theta)$. For any nonzero $y \in \mathbb{R}^K$,

$$\begin{aligned}
\left|y^\top S(r,\theta) y\right| &= \left| \sum_{i=1}^{K} \sum_{j=1}^{K} S_{i,j} y(i) y(j) \right| \\
&= \Bigg| p \cdot (p-1) \sum_i \frac{\pi_\theta(i)}{\theta(i)^2} \cdot (r(i) - \pi_\theta^\top r) \cdot y(i)^2 \\
&\qquad - 2 \cdot p^2 \sum_i \frac{\pi_\theta(i)}{\theta(i)} \cdot y(i) \sum_j \frac{\pi_\theta(j)}{\theta(j)} \cdot (r(j) - \pi_\theta^\top r) \cdot y(j) \Bigg| \\
&\le p \cdot (p-1) \cdot \left| \sum_i \frac{\pi_\theta(i)}{\theta(i)^2} \cdot (r(i) - \pi_\theta^\top r) \cdot y(i)^2 \right| \\
&\qquad + 2 \cdot p^2 \cdot \left| \sum_i \frac{\pi_\theta(i)}{\theta(i)} \cdot y(i) \sum_j \frac{\pi_\theta(j)}{\theta(j)} \cdot (r(j) - \pi_\theta^\top r) \cdot y(j) \right|, \qquad (26)
\end{aligned}$$

where the last inequality is by triangle inequality.

First part. For $p \ge 2$, the first term in Eq. (26) is upper bounded as,

$$\begin{aligned}
&\left| \sum_i \frac{\pi_\theta(i)}{\theta(i)^2} \cdot (r(i) - \pi_\theta^\top r) \cdot y(i)^2 \right| \le \sum_i \frac{\pi_\theta(i)}{\theta(i)^2} \cdot \left| r(i) - \pi_\theta^\top r \right| \cdot y(i)^2 \qquad \text{(by triangle inequality)} \\
&\quad \le \left(\max_a \frac{\pi_\theta(a)}{\theta(a)^2} \cdot |r(a) - \pi_\theta^\top r| \right) \cdot \|y\|_2^2 \qquad \text{(by Hölder's inequality)} \\
&\quad \le \left(\max_a \frac{\pi_\theta(a)}{\theta(a)^2} \right) \cdot \|y\|_2^2 \qquad (r(a) \in [0,1], \ \forall a) \\
&\quad = \frac{1}{\|\theta\|_p^2} \cdot \left(\max_a \pi_\theta(a)^{1-2/p} \right) \cdot \|y\|_2^2 \\
&\quad \le \frac{1}{\|\theta\|_p^2} \cdot \|y\|_2^2 . \qquad (p \ge 2) \qquad (27)
\end{aligned}$$

The last term in Eq. (26) is upper bounded as,

$$\begin{aligned}
&\left| \sum_i \frac{\pi_\theta(i)}{\theta(i)} \cdot y(i) \sum_j \frac{\pi_\theta(j)}{\theta(j)} \cdot (r(j) - \pi_\theta^\top r) \cdot y(j) \right| \\
&\quad \le \left\| \frac{\pi_\theta}{\theta} \right\|_2 \cdot \|y\|_2 \cdot \left\| \operatorname{diag}\left(\frac{\pi_\theta}{\theta} \right) (r - \pi_\theta^\top r \cdot \mathbf{1}) \right\|_2 \cdot \|y\|_2 \qquad \text{(by Cauchy-Schwarz)} \\
&\quad = \left\| \frac{\pi_\theta}{\theta} \right\|_2 \cdot \left[\sum_a \left(\frac{\pi_\theta(a)}{\theta(a)} \right)^2 \cdot (r(a) - \pi_\theta^\top r)^2 \right]^{\frac{1}{2}} \cdot \|y\|_2^2 \\
&\quad \le \left\| \frac{\pi_\theta}{\theta} \right\|_2 \cdot \left[\sum_a \left(\frac{\pi_\theta(a)}{\theta(a)} \right)^2 \right]^{\frac{1}{2}} \cdot \|y\|_2^2 \qquad (r(a) \in [0,1], \ \forall a) \\
&\quad = \sum_a \left(\frac{\pi_\theta(a)}{\theta(a)} \right)^2 \cdot \|y\|_2^2 \\
&\quad = \frac{1}{\|\theta\|_p^2} \sum_a \left(\pi_\theta(a)^{1-1/p} \right)^2 \cdot \|y\|_2^2 \\
&\quad \le \frac{1}{\|\theta\|_p^2} \cdot \left(\sum_a \pi_\theta(a)^{1-1/p} \right) \cdot \|y\|_2^2 . \qquad \left(\pi_\theta(a)^{1-1/p} \in [0,1] \right)
\end{aligned} \tag{28}$$

The intermediate term is then upper bounded as,

$$\begin{aligned}
\sum_a \pi_\theta(a)^{1-1/p} &= K^{1/p} \cdot \frac{1}{K} \sum_a (K \cdot \pi_\theta(a))^{1-1/p} \\
&\le K^{1/p} \cdot \left(\sum_a \frac{K \cdot \pi_\theta(a)}{K} \right)^{1-1/p} \qquad \text{(by Jensen's inequality)} \\
&= K^{1/p}.
\end{aligned} \tag{29}$$

Combining Eqs. (26) to (29), we have

$$\begin{aligned}
\left| y^\top S(r,\theta) y \right| &\le p \cdot (p-1) \cdot \frac{1}{\|\theta\|_p^2} \cdot \|y\|_2^2 + 2 \cdot p^2 \cdot \frac{1}{\|\theta\|_p^2} \cdot K^{1/p} \cdot \|y\|_2^2 \\
&\le \frac{3 \cdot p^2 \cdot K^{1/p}}{\|\theta\|_p^2} \cdot \|y\|_2^2 . \qquad \left(K^{1/p} \ge 1 \right)
\end{aligned} \tag{30}$$

According to Taylor's theorem, we have, for $p \ge 2$,

$$\begin{aligned}
\left| (\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d\pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle \right| &= \frac{1}{2} \cdot \left| (\theta' - \theta)^\top S(r, \theta_\xi) (\theta' - \theta) \right| \\
&\le \frac{3 \cdot p^2 \cdot K^{1/p}}{2 \cdot \|\theta_\xi\|_p^2} \cdot \|\theta' - \theta\|_2^2 . \qquad \text{(by Eq. (30))}
\end{aligned}$$

Second part. For $p = 1$, according to Eq. (26), Eqs. (28) and (29), we have,

$$\left| y^\top S(r,\theta) y \right| \le 2 \cdot p^2 \cdot \frac{1}{\|\theta\|_p^2} \cdot K^{1/p} \cdot \|y\|_2^2 = \frac{2 \cdot K}{\|\theta\|_1^2} \cdot \|y\|_2^2 . \tag{31}$$

According to Taylor's theorem, we have, for $p = 1$,

$$\left| (\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d\pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle \right| \le \frac{K}{\|\theta_\xi\|_1^2} \cdot \|\theta' - \theta\|_2^2 . \qquad \square$$

Lemma 3 (Non-uniform Łojasiewicz). *Let $\pi_\theta = f_p(\theta)$. For any $p > 0$, we have,*

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \ge \frac{p}{\|\theta\|_p} \cdot \pi_\theta(a^*)^{1-1/p} \cdot (\pi^* - \pi_\theta)^\top r,$$

where $\pi^* = \arg\max_{\pi \in \Delta} \pi^\top r$ is the optimal policy.

Proof. The result follows from calculating the gradient norm,

$$
\begin{aligned}
\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 &= \left[\sum_{a=1}^K \left(p \cdot \frac{\pi_\theta(a)}{\theta(a)} \cdot (r(a) - \pi_\theta^\top r) \right)^2 \right]^{\frac{1}{2}} \\
&\ge \left| p \cdot \frac{\pi_\theta(a^*)}{\theta(a^*)} \cdot (r(a^*) - \pi_\theta^\top r) \right| \\
&= p \cdot \frac{\pi_\theta(a^*)}{|\theta(a^*)|} \cdot (\pi^* - \pi_\theta)^\top r \\
&= \frac{p}{\|\theta\|_p} \cdot \pi_\theta(a^*)^{1-1/p} \cdot (\pi^* - \pi_\theta)^\top r. \qquad \left(\pi_\theta(a) = \frac{|\theta(a)|^p}{\sum_{a'} |\theta(a')|^p} \right) \qquad \square
\end{aligned}
$$

Theorem 2. For a single-state MDP, following the escort policy gradient with any initialization such that $|\theta_1(a)| > 0$, $\forall a$, we obtain the following upper bounds on the sub-optimality for all $t \ge 1$,

- (gradient ascent) for $p \ge 2$, with $\eta_t = \frac{2 \cdot \|\theta_t\|_p^2}{9 \cdot p^2 \cdot K^{1/p}}$, we have,

$$
(\pi^* - \pi_{\theta_t})^\top r \le \frac{9 \cdot K^{1/p}}{c^{2-2/p}} \cdot \frac{1}{t};
$$

- (gradient ascent) for $p = 1$, with $\eta_t = \frac{2 \cdot \|\theta_t\|_1^2}{9 \cdot K}$, we have,

$$
(\pi^* - \pi_{\theta_t})^\top r \le \frac{9 \cdot K}{t};
$$

- (gradient flow) for $p \ge 1$, with $\eta_t = \frac{\|\theta_t\|_p^2}{p^2}$, we have,

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

where $c \coloneqq \inf_{t \ge 1} \pi_{\theta_t}(a^*) > 0$ depends on the problem and initialization, but is time-independent.

Proof. **First part.** For $p \ge 2$, according to Lemma 2,

$$
\left| (\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| \le \frac{3 \cdot p^2 \cdot K^{1/p}}{2 \cdot \|\theta_{t,\xi}\|_p^2} \cdot \|\theta_{t+1} - \theta_t\|_2^2, \tag{32}
$$

where

$$
\theta_{t,\xi} \coloneqq \theta_t + \xi \cdot (\theta_{t+1} - \theta_t) = \theta_t + \xi \cdot \eta_t \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t},
$$

for some $\xi \in [0, 1]$. The ℓ_p gradient norm can be upper bounded as,

$$
\begin{aligned}
\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_p &= \left[\sum_{a=1}^K \left| p \cdot \frac{\pi_\theta(a)}{\theta(a)} \cdot (r(a) - \pi_\theta^\top r) \right|^p \right]^{\frac{1}{p}} \\
&= p \cdot \left[\sum_{a=1}^K \left(\frac{\pi_\theta(a)}{|\theta(a)|} \cdot |r(a) - \pi_\theta^\top r| \right)^p \right]^{\frac{1}{p}} \\
&= \frac{p}{\|\theta\|_p} \cdot \left[\sum_{a=1}^K \left(\pi_\theta(a)^{1-1/p} \cdot |r(a) - \pi_\theta^\top r| \right)^p \right]^{\frac{1}{p}} \\
&\le \frac{p}{\|\theta\|_p} \cdot \left[\sum_{a=1}^K (1 \cdot 1)^p \right]^{\frac{1}{p}} = \frac{p \cdot K^{1/p}}{\|\theta\|_p}.
\end{aligned} \tag{33}
$$

According to the triangle inequality, we have,

$$
\begin{aligned}
\|\theta_{t,\xi}\|_p &\geq \|\theta_t\|_p - \xi \cdot \eta_t \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_p \\
&\geq \|\theta_t\|_p - \xi \cdot \eta_t \cdot \frac{p \cdot K^{1/p}}{\|\theta_t\|_p}. \qquad \text{(by Eq. (33))} \\
&= \|\theta_t\|_p \cdot \left(1 - \xi \cdot \frac{2}{9 \cdot p} \right) \qquad \left(\eta_t = \frac{2 \cdot \|\theta_t\|_p^2}{9 \cdot p^2 \cdot K^{1/p}} \right) \\
&\geq \|\theta_t\|_p \cdot \left(1 - \frac{2}{9 \cdot p} \right) \qquad (\xi \in [0,1]) \\
&= \|\theta_t\|_p \cdot \left[\left(1 - \frac{2}{\sqrt{6}} \right) \cdot \left(1 - \frac{2 \cdot (3 + \sqrt{6})}{9 \cdot p} \right) + \frac{2}{\sqrt{6}} \right] \\
&\geq \frac{2}{\sqrt{6}} \cdot \|\theta_t\|_p. \qquad (p \geq 2) \qquad (34)
\end{aligned}
$$

Combining Eqs. (32) and (34), we have,

$$
\begin{aligned}
\left| (\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \Big\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \Big\rangle \right| &\leq \frac{3 \cdot p^2 \cdot K^{1/p}}{2 \cdot \|\theta_{t,\xi}\|_p^2} \cdot \|\theta_{t+1} - \theta_t\|_2^2 \\
&\leq \frac{9 \cdot p^2 \cdot K^{1/p}}{4 \cdot \|\theta_t\|_p^2} \cdot \|\theta_{t+1} - \theta_t\|_2^2,
\end{aligned}
$$

which implies,

$$
\begin{aligned}
\pi_{\theta_t}^\top r - \pi_{\theta_{t+1}}^\top r &\leq -\Big\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \Big\rangle + \frac{9 \cdot p^2 \cdot K^{1/p}}{4 \cdot \|\theta_t\|_p^2} \cdot \|\theta_{t+1} - \theta_t\|_2^2 \\
&= -\eta_t \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 + \frac{9 \cdot p^2 \cdot K^{1/p}}{4 \cdot \|\theta_t\|_p^2} \cdot \eta_t^2 \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \qquad \left(\theta_{t+1} = \theta_t + \eta_t \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right) \\
&= -\frac{\|\theta_t\|_p^2}{9 \cdot p^2 \cdot K^{1/p}} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \qquad \left(\eta_t = \frac{2 \cdot \|\theta_t\|_p^2}{9 \cdot p^2 \cdot K^{1/p}} \right) \\
&\leq -\frac{\cancel{\|\theta_t\|_p^2}}{9 \cdot \cancel{p^2} \cdot K^{1/p}} \cdot \left[\frac{\cancel{p}}{\cancel{\|\theta_t\|_p}} \cdot \pi_{\theta_t}(a^*)^{1-1/p} \cdot (\pi^* - \pi_{\theta_t})^\top r \right]^2 \qquad \text{(by Lemma 3)} \\
&= -\frac{\pi_{\theta_t}(a^*)^{2-2/p}}{9 \cdot K^{1/p}} \cdot \left[(\pi^* - \pi_{\theta_t})^\top r \right]^2 \\
&\leq -\frac{c_t^{2-2/p}}{9 \cdot K^{1/p}} \cdot \left[(\pi^* - \pi_{\theta_t})^\top r \right]^2, \qquad (35)
\end{aligned}
$$

where $c_t \coloneqq \min_{1 \leq s \leq t} \pi_{\theta_s}(a^*) > 0$. Eq. (35) is equivalent to,

$$
(\pi^* - \pi_{\theta_{t+1}})^\top r - (\pi^* - \pi_{\theta_t})^\top r \leq -\frac{c_t^{2-2/p}}{9 \cdot K^{1/p}} \cdot \left[(\pi^* - \pi_{\theta_t})^\top r \right]^2. \qquad (36)
$$

Denote $\delta_t \coloneqq (\pi^* - \pi_{\theta_t})^\top r$. We prove $\delta_t \leq \frac{9 \cdot K^{1/p}}{c_t^{2-2/p}} \cdot \frac{1}{t}$ by induction. For $t = 2$, since $c_2 \in (0,1)$,

$$
\delta_2 \leq 1 \leq \frac{9 \cdot K^{1/p}}{c_2^{2-2/p}} \cdot \frac{1}{2}.
$$

Suppose $\delta_t \le \frac{9 \cdot K^{1/p}}{c_t^{2-2/p}} \cdot \frac{1}{t}$, $t \ge 2$. Consider $f_t : \mathbb{R} \to \mathbb{R}$, $f_t(x) \coloneqq x - \frac{c_t^{2-2/p}}{9 \cdot K^{1/p}} \cdot x^2$. Clearly, f_t is monotonically increasing in $\left[0, \frac{9 \cdot K^{1/p}}{2 \cdot c_t^{2-2/p}}\right]$. We have,

$$\begin{aligned}
\delta_{t+1} &\le \delta_t - \frac{c_t^{2-2/p}}{9 \cdot K^{1/p}} \cdot \delta_t^2 \qquad \text{(by Eq. (36))} \\
&\le \frac{9 \cdot K^{1/p}}{c_t^{2-2/p}} \cdot \frac{1}{t} - \frac{c_t^{2-2/p}}{9 \cdot K^{1/p}} \cdot \left(\frac{9 \cdot K^{1/p}}{c_t^{2-2/p}} \cdot \frac{1}{t} \right)^2 \qquad \left(\delta_t \le \frac{9 \cdot K^{1/p}}{c_t^{2-2/p}} \cdot \frac{1}{t} \le \frac{9 \cdot K^{1/p}}{2 \cdot c_t^{2-2/p}},\ t \ge 2 \right) \\
&= \frac{9 \cdot K^{1/p}}{c_t^{2-2/p}} \cdot \left(\frac{1}{t} - \frac{1}{t^2} \right) \\
&\le \frac{9 \cdot K^{1/p}}{c_t^{2-2/p}} \cdot \frac{1}{t+1}, \qquad (37)
\end{aligned}$$

which completes the proof for $\delta_t \le \frac{9 \cdot K^{1/p}}{c_t^{2-2/p}} \cdot \frac{1}{t}$. Then we have, for all $t \ge 1$,

$$(\pi^* - \pi_{\theta_t})^\top r \le \frac{9 \cdot K^{1/p}}{c_t^{2-2/p}} \cdot \frac{1}{t} \le \frac{9 \cdot K^{1/p}}{(\inf_{t \ge 1} \pi_{\theta_t}(a^*))^{2-2/p}} \cdot \frac{1}{t}.$$

Second part. For $p = 1$, according to Lemma 2,

$$\left| (\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| \le \frac{K}{\|\theta_{t,\xi}\|_1^2} \cdot \|\theta_{t+1} - \theta_t\|_2^2, \qquad (38)$$

where

$$\theta_{t,\xi} \coloneqq \theta_t + \xi \cdot (\theta_{t+1} - \theta_t) = \theta_t + \xi \cdot \eta_t \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t},$$

for some $\xi \in [0, 1]$. The ℓ_1 norm can be upper bounded as

$$\begin{aligned}
\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_1 &= \sum_{a=1}^K \left| \frac{\pi_{\theta_t}(a)}{\theta_t(a)} \cdot \left(r(a) - \pi_\theta^\top r \right) \right| \\
&= \frac{1}{\|\theta_t\|_1} \sum_{a=1}^K |r(a) - \pi_\theta^\top r| \\
&\le \frac{K}{\|\theta_t\|_1}. \qquad \left(r \in [0,1]^K \right) \qquad (39)
\end{aligned}$$

According to triangle inequality, we have,

$$\begin{aligned}
\|\theta_{t,\xi}\|_1 &\ge \|\theta_t\|_1 - \xi \cdot \eta_t \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_1 \\
&\ge \|\theta_t\|_1 - \xi \cdot \eta_t \cdot \frac{K}{\|\theta_t\|_1}. \qquad \text{(by Eq. (39))} \\
&= \|\theta_t\|_1 \cdot \left(1 - \xi \cdot \frac{2}{9} \right) \qquad \left(\eta_t = \frac{2 \cdot \|\theta_t\|_1^2}{9 \cdot K} \right) \\
&\ge \frac{2}{3} \cdot \|\theta_t\|_1. \qquad (\xi \in [0,1]) \qquad (40)
\end{aligned}$$

Combining Eqs. (38) and (40), we have,

$$\begin{aligned}
\left| (\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| &\le \frac{K}{\|\theta_{t,\xi}\|_1^2} \cdot \|\theta_{t+1} - \theta_t\|_1^2 \\
&\le \frac{9 \cdot K}{4 \cdot \|\theta_t\|_1^2} \cdot \|\theta_{t+1} - \theta_t\|_1^2,
\end{aligned}$$

which implies,

$$\begin{aligned}
\pi_{\theta_t}^\top r - \pi_{\theta_{t+1}}^\top r &\le -\Big\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \Big\rangle + \frac{9 \cdot K}{4 \cdot \|\theta_t\|_1^2} \cdot \|\theta_{t+1} - \theta_t\|_1^2 \\
&= -\eta_t \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 + \frac{9 \cdot K}{4 \cdot \|\theta_t\|_1^2} \cdot \eta_t^2 \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \qquad \left(\theta_{t+1} = \theta_t + \eta_t \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right) \\
&= -\frac{\|\theta_t\|_1^2}{9 \cdot K} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \qquad \left(\eta_t = \frac{2 \cdot \|\theta_t\|_1^2}{9 \cdot K} \right) \\
&\le -\frac{\cancel{\|\theta_t\|_1^2}}{9 \cdot K} \cdot \left[\frac{1}{\cancel{\|\theta_t\|_1}} \cdot (\pi^* - \pi_{\theta_t})^\top r \right]^2 \qquad \text{(by Lemma 3)} \\
&= -\frac{1}{9 \cdot K} \cdot \left[(\pi^* - \pi_{\theta_t})^\top r \right]^2. \qquad (41)
\end{aligned}$$

Using a similar induction argument as in Eq. (37), we have

$$(\pi^* - \pi_{\theta_t})^\top r \le \frac{9 \cdot K}{t}.$$

Third part. For the gradient flow, we have,

$$\begin{aligned}
\frac{d\{(\pi^* - \pi_{\theta_t})^\top r\}}{dt} &= -\frac{d\pi_{\theta_t}^\top r}{dt} \\
&= -\left(\frac{d\theta_t}{dt} \right)^\top \left(\frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right) \\
&= -\eta_t \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \qquad \left(\frac{d\theta_t}{dt} = \eta_t \cdot \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right) \\
&\le -\eta_t \cdot \left[\frac{p}{\|\theta_t\|_p} \cdot \pi_{\theta_t}(a^*)^{1-1/p} \cdot (\pi^* - \pi_{\theta_t})^\top r \right]^2 \qquad \text{(by Lemma 3)} \\
&= -\pi_{\theta_t}(a^*)^{2-2/p} \cdot \left[(\pi^* - \pi_{\theta_t})^\top r \right]^2 \qquad \left(\eta_t = \frac{\|\theta_t\|_p^2}{p^2} \right) \\
&\le -c^{2-2p} \cdot [(\pi^* - \pi_{\theta_t})^\top r]^2,
\end{aligned}$$

which implies,

$$\frac{d}{dt} \left\{ \frac{1}{(\pi^* - \pi_{\theta_t})^\top r} \right\} = -\frac{1}{[(\pi^* - \pi_{\theta_t})^\top r]^2} \cdot \frac{d\{(\pi^* - \pi_{\theta_t})^\top r\}}{dt} = c^{2-2p}.$$

Taking integral, we have,

$$\begin{aligned}
\frac{1}{(\pi^* - \pi_{\theta_t})^\top r} &= \frac{1}{(\pi^* - \pi_{\theta_1})^\top r} + c^{2-2p} \cdot (t-1) \\
&\ge 1 + c^{2-2p} \cdot (t-1), \qquad \left((\pi^* - \pi_{\theta_1})^\top r \in (0, 1] \right)
\end{aligned}$$

which is equivalent to

$$(\pi^* - \pi_{\theta_t})^\top r \le \frac{1}{c^{2-2p} \cdot (t-1) + 1}.$$

□

C.3 General MDPs

Lemma 4 (Policy gradient theorem [19]). *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,*

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

Lemma 5. *The escort policy gradient w.r.t. θ is*

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

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

$$\begin{aligned} A^{\pi_\theta}(s,a) &= Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s), \\ Q^{\pi_\theta}(s,a) &= r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) V^{\pi_\theta}(s'). \end{aligned}$$

Proof. According to Lemma 4, we have,

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

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

$$\begin{aligned} \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} &= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \left[\sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)} \cdot Q^{\pi_\theta}(s,a) \right] \\ &= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot \left(\frac{d\pi(\cdot|s)}{d\theta(s,\cdot)} \right)^\top Q^{\pi_\theta}(s,\cdot) \\ &= \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot p \cdot \mathrm{diag}\left(\frac{\pi(\cdot|s)}{\theta(s,\cdot)} \right) \left(\mathbf{Id} - \mathbf{1}\pi_\theta^\top \right) Q^{\pi_\theta}(s,\cdot). \end{aligned}$$

For each component a, we have

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

□

Lemma 6 (Non-uniform smoothness)**.** *Suppose $r(s,a) \in [0,1]$ for all (s,a). Let $\pi_\theta \coloneqq f_p(\theta)$, and $\pi_{\theta'} \coloneqq f_p(\theta')$. Denote $\theta_\xi \coloneqq \theta + \xi \cdot (\theta' - \theta)$ with some $\xi \in [0,1]$. Denote $A \coloneqq |\mathcal{A}|$ as the total number of actions. Then we have,*

- *for $p \geq 2$, $V^{\pi_\theta}(\rho)$ is $\frac{8 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \frac{1}{\min_s \|\theta_\xi(s,\cdot)\|_p^2}$-smooth, i.e.,*

$$\left| V^{\pi_{\theta'}}(\rho) - V^{\pi_\theta}(\rho) - \left\langle \frac{\partial V^{\pi_\theta}(\rho)}{\partial \theta}, \theta' - \theta \right\rangle \right| \leq \frac{4 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \frac{\|\theta' - \theta\|_2^2}{\min_s \|\theta_\xi(s,\cdot)\|_p^2};$$

- *for $p = 1$, $V^{\pi_\theta}(\rho)$ is $\frac{8 \cdot A^2}{(1-\gamma)^3} \cdot \frac{1}{\min_s \|\theta_\xi(s,\cdot)\|_1^2}$-smooth, i.e.,*

$$\left| V^{\pi_{\theta'}}(\rho) - V^{\pi_\theta}(\rho) - \left\langle \frac{\partial V^{\pi_\theta}(\rho)}{\partial \theta}, \theta' - \theta \right\rangle \right| \leq \frac{4 \cdot A^2}{(1-\gamma)^3} \cdot \frac{\|\theta' - \theta\|_2^2}{\min_s \|\theta_\xi(s,\cdot)\|_1^2}.$$

Proof. 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 $s' \neq s$,

$$\begin{aligned}
\sum_a \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha}\Big|_{\alpha=0} \right| &= \sum_a \left| \Big\langle \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)}, u(s,\cdot) \Big\rangle \right| \\
&= \sum_a p \cdot \frac{\pi_\theta(a|s)}{|\theta(s,a)|} \cdot \left| u(s,a) - \pi_\theta(\cdot|s)^\top u(s,\cdot) \right| \\
&= \sum_a \frac{p}{\|\theta(s,\cdot)\|_p} \cdot \pi_\theta(a|s)^{1-1/p} \cdot \left| u(s,a) - \pi_\theta(\cdot|s)^\top u(s,\cdot) \right| \\
&\le \frac{p}{\|\theta(s,\cdot)\|_p} \cdot \max_a \left| u(s,a) - \pi_\theta(\cdot|s)^\top u(s,\cdot) \right| \cdot \sum_a \pi_\theta(a|s)^{1-1/p} \\
&\le \frac{p}{\|\theta(s,\cdot)\|_p} \cdot 2 \cdot \|u\|_\infty \cdot A^{1/p} \qquad \text{(by Eq. (29))} \\
&\le \frac{2 \cdot p \cdot A^{1/p}}{\|\theta(s,\cdot)\|_p} \cdot \|u\|_2. \qquad (42)
\end{aligned}$$

Similarly, the second derivative is,

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

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,

$$\begin{aligned}
S_{i,j} &= p \cdot \frac{\partial \{ \delta_{ia} \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)} - \pi_\theta(a|s) \cdot \frac{\pi_\theta(i|s)}{\theta(s,i)} \}}{\partial \theta(s,j)} \\
&= p \cdot \delta_{ia} \cdot \frac{\frac{p}{\theta(s,j)} \cdot [\delta_{ja} \pi_\theta(a|s) - \pi_\theta(a|s)\pi_\theta(j|s)] \cdot \theta(s,a) - \delta_{ja}\pi_\theta(a|s)}{\theta(s,a)^2} \\
&\quad - \frac{p^2}{\theta(s,j)} \cdot [\delta_{ja}\pi_\theta(a|s) - \pi_\theta(a|s)\pi_\theta(j|s)] \cdot \frac{\pi_\theta(i|s)}{\theta(s,i)} \\
&\quad - p \cdot \pi_\theta(a|s) \cdot \frac{\frac{p}{\theta(s,j)} \cdot [\delta_{ij}\pi_\theta(i|s) - \pi_\theta(i|s)\pi_\theta(j|s)] \cdot \theta(s,i) - \delta_{ij}\pi_\theta(i|s)}{\theta(s,i)^2} \\
&= \delta_{ia} \cdot \delta_{ja} \cdot p \cdot (p-1) \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)^2} - \delta_{ia} \cdot p^2 \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)} \cdot \frac{\pi_\theta(j|s)}{\theta(s,j)} - \delta_{ja} \cdot p^2 \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)} \cdot \frac{\pi_\theta(i|s)}{\theta(s,i)} \\
&\quad + p \cdot \pi_\theta(a|s) \cdot \left[2 \cdot p \cdot \frac{\pi_\theta(i|s)}{\theta(s,i)} \cdot \frac{\pi_\theta(j|s)}{\theta(s,j)} - \delta_{ij} \cdot (p-1) \cdot \frac{\pi_\theta(i|s)}{\theta(s,i)^2} \right],
\end{aligned}$$

where the δ notation is as defined in Eq. (23). Then we have,

$$
\begin{aligned}
&\left| \left\langle \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} u(s,\cdot), u(s,\cdot) \right\rangle \right| = \left| \sum_{i=1}^{A} \sum_{j=1}^{A} S_{i,j} u(s,i) u(s,j) \right| \\
&\quad \le p \cdot (p-1) \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)^2} \cdot u(s,a)^2 + 2 \cdot p^2 \cdot \frac{\pi_\theta(a|s)}{|\theta(s,a)|} \cdot |u(s,a)| \cdot \left| \left(\frac{\pi_\theta(\cdot|s)}{\theta(s,\cdot)} \right)^\top u(s,\cdot) \right| \\
&\qquad + \pi_\theta(a|s) \cdot \left[2 \cdot p^2 \cdot \left| \left(\frac{\pi_\theta(\cdot|s)}{\theta(s,\cdot)} \right)^\top u(s,\cdot) \right|^2 + p \cdot (p-1) \cdot \left| \left(\frac{\pi_\theta(\cdot|s)}{\theta(s,\cdot)^2} \right)^\top (u(s,\cdot) \odot u(s,\cdot)) \right| \right] \\
&\quad = \frac{p \cdot (p-1)}{\|\theta(s,\cdot)\|_p^2} \cdot \pi_\theta(a|s)^{1-2/p} \cdot u(s,a)^2 \\
&\qquad + \frac{2 \cdot p^2}{\|\theta(s,\cdot)\|_p^2} \cdot \pi_\theta(a|s)^{1-1/p} \cdot |u(s,a)| \cdot \left| \left(\pi_\theta(\cdot|s)^{1-1/p} \right)^\top u(s,\cdot) \right| \\
&\qquad + \frac{2 \cdot p^2}{\|\theta(s,\cdot)\|_p^2} \cdot \pi_\theta(a|s) \cdot \left| \left(\pi_\theta(\cdot|s)^{1-1/p} \right)^\top u(s,\cdot) \right|^2 \\
&\qquad + \frac{p \cdot (p-1)}{\|\theta(s,\cdot)\|_p^2} \cdot \pi_\theta(a|s) \cdot \left| \left(\pi_\theta(\cdot|s)^{1-2/p} \right)^\top (u(s,\cdot) \odot u(s,\cdot)) \right|. \qquad (43)
\end{aligned}
$$

First part. For $p \ge 2$, according to the Cauchy-Schwarz inequality, we have,

$$
\begin{aligned}
\sum_a \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2} \Big|_{\alpha=0} \right| &\le \frac{p \cdot (p-1)}{\|\theta(s,\cdot)\|_p^2} \cdot \sum_a u(s,a)^2 + \frac{2 \cdot p^2}{\|\theta(s,\cdot)\|_p^2} \cdot \|u(s,\cdot)\|_2^2 \cdot \|\pi_\theta(\cdot|s)^{1-1/p}\|_2^2 \\
&\quad + \frac{2 \cdot p^2}{\|\theta(s,\cdot)\|_p^2} \cdot \|u(s,\cdot)\|_2^2 \cdot \|\pi_\theta(\cdot|s)^{1-1/p}\|_2^2 + \frac{p \cdot (p-1)}{\|\theta(s,\cdot)\|_p^2} \cdot \| \cdot \|\pi_\theta(\cdot|s)^{1-2/p}\|_\infty \cdot \|u(s,\cdot) \odot u(s,\cdot)\|_1 \\
&\le \frac{2 \cdot p \cdot (p-1)}{\|\theta(s,\cdot)\|_p^2} \cdot \|u(s,\cdot)\|_2^2 + \frac{4 \cdot p^2}{\|\theta(s,\cdot)\|_p^2} \cdot A^{1/p} \cdot \|u(s,\cdot)\|_2^2 \qquad \text{(by Eq. (29))} \\
&\le \frac{2 \cdot p^2 \cdot (1 + 2 \cdot A^{1/p})}{\|\theta(s,\cdot)\|_p^2} \cdot \|u\|_2^2. \qquad (44)
\end{aligned}
$$

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

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

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,

$$
\begin{aligned}
\left\| \frac{\partial P(\alpha)}{\partial \alpha} \Big|_{\alpha=0} x \right\|_\infty &= \max_s \left| \sum_{s'} \sum_a \left[\frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \Big|_{\alpha=0} \right] \cdot \mathcal{P}(s'|s,a) \cdot x(s') \right| \\
&\le \max_s \sum_a \sum_{s'} \mathcal{P}(s'|s,a) \cdot \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \Big|_{\alpha=0} \right| \cdot \|x\|_\infty \\
&= \max_s \sum_a \left| \frac{\partial \pi_{\theta_\alpha}(a|s)}{\partial \alpha} \Big|_{\alpha=0} \right| \cdot \|x\|_\infty \\
&\le \max_s \frac{2 \cdot p \cdot A^{1/p}}{\|\theta(s,\cdot)\|_p} \cdot \|u\|_2 \cdot \|x\|_\infty. \qquad \text{(by Eq. (42))} \qquad (46)
\end{aligned}
$$

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

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

The ℓ_∞ norm is upper bounded as,

$$\begin{aligned} \left\| \frac{\partial^2 P(\alpha)}{\partial \alpha^2} \Big|_{\alpha=0} x \right\|_\infty &= \max_s \left| \sum_{s'} \sum_a \left[\frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2} \Big|_{\alpha=0} \right] \cdot \mathcal{P}(s'|s,a) \cdot x(s') \right| \\ &\le \max_s \sum_a \sum_{s'} \mathcal{P}(s'|s,a) \cdot \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2} \Big|_{\alpha=0} \right| \cdot \|x\|_\infty \\ &= \max_s \sum_a \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2} \Big|_{\alpha=0} \right| \cdot \|x\|_\infty \\ &\le \max_s \frac{2 \cdot p^2 \cdot (1 + 2 \cdot A^{1/p})}{\|\theta(s,\cdot)\|_p^2} \cdot \|u\|_2^2 \cdot \|x\|_\infty. \qquad \text{(by Eq. (44))} \end{aligned} \tag{47}$$

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

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

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

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

which implies, $\forall i$,

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

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

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

Since $r(s,a) \in [0,1], \forall (s,a)$, we have,

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

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

$$\begin{aligned}
\left| \frac{\partial r_{\theta_\alpha}(s)}{\partial \alpha} \right| &= \left| \left(\frac{\partial r_{\theta_\alpha}(s)}{\partial \theta_\alpha} \right)^\top \frac{\partial \theta_\alpha}{\partial \alpha} \right| \\
&= \left| \left(\frac{\partial \{ \pi_{\theta_\alpha}(\cdot|s)^\top r(s,\cdot) \}}{\partial \theta_\alpha(s,\cdot)} \right)^\top u(s,\cdot) \right| \\
&= \left| p \cdot \left(\text{diag}(1/\theta_\alpha(s,\cdot)) \left(\text{diag}(\pi_{\theta_\alpha}(\cdot|s)) - \pi_{\theta_\alpha}(\cdot|s) \pi_{\theta_\alpha}(\cdot|s)^\top \right) r(s,\cdot) \right)^\top u(s,\cdot) \right| \\
&\leq p \cdot \left\| \text{diag}(1/\theta_\alpha(s,\cdot)) \left(\text{diag}(\pi_{\theta_\alpha}(\cdot|s)) - \pi_{\theta_\alpha}(\cdot|s) \pi_{\theta_\alpha}(\cdot|s)^\top \right) r(s,\cdot) \right\|_1 \cdot \| u(s,\cdot) \|_\infty . \qquad (52)
\end{aligned}$$

The ℓ_1 norm is upper bounded as,

$$\begin{aligned}
&\left\| \text{diag}(1/\theta_\alpha(s,\cdot)) \left(\text{diag}(\pi_{\theta_\alpha}(\cdot|s)) - \pi_{\theta_\alpha}(\cdot|s) \pi_{\theta_\alpha}(\cdot|s)^\top \right) r(s,\cdot) \right\|_1 \\
&= \sum_a \frac{\pi_{\theta_\alpha}(a|s)^{1-1/p}}{\| \theta_\alpha(s,\cdot) \|_p} \cdot \left| r(s,a) - \pi_{\theta_\alpha}(\cdot|s)^\top r(s,\cdot) \right| \\
&\leq \frac{1}{\| \theta_\alpha(s,\cdot) \|_p} \cdot \max_a \left| r(s,a) - \pi_{\theta_\alpha}(\cdot|s)^\top r(s,\cdot) \right| \cdot \sum_a \pi_{\theta_\alpha}(a|s)^{1-1/p} \\
&\leq \frac{1}{\| \theta_\alpha(s,\cdot) \|_p} \cdot \sum_a \pi_{\theta_\alpha}(a|s)^{1-1/p} \qquad (r(s,a) \in [0,1]) \\
&\leq \frac{A^{1/p}}{\| \theta_\alpha(s,\cdot) \|_p}. \qquad \text{(by Eq. (29))} \qquad (53)
\end{aligned}$$

Combining Eqs. (52) and (53), we have,

$$\begin{aligned}
\left\| \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \right\|_\infty &= \max_s \left| \frac{\partial r_{\theta_\alpha}(s)}{\partial \alpha} \right| \\
&\leq \max_s p \cdot \left\| \text{diag}(1/\theta_\alpha(s,\cdot)) \left(\text{diag}(\pi_{\theta_\alpha}(\cdot|s)) - \pi_{\theta_\alpha}(\cdot|s) \pi_{\theta_\alpha}(\cdot|s)^\top \right) r(s,\cdot) \right\|_1 \cdot \| u(s,\cdot) \|_\infty \\
&\leq \max_s \frac{p \cdot A^{1/p}}{\| \theta_\alpha(s,\cdot) \|_p} \cdot \| u \|_2. \qquad (54)
\end{aligned}$$

Similarly, for the second derivative, we have,

$$\begin{aligned}
\left\| \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2} \right\|_\infty &= \max_s \left| \frac{\partial^2 r_{\theta_\alpha}(s)}{\partial \alpha^2} \right| \\
&= \max_s \left| \left(\frac{\partial}{\partial \theta_\alpha} \left\{ \frac{\partial r_{\theta_\alpha}(s)}{\partial \alpha} \right\} \right)^\top \frac{\partial \theta_\alpha}{\partial \alpha} \right| \\
&= \max_s \left| \left(\frac{\partial^2 r_{\theta_\alpha}(s)}{\partial \theta_\alpha^2} \frac{\partial \theta_\alpha}{\partial \alpha} \right)^\top \frac{\partial \theta_\alpha}{\partial \alpha} \right| \\
&= \max_s \left| u(s,\cdot)^\top \frac{\partial^2 \{ \pi_{\theta_\alpha}(\cdot|s)^\top r(s,\cdot) \}}{\partial \theta_\alpha(s,\cdot)^2} u(s,\cdot) \right| \\
&\leq \max_s \frac{3 \cdot p^2 \cdot A^{1/p}}{\| \theta_\alpha(s,\cdot) \|_p^2} \cdot \| u(s,\cdot) \|_2^2 \qquad \text{(by Eq. (30))} \\
&\leq \max_s \frac{3 \cdot p^2 \cdot A^{1/p}}{\| \theta_\alpha(s,\cdot) \|_p^2} \cdot \| u \|_2^2. \qquad (55)
\end{aligned}$$

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

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

Taking second derivative w.r.t. α, we have,

$$\begin{aligned}\frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2} &= 2\gamma^2 \cdot e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} + \gamma \cdot e_s^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha} \\ &\quad + 2\gamma \cdot e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} + e_s^\top M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}. \end{aligned} \tag{56}$$

For the last term,

$$\begin{aligned}\left| e_s^\top M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0} \right| &\le \|e_s\|_1 \cdot \left\| M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0} \right\|_\infty \\ &\le \frac{1}{1-\gamma} \cdot \left\| \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0} \right\|_\infty \qquad \text{(by Eq. (50))} \\ &\le \frac{1}{1-\gamma} \cdot \max_s \frac{3 \cdot p^2 \cdot A^{1/p}}{\|\theta_\alpha(s,\cdot)\|_p^2} \cdot \|u\|_2^2. \qquad \text{(by Eq. (55))} \end{aligned} \tag{57}$$

For the second last term,

$$\begin{aligned}&\left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} \right| \le \left\| M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} \right\|_\infty \\ &\le \frac{1}{1-\gamma} \cdot \left\| \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} \right\|_\infty \qquad \text{(by Eq. (50))} \\ &\le \frac{2 \cdot p \cdot A^{1/p} \cdot \|u\|_2}{1-\gamma} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p} \cdot \left\| M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} \right\|_\infty \qquad \text{(by Eq. (46))} \\ &\le \frac{2 \cdot p \cdot A^{1/p} \cdot \|u\|_2}{(1-\gamma)^2} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p} \cdot \left\| \frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} \right\|_\infty \qquad \text{(by Eq. (50))} \\ &\le \frac{2 \cdot p^2 \cdot A^{2/p} \cdot \|u\|_2}{(1-\gamma)^2} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p^2} \cdot \|u\|_2. \qquad \text{(by Eq. (54))} \end{aligned} \tag{58}$$

For the second term,

$$\begin{aligned}&\left| e_s^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha}\Big|_{\alpha=0} \right| \le \left\| M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha}\Big|_{\alpha=0} \right\|_\infty \\ &\le \frac{1}{1-\gamma} \cdot \left\| \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha}\Big|_{\alpha=0} \right\|_\infty \qquad \text{(by Eq. (50))} \\ &\le \frac{2 \cdot p^2 \cdot (1 + 2 \cdot A^{1/p}) \cdot \|u\|_2^2}{1-\gamma} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p^2} \cdot \left\| M(\alpha) r_{\theta_\alpha}\Big|_{\alpha=0} \right\|_\infty \qquad \text{(by Eq. (47))} \\ &\le \frac{2 \cdot p^2 \cdot (1 + 2 \cdot A^{1/p}) \cdot \|u\|_2^2}{(1-\gamma)^2} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p^2} \cdot \left\| r_{\theta_\alpha}\Big|_{\alpha=0} \right\|_\infty \qquad \text{(by Eq. (50))} \\ &\le \frac{2 \cdot p^2 \cdot (1 + 2 \cdot A^{1/p})}{(1-\gamma)^2} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p^2} \cdot \|u\|_2^2. \qquad \text{(by Eq. (51))} \end{aligned} \tag{59}$$

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

$$\begin{aligned}&\left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha}\Big|_{\alpha=0} \right| \le \left\| M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha}\Big|_{\alpha=0} \right\|_\infty \\ &\le \frac{1}{1-\gamma} \cdot \max_s \frac{2 \cdot p \cdot A^{1/p}}{\|\theta(s,\cdot)\|_p} \cdot \|u\|_2 \cdot \frac{1}{1-\gamma} \cdot \max_s \frac{2 \cdot p \cdot A^{1/p}}{\|\theta(s,\cdot)\|_p} \cdot \|u\|_2 \cdot \frac{1}{1-\gamma} \cdot 1 \\ &= \frac{4 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p^2} \cdot \|u\|_2^2. \end{aligned} \tag{60}$$

Combining Eqs. (57) to (60) with Eq. (56), we have,

$$\begin{aligned}\left|\frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2}\Big|_{\alpha=0}\right| &\le 2\gamma^2 \cdot \left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha} \Big|_{\alpha=0}\right| \\ &\quad + \gamma \cdot \left| e_s^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha} \Big|_{\alpha=0}\right| \\ &\quad + 2\gamma \cdot \left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha} \Big|_{\alpha=0}\right| + \left| e_s^\top M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2} \Big|_{\alpha=0}\right| \\ &\le \left(\frac{8 \cdot \gamma^2 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} + \frac{2 \cdot \gamma \cdot p^2 \cdot (1 + 2 \cdot A^{1/p})}{(1-\gamma)^2} + \frac{4 \cdot \gamma \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^2} + \frac{3 \cdot p^2 \cdot A^{1/p}}{1-\gamma} \right) \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p^2} \cdot \|u\|_2^2 \\ &\le \frac{8 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p^2} \cdot \|u\|_2^2, \qquad (61)\end{aligned}$$

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

$$\begin{aligned}\left| y^\top \frac{\partial^2 V^{\pi_\theta}(s)}{\partial \theta^2} y \right| &= \left| \left(\frac{y}{\|y\|_2} \right)^\top \frac{\partial^2 V^{\pi_\theta}(s)}{\partial \theta^2} \left(\frac{y}{\|y\|_2} \right) \right| \cdot \|y\|_2^2 \\ &\le \max_{\|u\|_2=1} \left| \left\langle \frac{\partial^2 V^{\pi_\theta}(s)}{\partial \theta^2} u, u \right\rangle \right| \cdot \|y\|_2^2 \\ &= \max_{\|u\|_2=1} \left| \left\langle \frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \theta_\alpha^2}\Big|_{\alpha=0} \frac{\partial \theta_\alpha}{\partial \alpha}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \cdot \|y\|_2^2 \\ &= \max_{\|u\|_2=1} \left| \left\langle \frac{\partial}{\partial \theta_\alpha} \left\{ \frac{\partial V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha} \right\} \Big|_{\alpha=0}, \frac{\partial \theta_\alpha}{\partial \alpha} \right\rangle \right| \cdot \|y\|_2^2 \\ &= \max_{\|u\|_2=1} \left| \frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2}\Big|_{\alpha=0} \right| \cdot \|y\|_2^2 \\ &\le \frac{8 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p^2} \cdot \|y\|_2^2. \qquad \text{(by Eq. (61))} \qquad (62)\end{aligned}$$

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

$$\begin{aligned}\left| V^{\pi_{\theta'}}(s) - V^{\pi_\theta}(s) - \left\langle \frac{\partial V^{\pi_\theta}(s)}{\partial \theta}, \theta' - \theta \right\rangle \right| &= \frac{1}{2} \cdot \left| (\theta' - \theta)^\top \frac{\partial^2 V^{\pi_{\theta_\xi}}(s)}{\partial \theta_\xi^2} (\theta' - \theta) \right| \\ &\le \frac{4 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \max_s \frac{1}{\|\theta_\xi(s,\cdot)\|_p^2} \cdot \|\theta' - \theta\|_2^2 \qquad \text{(by Eq. (62))} \\ &= \frac{4 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \frac{1}{\min_s \|\theta_\xi(s,\cdot)\|_p^2} \cdot \|\theta' - \theta\|_2^2. \qquad (63)\end{aligned}$$

Since $V^{\pi_\theta}(s)$ is $\frac{8 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \frac{1}{\min_s \|\theta_\xi(s,\cdot)\|_p^2}$-smooth, for any state s, $V^{\pi_\theta}(\rho) = \mathbb{E}_{s\sim\rho}\left[V^{\pi_\theta}(s)\right]$ is also $\frac{8 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \frac{1}{\min_s \|\theta_\xi(s,\cdot)\|_p^2}$-smooth.

Second part. For $p = 1$, we have,

$$\begin{aligned}\left| \left\langle \frac{\partial^2 \pi_\theta(a|s)}{\partial \theta^2(s,\cdot)} u(s,\cdot), u(s,\cdot) \right\rangle \right| &\le \frac{2 \cdot p^2}{\|\theta(s,\cdot)\|_p^2} \cdot \pi_\theta(a|s)^{1-1/p} \cdot |u(s,a)| \cdot \left| \left(\pi_\theta(\cdot|s)^{1-1/p} \right)^\top u(s,\cdot) \right| \\ &\quad + \frac{2 \cdot p^2}{\|\theta(s,\cdot)\|_p^2} \cdot \pi_\theta(a|s) \cdot \left| \left(\pi_\theta(\cdot|s)^{1-1/p} \right)^\top u(s,\cdot) \right|^2. \qquad \text{(by Eq. (43))}\end{aligned}$$

Therefore we have,

$$\begin{aligned}\sum_a \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\Big|_{\alpha=0} \right| &\le \frac{4 \cdot p^2}{\|\theta(s,\cdot)\|_p^2} \cdot \|u(s,\cdot)\|_2^2 \cdot \|\pi_\theta(\cdot|s)^{1-1/p}\|_2^2 \\ &\le \frac{4 \cdot p^2 \cdot A}{\|\theta(s,\cdot)\|_p^2} \cdot \|u\|_2^2. \qquad (64)\end{aligned}$$

Similar to Eq. (47), we have,

$$\begin{aligned}\left\| \frac{\partial^2 P(\alpha)}{\partial \alpha^2}\Big|_{\alpha=0} x \right\|_\infty &\le \max_s \sum_a \left| \frac{\partial^2 \pi_{\theta_\alpha}(a|s)}{\partial \alpha^2}\Big|_{\alpha=0} \right| \cdot \|x\|_\infty \\ &\le \max_s \frac{4 \cdot p^2 \cdot A}{\|\theta(s,\cdot)\|_p^2} \cdot \|u\|_2^2 \cdot \|x\|_\infty. \qquad \text{(by Eq. (64))}\end{aligned} \tag{65}$$

Similar to Eq. (55), for the second derivative, we have,

$$\begin{aligned}\left\| \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2} \right\|_\infty &= \max_s \left| u(s,\cdot)^\top \frac{\partial^2 \{\pi_{\theta_\alpha}(\cdot|s)^\top r(s,\cdot)\}}{\partial \theta_\alpha(s,\cdot)^2} u(s,\cdot) \right| \\ &\le \max_s \frac{2 \cdot p^2 \cdot A^{1/p}}{\|\theta_\alpha(s,\cdot)\|_p^2} \cdot \|u(s,\cdot)\|_2^2 \qquad \text{(by Eq. (31))} \\ &\le \max_s \frac{2 \cdot p^2 \cdot A^{1/p}}{\|\theta_\alpha(s,\cdot)\|_p^2} \cdot \|u\|_2^2.\end{aligned} \tag{66}$$

Similar to Eq. (57), we have,

$$\begin{aligned}\left| e_s^\top M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0} \right| &\le \|e_s\|_1 \cdot \left\| M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0} \right\|_\infty \\ &\le \frac{1}{1-\gamma} \cdot \left\| \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0} \right\|_\infty \qquad \text{(by Eq. (50))} \\ &\le \frac{1}{1-\gamma} \cdot \max_s \frac{2 \cdot p^2 \cdot A^{1/p}}{\|\theta_\alpha(s,\cdot)\|_p^2} \cdot \|u\|_2^2. \qquad \text{(by Eq. (66))}\end{aligned} \tag{67}$$

Similar to Eq. (59), we have,

$$\begin{aligned}\left| e_s^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha}\Big|_{\alpha=0} \right| &\le \frac{1}{1-\gamma} \cdot \left\| \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha}\Big|_{\alpha=0} \right\|_\infty \qquad \text{(by Eq. (50))} \\ &\le \frac{4 \cdot p^2 \cdot A \cdot \|u\|_2^2}{1-\gamma} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p^2} \cdot \left\| M(\alpha) r_{\theta_\alpha}\Big|_{\alpha=0} \right\|_\infty \qquad \text{(by Eq. (65))} \\ &\le \frac{4 \cdot p^2 \cdot A \cdot \|u\|_2^2}{(1-\gamma)^2} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p^2} \cdot \left\| r_{\theta_\alpha}\Big|_{\alpha=0} \right\|_\infty \qquad \text{(by Eq. (50))} \\ &\le \frac{4 \cdot p^2 \cdot A}{(1-\gamma)^2} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p^2} \cdot \|u\|_2^2. \qquad \text{(by Eq. (51))}\end{aligned} \tag{68}$$

Combining Eqs. (58) and (60), Eqs. (67) and (68) with Eq. (56), we have,

$$\begin{aligned}\left| \frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2}\Big|_{\alpha=0} \right| &\le 2\gamma^2 \cdot \left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) r_{\theta_\alpha}\Big|_{\alpha=0} \right| \\ &\quad + \gamma \cdot \left| e_s^\top M(\alpha) \frac{\partial^2 P(\alpha)}{\partial \alpha^2} M(\alpha) r_{\theta_\alpha}\Big|_{\alpha=0} \right| \\ &\quad + 2\gamma \cdot \left| e_s^\top M(\alpha) \frac{\partial P(\alpha)}{\partial \alpha} M(\alpha) \frac{\partial r_{\theta_\alpha}}{\partial \alpha}\Big|_{\alpha=0} \right| + \left| e_s^\top M(\alpha) \frac{\partial^2 r_{\theta_\alpha}}{\partial \alpha^2}\Big|_{\alpha=0} \right| \\ &\le \left(\frac{8 \cdot \gamma^2 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} + \frac{4 \cdot \gamma \cdot p^2 \cdot A}{(1-\gamma)^2} + \frac{4 \cdot \gamma \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^2} + \frac{2 \cdot p^2 \cdot A^{1/p}}{1-\gamma} \right) \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p^2} \cdot \|u\|_2^2 \\ &\le \frac{8 \cdot A^2}{(1-\gamma)^3} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_1^2} \cdot \|u\|_2^2. \qquad (p = 1)\end{aligned} \tag{69}$$

Similar to Eq. (62), Eq. (69) implies for all $y \in \mathbb{R}^{SA}$ and θ,

$$\begin{aligned}\left| y^\top \frac{\partial^2 V^{\pi_\theta}(s)}{\partial \theta^2} y \right| &\le \max_{\|u\|_2=1} \left| \frac{\partial^2 V^{\pi_{\theta_\alpha}}(s)}{\partial \alpha^2}\Big|_{\alpha=0} \right| \cdot \|y\|_2^2 \\ &\le \frac{8 \cdot A^2}{(1-\gamma)^3} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_1^2} \cdot \|y\|_2^2. \qquad \text{(Eq. (69))}\end{aligned} \tag{70}$$

Similar to Eq. (63), we have, $\forall s, \forall \theta,\ \theta'$,

$$\left| V^{\pi_{\theta'}}(s) - V^{\pi_\theta}(s) - \left\langle \frac{\partial V^{\pi_\theta}(s)}{\partial \theta}, \theta' - \theta \right\rangle \right| \le \frac{4 \cdot A^2}{(1-\gamma)^3} \cdot \max_s \frac{\|\theta' - \theta\|_2^2}{\|\theta_\xi(s,\cdot)\|_1^2} \qquad \text{(Eq. (70))}$$
$$= \frac{4 \cdot A^2}{(1-\gamma)^3} \cdot \frac{\|\theta' - \theta\|_2^2}{\min_s \|\theta_\xi(s,\cdot)\|_1^2}.$$

Since $V^{\pi_\theta}(s)$ is $\frac{8 \cdot A^2}{(1-\gamma)^3} \cdot \frac{1}{\min_s \|\theta_\xi(s,\cdot)\|_1^2}$-smooth, for any state s, $V^{\pi_\theta}(\rho) = \mathbb{E}_{s\sim\rho}\left[V^{\pi_\theta}(s)\right]$ is also $\frac{8 \cdot A^2}{(1-\gamma)^3} \cdot \frac{1}{\min_s \|\theta_\xi(s,\cdot)\|_1^2}$-smooth. □

Lemma 7 (Non-uniform Łojasiewicz). *Suppose $\mu(s) > 0$ for all state s and $\pi_\theta \coloneqq f_p(\theta)$. Then,*

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

where $a^*(s) \coloneqq \arg\max_a \pi^*(a|s)$, $\forall s \in \mathcal{S}$, is the action that the optimal policy π^* takes under s.

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

$$\begin{aligned}
\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 &= \left[\sum_{s,a} \left(\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a)} \right)^2 \right]^{\frac{1}{2}} \\
&\ge \left[\sum_{s} \left(\frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a^*(s))} \right)^2 \right]^{\frac{1}{2}} \\
&\ge \frac{1}{\sqrt{S}} \sum_s \left| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s,a^*(s))} \right| \qquad \text{(by Cauchy-Schwarz, } \|x\|_1 = |\langle \mathbf{1}, |x| \rangle| \le \|\mathbf{1}\|_2 \cdot \|x\|_2) \\
&= \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_s \left| d_\mu^{\pi_\theta}(s) \cdot p \cdot \frac{\pi_\theta(a^*(s)|s)}{\theta(s,a^*(s))} \cdot A^{\pi_\theta}(s,a^*(s)) \right| \qquad \text{(by Lemma 5)} \\
&= \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_s d_\mu^{\pi_\theta}(s) \cdot p \cdot \frac{\pi_\theta(a^*(s)|s)}{|\theta(s,a^*(s))|} \cdot |A^{\pi_\theta}(s,a^*(s))|. \qquad \left(d_\mu^{\pi_\theta}(s) \ge 0,\ \pi_\theta(a^*(s)|s) \ge 0\right)
\end{aligned}$$

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

$$\begin{aligned}
\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 &\ge \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_s \frac{d_\mu^{\pi_\theta}(s)}{d_\rho^{\pi^*}(s)} \cdot d_\rho^{\pi^*}(s) \cdot p \cdot \frac{\pi_\theta(a^*(s)|s)}{|\theta(s,a^*(s))|} \cdot |A^{\pi_\theta}(s,a^*(s))| \\
&= \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \sum_s \frac{d_\mu^{\pi_\theta}(s)}{d_\rho^{\pi^*}(s)} \cdot d_\rho^{\pi^*}(s) \cdot p \cdot \frac{1}{\|\theta(s,\cdot)\|_p} \cdot \left(\pi_\theta(a^*(s)|s)\right)^{1-1/p} \cdot |A^{\pi_\theta}(s,a^*(s))| \\
&\ge \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot p \cdot \min_s \frac{1}{\|\theta(s,\cdot)\|_p} \cdot \min_s \pi_\theta(a^*(s)|s)^{1-1/p} \cdot \sum_s d_\rho^{\pi^*}(s) \cdot |A^{\pi_\theta}(s,a^*(s))| \\
&\ge \frac{1}{1-\gamma} \cdot \frac{1}{\sqrt{S}} \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot p \cdot \min_s \frac{1}{\|\theta(s,\cdot)\|_p} \cdot \min_s \pi_\theta(a^*(s)|s)^{1-1/p} \cdot \sum_s d_\rho^{\pi^*}(s) \cdot A^{\pi_\theta}(s,a^*(s)) \\
&= \frac{p}{\sqrt{S}} \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot \frac{\min_s \pi_\theta(a^*(s)|s)^{1-1/p}}{\max_s \|\theta(s,\cdot)\|_p} \cdot \frac{1}{1-\gamma} \sum_s d_\rho^{\pi^*}(s) \sum_a \pi^*(a|s) \cdot A^{\pi_\theta}(s,a) \\
&= \frac{p}{\sqrt{S}} \cdot \left\| \frac{d_\rho^{\pi^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-1} \cdot \frac{\min_s \pi_\theta(a^*(s)|s)^{1-1/p}}{\max_s \|\theta(s,\cdot)\|_p} \cdot \left[V^*(\rho) - V^{\pi_\theta}(\rho)\right],
\end{aligned}$$

where the last equation is according to the performance difference lemma of Lemma 8. □

Lemma 8 (Performance difference lemma [9])**.** *For any policies π and π',*

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

C.3.1 An equivalent algorithm

For convenience of analysis, we introduce Algorithm 2, which is equivalent to Algorithm 1 as shown in Lemma 9.

```
Algorithm 2 Escort Policy Gradient Method with Parameter Normalization
  Input: Learning rate $\eta > 0$.
  Output: Policies $\pi_{\theta_t} = f_p(\theta_t)$.
  Initialize parameter $\theta_1(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.
  Normalize parameter $\tilde{\theta}_1(s,a) \leftarrow \frac{\theta_1(s,a)}{\|\theta_1(s,\cdot)\|_p}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.
  for $t = 1$ to $T$ do
    $\tilde{\zeta}_{t+1}(s,a) \leftarrow \tilde{\theta}_t(s,a) + \eta \cdot \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t(s,a)}$ for all $(s,a)$.
    $\tilde{\theta}_{t+1}(s,a) \leftarrow \frac{\tilde{\zeta}_{t+1}(s,a)}{\|\tilde{\zeta}_{t+1}(s,\cdot)\|_p}$ for all $(s,a)$.
  end for
```

Lemma 9. *Using the escort transform $\pi_\theta = f_p(\theta)$, Algorithm 2 with constant learning rate η and Algorithm 1 with learning rate $\eta_t(s) = \eta \cdot \|\theta_t(s,\cdot)\|_p^2$ are equivalent, i.e., for all (s,a),*

$$\begin{aligned} \tilde{\theta}_t(s,a) &= \frac{\theta_t(s,a)}{\|\theta_t(s,\cdot)\|_p}, \text{ and} \\ \pi_{\tilde{\theta}_t}(a|s) &= \pi_{\theta_t}(a|s). \end{aligned}$$

Proof. For $t = 1$, according to Algorithm 2, we have, for all (s,a), $\tilde{\theta}_1(s,a) = \frac{\theta_1(s,a)}{\|\theta_1(s,\cdot)\|_p}$, and,

$$\pi_{\tilde{\theta}_1}(a|s) = \frac{|\tilde{\theta}_1(s,a)|^p}{\sum_{a'}|\tilde{\theta}_1(s,a')|^p} = \frac{|\theta_1(s,a)|^p}{\sum_{a'}|\theta_1(s,a')|^p}\cdot\frac{1}{\cancel{\|\theta_1(s,\cdot)\|_p^p}}\cdot\cancel{\|\theta_1(s,\cdot)\|_p^p} = \pi_{\theta_1}(a|s). \tag{71}$$

Suppose $\tilde{\theta}_t(s,a) = \frac{\theta_t(s,a)}{\|\theta_t(s,\cdot)\|_p}$ for some $t \geq 1$. Using similar calculation as in Eq. (71), we have, for all (s,a), $\pi_{\tilde{\theta}_t}(a|s) = \pi_{\theta_t}(a|s)$, and,

$$\begin{aligned} \tilde{\zeta}_{t+1}(s,a) &\leftarrow \tilde{\theta}_t(s,a) + \eta\cdot\frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial\tilde{\theta}_t(s,a)} \qquad \text{(Algorithm 2)} \\ &= \frac{\theta_t(s,a)}{\|\theta_t(s,\cdot)\|_p} + \eta\cdot\|\theta_t(s,\cdot)\|_p\cdot\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial\theta_t(s,a)} \qquad \big(\text{induction hypothesis and } \pi_{\tilde{\theta}_t} = \pi_{\theta_t}\big) \\ &= \frac{\theta_t(s,a)}{\|\theta_t(s,\cdot)\|_p} + \eta_t(s)\cdot\frac{1}{\|\theta_t(s,\cdot)\|_p}\cdot\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial\theta_t(s,a)} \qquad \big(\eta_t(s) = \eta\cdot\|\theta_t(s,\cdot)\|_p^2\big) \\ &= \frac{1}{\|\theta_t(s,\cdot)\|_p}\cdot\left(\theta_t(s,a) + \eta_t(s)\cdot\frac{\partial V^{\pi_{\theta_t}}(\mu)}{\partial\theta_t(s,a)}\right) \\ &= \frac{1}{\|\theta_t(s,\cdot)\|_p}\cdot\theta_{t+1}(s,a). \qquad \text{(Algorithm 1)} \end{aligned} \tag{72}$$

Therefore we have,

$$\begin{aligned} \tilde{\theta}_{t+1}(s,a) &\leftarrow \frac{\tilde{\zeta}_{t+1}(s,a)}{\|\tilde{\zeta}_{t+1}(s,\cdot)\|_p} \qquad \text{(Algorithm 2)} \\ &= \frac{1}{\cancel{\|\theta_t(s,\cdot)\|_p}}\cdot\theta_{t+1}(s,a)\cdot\frac{\cancel{\|\theta_t(s,\cdot)\|_p}}{\|\theta_{t+1}(s,\cdot)\|_p} \qquad \text{(by Eq. (72))} \\ &= \frac{\theta_{t+1}(s,a)}{\|\theta_{t+1}(s,\cdot)\|_p}. \end{aligned}$$

Using similar calculation as in Eq. (71), we have, for all (s,a), $\pi_{\tilde{\theta}_{t+1}}(a|s) = \pi_{\theta_{t+1}}(a|s)$. □

Theorem 3. Following the escort policy gradient Algorithm 1 with any initialization such that $|\theta_1(s,a)| > 0$, $\forall (s,a)$, and $\eta_t(s) = \frac{(1-\gamma)^3 \cdot \|\theta_t(s,\cdot)\|_p^2}{10 \cdot p^2 \cdot A^{2/p}}$ to get $\{\theta_t\}_{t\geq 1}$, for all $t \geq 1$, the following sub-optimality upper bounds hold for π_{θ_t},

- for $p \geq 2$, we have,

$$V^*(\rho) - V^{\pi_{\theta_t}}(\rho) \leq \frac{20 \cdot A^{2/p} \cdot S}{c^{2-2/p} \cdot (1-\gamma)^6 \cdot t} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^2 \cdot \left\| \frac{1}{\mu} \right\|_\infty ;$$

- for $p = 1$, we have,

$$V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \leq \frac{20 \cdot A^2 \cdot S}{(1-\gamma)^6 \cdot t} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^2 \cdot \left\| \frac{1}{\mu} \right\|_\infty ,$$

where $c \coloneqq \inf_{s \in \mathcal{S}} \inf_{t \geq 1} \pi_{\theta_t}(a^*(s)|s) > 0$ is problem- and initialization-dependent constant, $A \coloneqq |\mathcal{A}|$ and $S \coloneqq |\mathcal{S}|$ are the total number of actions and states, respectively, and $\mu \in \Delta(\mathcal{S})$ is an initial state distribution which provides initial states for the policy gradient Algorithm 1.

Proof. Note that for any θ and μ,

$$\begin{aligned} d_\mu^{\pi_\theta}(s) &= \mathop{\mathbb{E}}_{s_0 \sim \mu} \left[d_\mu^{\pi_\theta}(s) \right] \\ &= \mathop{\mathbb{E}}_{s_0 \sim \mu} \left[(1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s | s_0, \pi_\theta, \mathcal{P}) \right] \\ &\geq \mathop{\mathbb{E}}_{s_0 \sim \mu} \left[(1-\gamma) \Pr(s_0 = s | s_0) \right] \\ &= (1-\gamma) \cdot \mu(s) . \end{aligned} \tag{73}$$

According to the value sub-optimality lemma of Lemma 10,

$$\begin{aligned} V^*(\rho) - V^{\pi_\theta}(\rho) &= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\theta}(s) \sum_a \left(\pi^*(a|s) - \pi_\theta(a|s) \right) \cdot Q^*(s,a) \\ &= \frac{1}{1-\gamma} \sum_s \frac{d_\rho^{\pi_\theta}(s)}{d_\mu^{\pi_\theta}(s)} \cdot d_\mu^{\pi_\theta}(s) \sum_a \left(\pi^*(a|s) - \pi_\theta(a|s) \right) \cdot Q^*(s,a) \\ &\leq \frac{1}{1-\gamma} \cdot \left\| \frac{1}{d_\mu^{\pi_\theta}} \right\|_\infty \sum_s d_\mu^{\pi_\theta}(s) \sum_a \left(\pi^*(a|s) - \pi_\theta(a|s) \right) \cdot Q^*(s,a) \\ &\leq \frac{1}{(1-\gamma)^2} \cdot \left\| \frac{1}{\mu} \right\|_\infty \sum_s d_\mu^{\pi_\theta}(s) \sum_a \left(\pi^*(a|s) - \pi_\theta(a|s) \right) \cdot Q^*(s,a) \qquad \left(\text{by Eq. (73) and } \min_s \mu(s) > 0 \right) \\ &= \frac{1}{1-\gamma} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot \left[V^*(\mu) - V^{\pi_\theta}(\mu) \right] , \end{aligned} \tag{74}$$

where the first inequality is because of

$$\sum_a \left(\pi^*(a|s) - \pi_\theta(a|s) \right) \cdot Q^*(s,a) \geq 0,$$

and the last equation is again by Lemma 10.

For $p \geq 2$ and $p = 1$, according to Lemma 6, $V^{\pi_\theta}(\mu)$ is β-smooth with $\beta = \frac{8 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \frac{1}{\min_s \|\theta_\xi(s,\cdot)\|_p^2}$, i.e., we have, in Algorithm 2,

$$\left| V^{\pi_{\tilde{\zeta}_{t+1}}}(\mu) - V^{\pi_{\tilde{\theta}_t}}(\mu) - \left\langle \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \tilde{\zeta}_{t+1} - \tilde{\theta}_t \right\rangle \right| \leq \frac{4 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \frac{\|\tilde{\zeta}_{t+1} - \tilde{\theta}_t\|_2^2}{\min_s \|\tilde{\theta}_{t,\xi}(s,\cdot)\|_p^2}, \tag{75}$$

where

$$\begin{aligned} \tilde{\theta}_{t,\xi} &\coloneqq \tilde{\theta}_t + \xi \cdot (\tilde{\zeta}_{t+1} - \tilde{\theta}_t) \\ &= \tilde{\theta}_t + \xi \cdot \eta \cdot \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \qquad \text{(Algorithm 2)} \end{aligned}$$

for some $\xi \in [0, 1]$. Denote $s_\xi \coloneqq \arg\min_s \|\tilde{\theta}_{t,\xi}(s, \cdot)\|_p^2$. We have,

$$\begin{aligned}
\|\tilde{\theta}_{t,\xi}(s_\xi, \cdot)\|_p &\ge \|\tilde{\theta}_t(s_\xi, \cdot)\|_p - \xi \cdot \eta \cdot \left\| \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t(s_\xi, \cdot)} \right\|_p \qquad \text{(triangle inequality)} \\
&\ge \min_s \|\tilde{\theta}_t(s, \cdot)\|_p - \xi \cdot \eta \cdot \left\| \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t(s_\xi, \cdot)} \right\|_p . \qquad (76)
\end{aligned}$$

The ℓ_p gradient norm can be upper bounded as,

$$\begin{aligned}
\left\| \frac{\partial V^{\pi_\theta}(\mu)}{\partial \theta(s, \cdot)} \right\|_p &= \left[\sum_a \left| \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot p \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)} \cdot A^{\pi_\theta}(s,a) \right|^p \right]^{\frac{1}{p}} \qquad \text{(by Lemma 5)} \\
&\le \frac{p}{1-\gamma} \cdot \left[\sum_a \left| \frac{\pi_\theta(a|s)}{\theta(s,a)} \cdot A^{\pi_\theta}(s,a) \right|^p \right]^{\frac{1}{p}} \\
&= \frac{p}{1-\gamma} \cdot \frac{1}{\|\theta(s, \cdot)\|_p} \cdot \left[\sum_a \left(\pi_\theta(a|s)^{1-1/p} \cdot |A^{\pi_\theta}(s,a)| \right)^p \right]^{\frac{1}{p}} \\
&\le \frac{p}{1-\gamma} \cdot \frac{1}{\|\theta(s, \cdot)\|_p} \cdot \left[\sum_a \left(1 \cdot \frac{1}{1-\gamma} \right)^p \right]^{\frac{1}{p}} \\
&\le \frac{p \cdot A^{1/p}}{(1-\gamma)^2} \cdot \max_s \frac{1}{\|\theta(s, \cdot)\|_p}. \qquad (77)
\end{aligned}$$

Combining Eqs. (76) and (77), we have,

$$\begin{aligned}
\min_s \|\tilde{\theta}_{t,\xi}(s, \cdot)\|_p &\ge \min_s \|\tilde{\theta}_t(s, \cdot)\|_p - \xi \cdot \eta \cdot \frac{p \cdot A^{1/p}}{(1-\gamma)^2} \cdot \frac{1}{\min_s \|\tilde{\theta}_t(s, \cdot)\|_p} \\
&= 1 - \xi \cdot \eta \cdot \frac{p \cdot A^{1/p}}{(1-\gamma)^2}. \qquad \left(\|\tilde{\theta}_t(s, \cdot)\|_p = 1, \text{ for all } s, \text{ Algorithm 2} \right) \\
&= 1 - \xi \cdot \frac{1-\gamma}{10 \cdot p \cdot A^{1/p}} \qquad \left(\eta = \frac{(1-\gamma)^3}{10 \cdot p^2 \cdot A^{2/p}}, \text{ by Lemma 9} \right) \\
&\ge 1 - \frac{1-\gamma}{10 \cdot p \cdot A^{1/p}} \qquad (\xi \in [0, 1]) \\
&= \left(1 - \frac{2}{\sqrt{5}} \right) \cdot \left(1 - \frac{5 + 2\sqrt{5}}{10} \cdot \frac{1-\gamma}{p \cdot A^{1/p}} \right) + \frac{2}{\sqrt{5}} \\
&\ge \frac{2}{\sqrt{5}}. \qquad \left(p \ge 2, \ A^{1/p} \ge 1, \ 1-\gamma \in (0, 1] \right) \qquad (78)
\end{aligned}$$

Combining Eqs. (75) and (78), we have,

$$\begin{aligned}
\left| V^{\pi_{\tilde{\zeta}_{t+1}}}(\mu) - V^{\pi_{\tilde{\theta}_t}}(\mu) - \left\langle \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \tilde{\zeta}_{t+1} - \tilde{\theta}_t \right\rangle \right| &\le \frac{4 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \frac{\|\tilde{\zeta}_{t+1} - \tilde{\theta}_t\|_2^2}{\min_s \|\tilde{\theta}_{t,\xi}(s, \cdot)\|_p^2} \\
&\le \frac{5 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \|\tilde{\zeta}_{t+1} - \tilde{\theta}_t\|_2^2, \qquad (79)
\end{aligned}$$

which implies,

$$
\begin{aligned}
V^{\pi_{\tilde{\theta}_t}}(\mu) - V^{\pi_{\tilde{\theta}_{t+1}}}(\mu) &= V^{\pi_{\tilde{\theta}_t}}(\mu) - V^{\pi_{\tilde{\zeta}_{t+1}}}(\mu) \qquad \left(\tilde{\theta}_{t+1}(s,a) = \frac{\tilde{\zeta}_{t+1}(s,a)}{\| \tilde{\zeta}_{t+1}(s,\cdot) \|_p}, \text{ Algorithm 2} \right) \\
&\le - \Big\langle \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \tilde{\zeta}_{t+1} - \tilde{\theta}_t \Big\rangle + \frac{5 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \| \tilde{\zeta}_{t+1} - \tilde{\theta}_t \|_2^2 \qquad \text{(Eq. (79))} \\
&= -\eta \cdot \left\| \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t} \right\|_2^2 + \frac{5 \cdot p^2 \cdot A^{2/p}}{(1-\gamma)^3} \cdot \eta^2 \cdot \left\| \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t} \right\|_2^2 \qquad \left(\tilde{\zeta}_{t+1} = \tilde{\theta}_t + \eta \cdot \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \text{ Algorithm 2} \right) \\
&= - \frac{(1-\gamma)^3}{20 \cdot p^2 \cdot A^{2/p}} \cdot \left\| \frac{\partial V^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t} \right\|_2^2 \qquad \left(\eta = \frac{(1-\gamma)^3}{10 \cdot p^2 \cdot A^{2/p}} \right) \\
&\le - \frac{(1-\gamma)^3}{20 \cdot \cancel{p^2} \cdot A^{2/p}} \cdot \left[\frac{\cancel{p}}{\sqrt{S}} \cdot \left\| \frac{d_\mu^{\pi^*}}{d_\mu^{\pi_{\tilde{\theta}_t}}} \right\|_\infty^{-1} \cdot \frac{\min_s \pi_{\tilde{\theta}_t}(a^*(s)|s)^{1-1/p}}{\max_s \| \tilde{\theta}_t(s,\cdot) \|_p} \cdot \left[V^*(\mu) - V^{\pi_{\tilde{\theta}_t}}(\mu) \right] \right]^2 \qquad \text{(Lemma 7)} \\
&= - \frac{(1-\gamma)^3}{20 \cdot A^{2/p} \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{d_\mu^{\pi_{\tilde{\theta}_t}}} \right\|_\infty^{-2} \cdot \min_s \pi_{\tilde{\theta}_t}(a^*(s)|s)^{2-2/p} \cdot \left[V^*(\mu) - V^{\pi_{\tilde{\theta}_t}}(\mu) \right]^2 \qquad \left(\| \tilde{\theta}_t(s,\cdot) \|_p = 1, \text{ for all } s \right) \\
&\le - \frac{(1-\gamma)^5}{20 \cdot A^{2/p} \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-2} \cdot \min_s \pi_{\tilde{\theta}_t}(a^*(s)|s)^{2-2/p} \cdot \left[V^*(\mu) - V^{\pi_{\tilde{\theta}_t}}(\mu) \right]^2, \qquad (80)
\end{aligned}
$$

where the last inequality is by Eq. (73). Then we have,

$$
\begin{aligned}
V^{\pi_{\theta_t}}(\mu) - V^{\pi_{\theta_{t+1}}}(\mu) &= V^{\pi_{\tilde{\theta}_t}}(\mu) - V^{\pi_{\tilde{\theta}_{t+1}}}(\mu) \qquad \text{(by Lemma 9)} \\
&\le - \frac{(1-\gamma)^5}{20 \cdot A^{2/p} \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-2} \cdot \min_s \pi_{\theta_t}(a^*(s)|s)^{2-2/p} \cdot \left[V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \right]^2, \qquad \text{(by Eq. (80) and Lemma 9)} \\
&\le - \frac{(1-\gamma)^5}{20 \cdot A^{2/p} \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-2} \cdot c^{2-2/p} \cdot \left[V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \right]^2,
\end{aligned}
$$

which is equivalent to,

$$
\delta_{t+1} - \delta_t \le - \frac{(1-\gamma)^5}{20 \cdot A^{2/p} \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-2} \cdot c^{2-2/p} \cdot \delta_t^2,
$$

where $\delta_t = V^*(\mu) - V^{\pi_{\theta_t}}(\mu)$. Using the similar induction argument as in Eq. (37), we have,

$$
V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \le \frac{20 \cdot A^{2/p} \cdot S}{(1-\gamma)^5 \cdot t} \cdot \frac{1}{c^{2-2/p}} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^2,
$$

which leads to the final result,

$$
\begin{aligned}
V^*(\rho) - V^{\pi_{\theta_t}}(\rho) &\le \frac{1}{1-\gamma} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot \left[V^*(\mu) - V^{\pi_{\theta_t}}(\mu) \right] \qquad \text{(by Eq. (74))} \\
&\le \frac{20 \cdot A^{2/p} \cdot S}{c^{2-2/p} \cdot (1-\gamma)^6 \cdot t} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^2 \cdot \left\| \frac{1}{\mu} \right\|_\infty.
\end{aligned}
$$

□

Lemma 10 (Value sub-optimality lemma [14]). *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. See the proof in [14, Lemma 21]. □

C.4 Entropy Regularized MDPs

The objective for the entropy regularized policy gradient method is,

$$\tilde{V}^{\pi}(\rho) \coloneqq V^{\pi}(\rho) + \tau \cdot \mathbb{H}(\rho, \pi), \tag{81}$$

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

$$\mathbb{H}(\rho, \pi) \coloneqq \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, a_t \sim \pi(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t|s_t) \right], \tag{82}$$

and $\tau \geq 0$ is "temperature" of the regularization.

Lemma 11. *The entropy regularized escort policy gradient w.r.t. θ is*

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

$$\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} = \frac{1}{1-\gamma} \cdot d_{\mu}^{\pi_\theta}(s) \cdot p \cdot \textit{diag}\left(\frac{\pi_\theta(\cdot|s)}{\theta(s,\cdot)} \right) \left(\mathbf{Id} - \mathbf{1}\pi_\theta(\cdot|s)^\top \right) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(\cdot|s) \right].$$

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

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

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

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

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

Taking derivative w.r.t. θ,

$$\begin{aligned}
\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} &= \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) \right] \\
&\quad + \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \pi_\theta(a|s) \cdot \left[\frac{\partial \tilde{Q}^{\pi_\theta}(s,a)}{\partial \theta} - \tau \frac{1}{\pi_\theta(a|s)} \frac{\partial \pi_\theta(a|s)}{\partial \theta} \right] \\
&= \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) \right] + \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \pi_\theta(a|s) \cdot \frac{\partial \tilde{Q}^{\pi_\theta}(s,a)}{\partial \theta} \\
&= \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) \right] \\
&\quad + \gamma \cdot \mathop{\mathbb{E}}_{s\sim\mu} \sum_a \pi_\theta(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \cdot \frac{\partial \tilde{V}^{\pi_\theta}(s')}{\partial \theta} \\
&= \frac{1}{1-\gamma} \sum_s d_{\mu}^{\pi_\theta}(s) \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) \right],
\end{aligned}$$

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 5, i.e., for $s' \neq s$, $\frac{\partial \pi_\theta(a|s)}{\partial \theta(s',\cdot)} = \mathbf{0}$,

$$\begin{aligned}
\frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} &= \frac{1}{1-\gamma} \cdot d_{\mu}^{\pi_\theta}(s) \cdot \left[\sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta(s,\cdot)} \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) \right] \right] \\
&= \frac{1}{1-\gamma} \cdot d_{\mu}^{\pi_\theta}(s) \cdot \left(\frac{d\pi_\theta(\cdot|s)}{d\theta(s,\cdot)} \right)^\top \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(\cdot|s) \right] \\
&= \frac{1}{1-\gamma} \cdot d_{\mu}^{\pi_\theta}(s) \cdot p \cdot \operatorname{diag}\left(\frac{\pi_\theta(\cdot|s)}{\theta(s,\cdot)} \right) \left(\mathbf{Id} - \mathbf{1}\pi_\theta(\cdot|s)^\top \right) \left[\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(\cdot|s) \right].
\end{aligned}$$

For each component a, we have

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

Lemma 12 (Non-uniform Łojasiewicz). *Suppose $\mu(s) > 0$ for all $s \in \mathcal{S}$ and $\pi_\theta = f_p(\theta)$. Then,*

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

Proof. According to the definition of soft value functions,

$$\begin{aligned}\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) &= \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, a_t \sim \pi_\tau^*(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t,a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t (r(s_t,a_t) - \tau \log \pi_\tau^*(a_t|s_t)) \right] - \tilde{V}^{\pi_\theta}(\rho) \\ &= \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, a_t \sim \pi_\tau^*(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t,a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t (r(s_t,a_t) - \tau \log \pi_\tau^*(a_t|s_t) + \tilde{V}^{\pi_\theta}(s_t) - \tilde{V}^{\pi_\theta}(s_t)) \right] - \tilde{V}^{\pi_\theta}(\rho) \\ &= \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, a_t \sim \pi_\tau^*(\cdot|s_t), \\ s_{t+1} \sim \mathcal{P}(\cdot|s_t,a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t (r(s_t,a_t) - \tau \log \pi_\tau^*(a_t|s_t) + \gamma \tilde{V}^{\pi_\theta}(s_{t+1}) - \tilde{V}^{\pi_\theta}(s_t)) \right] \\ &= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \left[\sum_a \pi_\tau^*(a|s) \cdot \left(r(s,a) - \tau \log \pi_\tau^*(a|s) + \gamma \sum_{s'} \mathcal{P}(s'|s,a) \tilde{V}^{\pi_\theta}(s') - \tilde{V}^{\pi_\theta}(s) \right) \right] \\ &= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \left[\sum_a \pi_\tau^*(a|s) \cdot \Big[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\tau^*(a|s) \Big] - \tilde{V}^{\pi_\theta}(s) \right]. \qquad (83)\end{aligned}$$

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

$$\bar{\pi}_\theta(a|s) = \frac{\exp\left\{ \tilde{Q}^{\pi_\theta}(s,a)/\tau \right\}}{\sum_{a'} \exp\left\{ \tilde{Q}^{\pi_\theta}(s,a')/\tau \right\}}, \ \forall a. \qquad (84)$$

We have, $\forall s$,

$$\begin{aligned}\sum_a \pi_\tau^*(a|s) \cdot \Big[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\tau^*(a|s) \Big] &\leq \max_{\pi(\cdot|s)} \sum_a \pi(a|s) \cdot \Big[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi(a|s) \Big] \\ &= \sum_a \bar{\pi}_\theta(a|s) \cdot \Big[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \bar{\pi}_\theta(a|s) \Big] \\ &= \tau \log \sum_a \exp\left\{ \tilde{Q}^{\pi_\theta}(s,a)/\tau \right\}. \qquad (85)\end{aligned}$$

Also note that,

$$
\begin{aligned}
\tilde{V}^{\pi_\theta}(s) &= \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\theta(a|s) \right] \\
&= \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \bar{\pi}_\theta(a|s) + \tau \log \bar{\pi}_\theta(a|s) - \tau \log \pi_\theta(a|s) \right] \\
&= \sum_a \pi_\theta(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \bar{\pi}_\theta(a|s) \right] - \tau D_{\mathrm{KL}}(\pi_\theta(\cdot|s) \| \bar{\pi}_\theta(\cdot|s)) \\
&= \tau \log \sum_a \exp\left\{ \tilde{Q}^{\pi_\theta}(s,a)/\tau \right\} - \tau \cdot D_{\mathrm{KL}}(\pi_\theta(\cdot|s) \| \bar{\pi}_\theta(\cdot|s)). \qquad (86)
\end{aligned}
$$

Combining Eq. (83), Eqs. (85) and (86), we have,

$$
\begin{aligned}
\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) &= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \left[\sum_a \pi_\tau^*(a|s) \cdot \left[\tilde{Q}^{\pi_\theta}(s,a) - \tau \log \pi_\tau^*(a|s) \right] - \tilde{V}^{\pi_\theta}(s) \right] \\
&\le \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \left[\tau \log \sum_a \exp\left\{ \tilde{Q}^{\pi_\theta}(s,a)/\tau \right\} - \tilde{V}^{\pi_\theta}(s) \right] \\
&= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_\theta(\cdot|s) \| \bar{\pi}_\theta(\cdot|s)) \\
&\le \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \frac{\tau}{2} \cdot \left\| \frac{\tilde{Q}^{\pi_\theta}(s,\cdot)}{\tau} - \log \pi_\theta(s,\cdot) - \frac{c_\theta(s)}{\tau} \cdot \mathbf{1} \right\|_\infty^2 \qquad \text{(by Lemma 13)} \\
&= \frac{1}{1-\gamma} \sum_s d_\rho^{\pi_\tau^*}(s) \cdot \frac{1}{2\tau} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(s,\cdot) - c_\theta(s) \cdot \mathbf{1} \right\|_\infty^2, \qquad (87)
\end{aligned}
$$

where $c_\theta(s) = \frac{\left(\tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(s,\cdot) \right)^\top \mathbf{1}}{A}$. Taking square root of Eq. (87), we have,

$$
\begin{aligned}
\left[\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \right]^{\frac{1}{2}} &\le \frac{1}{\sqrt{1-\gamma}} \cdot \left[\sum_s d_\rho^{\pi_\tau^*}(s) \cdot \frac{1}{2\tau} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(s,\cdot) - c_\theta(s) \cdot \mathbf{1} \right\|_\infty^2 \right]^{\frac{1}{2}} \\
&= \frac{1}{\sqrt{1-\gamma}} \cdot \left[\sum_s \left(\sqrt{d_\rho^{\pi_\tau^*}(s)} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(s,\cdot) - c_\theta(s) \cdot \mathbf{1} \right\|_\infty \right)^2 \right]^{\frac{1}{2}} \\
&\le \frac{1}{\sqrt{1-\gamma}} \cdot \sum_s \sqrt{d_\rho^{\pi_\tau^*}(s)} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(s,\cdot) - c_\theta(s) \cdot \mathbf{1} \right\|_\infty \qquad (\|x\|_2 \le \|x\|_1) \\
&\le \frac{1}{\sqrt{1-\gamma}} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{\frac{1}{2}} \sum_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \left\| \tilde{Q}^{\pi_\theta}(s,\cdot) - \tau \log \pi_\theta(s,\cdot) - c_\theta(s) \cdot \mathbf{1} \right\|_\infty. \\
&\qquad (88)
\end{aligned}
$$

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

$$\begin{aligned}
\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 &= \left[\sum_s \left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s, \cdot)} \right\|_2^2 \right]^{\frac{1}{2}} \\
&\geq \frac{1}{\sqrt{S}} \sum_s \left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s, \cdot)} \right\|_2, \qquad \text{(by Cauchy-Schwarz , } \|x\|_1 = |\langle \mathbf{1}, \ |x| \rangle| \leq \|\mathbf{1}\|_2 \cdot \|x\|_2\text{)} \\
&\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \\
&\quad \cdot \left\| p \cdot \text{diag}\left(\frac{\pi_\theta(\cdot|s)}{\theta(s, \cdot)} \right) \left(\mathbf{Id} - \mathbf{1} \pi_\theta(\cdot|s)^\top \right) \left[\tilde{Q}^{\pi_\theta}(s, \cdot) - \tau \log \pi_\theta(s, \cdot) - c_\theta(s) \cdot \mathbf{1} \right] \right\|_2 \qquad \text{(by Lemma 11)} \\
&= \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \frac{p}{\|\theta(s, \cdot)\|_p} \\
&\quad \cdot \left\| \text{diag}\left(\pi_\theta(\cdot|s)^{1-1/p} \right) \left(\mathbf{Id} - \mathbf{1} \pi_\theta(\cdot|s)^\top \right) \left[\tilde{Q}^{\pi_\theta}(s, \cdot) - \tau \log \pi_\theta(s, \cdot) - c_\theta(s) \cdot \mathbf{1} \right] \right\|_2 \\
&\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \frac{p}{\|\theta(s, \cdot)\|_p} \cdot \min_a \pi_\theta(a|s)^{1-1/p} \cdot \left\| \tilde{Q}^{\pi_\theta}(s, \cdot) - \tau \log \pi_\theta(s, \cdot) - c_\theta(s) \cdot \mathbf{1} \right\|_2 . \quad \text{(by Lemma 14)}
\end{aligned}$$

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

$$\begin{aligned}
\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta} \right\|_2 &\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \frac{p}{\|\theta(s, \cdot)\|_p} \cdot \min_a \pi_\theta(a|s)^{1-1/p} \cdot \|\zeta_\theta(s)\|_2 \\
&\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{1-\gamma} \sum_s d_\mu^{\pi_\theta}(s) \cdot \frac{p}{\|\theta(s, \cdot)\|_p} \cdot \min_a \pi_\theta(a|s)^{1-1/p} \cdot \|\zeta_\theta(s)\|_\infty \\
&\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{\sqrt{1-\gamma}} \cdot \min_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \min_s \frac{p}{\|\theta(s, \cdot)\|_p} \cdot \min_{s,a} \pi_\theta(a|s)^{1-1/p} \cdot \sqrt{2\tau} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}} \\
&\quad \cdot \left[\frac{1}{\sqrt{1-\gamma}} \cdot \frac{1}{\sqrt{2\tau}} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{\frac{1}{2}} \sum_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \|\zeta_\theta(s)\|_\infty \right] \\
&\geq \frac{1}{\sqrt{S}} \cdot \frac{1}{\sqrt{1-\gamma}} \cdot \min_s \sqrt{d_\mu^{\pi_\theta}(s)} \cdot \min_s \frac{p}{\|\theta(s, \cdot)\|_p} \cdot \min_{s,a} \pi_\theta(a|s)^{1-1/p} \cdot \sqrt{2\tau} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}} \\
&\quad \cdot \left[\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \right]^{\frac{1}{2}} \qquad \text{(by Eq. (88))} \\
&\geq \frac{\sqrt{2\tau}}{\sqrt{S}} \cdot \min_s \sqrt{\mu(s)} \cdot \min_s \frac{p}{\|\theta(s, \cdot)\|_p} \cdot \min_{s,a} \pi_\theta(a|s)^{1-1/p} \cdot \left\| \frac{d_\rho^{\pi_\tau^*}}{d_\mu^{\pi_\theta}} \right\|_\infty^{-\frac{1}{2}} \cdot \left[\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_\theta}(\rho) \right]^{\frac{1}{2}},
\end{aligned}$$

where the last inequality is by $d_\mu^{\pi_\theta}(s) \geq (1-\gamma) \cdot \mu(s)$ (cf. Eq. (73)). Note that $\min_{s,a} \pi_\theta(a|s)^{1-1/p} \geq \min_{s,a} \pi_\theta(a|s)$, which is a better dependence than [14, Lemma 15]. □

Lemma 13 (KL-Logit inequality [14])**.** *Let $\pi_\theta = \mathrm{softmax}(\theta)$ and $\pi_{\theta'} = \mathrm{softmax}(\theta')$. Then for any constant $c \in \mathbb{R}$,*

$$D_{\mathrm{KL}}(\pi_\theta \| \pi_{\theta'}) \leq \frac{1}{2} \cdot \|\theta' - \theta - c \cdot \mathbf{1}\|_\infty^2 .$$

Proof. See the proof in [14, Lemma 27]. □

Lemma 14. *Let $\pi \in \Delta(\mathcal{A})$ and $q \geq 0$. For any vector $x \in \mathbb{R}^K$, we have,*

$$\left\| diag(q) \left(\mathbf{Id} - \mathbf{1}\pi^\top \right) \left(x - \frac{x^\top \mathbf{1}}{K} \cdot \mathbf{1} \right) \right\|_2 \geq \min_a q(a) \cdot \left\| x - \frac{x^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_2 .$$

Proof. Denote $G = G(\pi, q) = \text{diag}(q)\left(\mathbf{Id} - \mathbf{1}\pi^\top\right) \in \mathbb{R}^{K\times K}$. Denote the eigenvalues of $G^\top G$ as

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

First, we show that $\lambda_1 = 0$.

$$\begin{aligned} G^\top G\mathbf{1} &= G^\top \text{diag}(q)\left(\mathbf{Id} - \mathbf{1}\pi^\top\right)\mathbf{1} \\ &= G^\top \text{diag}(q)\left(\mathbf{1} - \mathbf{1}\right) = 0 \cdot \mathbf{1}, \end{aligned}$$

which means $\mathbf{1}$ is an eigenvector of $G^\top G$ with eigenvalue 0. And for any vector $x \in \mathbb{R}^K$, we have,

$$x^\top G^\top G x = \|Gx\|_2^2 \ge 0,$$

which means $G^\top G$ is semi-positive definite. Therefore $\lambda_1 = 0$.

Second, for any vector $x \in \mathbb{R}^K$, x can be written as linear combination of eigenvectors of $G^\top G$,

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

Since $G^\top G$ is symmetric, $\left\{\frac{\mathbf{1}}{\sqrt{K}}, v_2, \ldots, v_K\right\}$ are orthonormal. The last equation is because the representation is unique, and

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

Denote

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

We have,

$$\|x'\|_2^2 = a_2^2 + \cdots + a_K^2. \tag{89}$$

Since $v_2, \ldots, v_K$ are eigenvectors of $G^\top G$,

$$G^\top G x' = a_2 \cdot \lambda_2 \cdot v_2 + \cdots + a_K \cdot \lambda_K \cdot v_K.$$

Therefore we have,

$$\begin{aligned} \|Gx'\|_2 = \left(\|Gx'\|_2^2\right)^{\frac{1}{2}} &= \left({x'}^\top G^\top G x'\right)^{\frac{1}{2}} \\ &= \left(a_2^2 \cdot \lambda_2 + \cdots + a_K^2 \cdot \lambda_K\right)^{\frac{1}{2}} \\ &\ge \left(a_2^2 \cdot \lambda_2 + \cdots + a_K^2 \cdot \lambda_2\right)^{\frac{1}{2}} \\ &= \sqrt{\lambda_2} \cdot \|x'\|_2. \qquad \text{(by Eq. (89))} \end{aligned} \tag{90}$$

Next, we have,

$$\begin{aligned} \lambda_2 = \frac{v_2^\top G^\top G v_2}{v_2^\top v_2} = \frac{1}{v_2^\top v_2} \cdot \|Gv_2\|_2^2 &= \frac{1}{v_2^\top v_2} \cdot \left\|\text{diag}(q)\left(v_2 - \pi^\top v_2 \cdot \mathbf{1}\right)\right\|_2^2 \\ &= \frac{1}{v_2^\top v_2} \cdot \left[\sum_{a=1}^K q(a)^2 \cdot \left(v_2(a) - \pi^\top v_2\right)^2\right] \\ &\ge \frac{1}{v_2^\top v_2} \cdot \min_a q(a)^2 \cdot \left\|v_2 - \pi^\top v_2 \cdot \mathbf{1}\right\|_2^2 \\ &= \min_a q(a)^2 \cdot \frac{v_2^\top v_2 + K \cdot (\pi^\top v_2)^2}{v_2^\top v_2} \qquad \left(v_2^\top \mathbf{1} = 0\right) \\ &\ge \min_a q(a)^2. \end{aligned} \tag{91}$$

Combining Eqs. (90) and (91), we have,

$$\begin{aligned}\left\| \mathrm{diag}(q) \left(\mathbf{Id} - \mathbf{1}\pi^\top \right) \left(x - \frac{x^\top \mathbf{1}}{K} \cdot \mathbf{1} \right) \right\|_2 &= \| G x' \|_2 \\ &\ge \sqrt{\lambda_2} \cdot \| x' \|_2 \\ &\ge \min_a q(a) \cdot \left\| x - \frac{x^\top \mathbf{1}}{K} \cdot \mathbf{1} \right\|_2 . \qquad \square\end{aligned}$$

Theorem 4. For an entropy regularized MDP with finite states and actions, following the escort policy gradient with any initialization such that $|\theta_1(s,a)| > 0, \forall (s,a)$, and $\eta_t = (1-\gamma)^3 / (10 \cdot p^2 \cdot A^{1/p} + c_\tau)$ to get $\{\theta_t\}_{t \ge 1}$, for all $t \ge 1$, the following sub-optimality upper bounds hold for π_{θ_t}:

$$\text{for } p \ge 2, \qquad \tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) \le \frac{\| 1/\mu \|_\infty}{\exp\{ C_\tau \cdot c'^2 \cdot t \}} \cdot \frac{1 + \tau \log A}{(1-\gamma)^2} , \tag{92}$$

where $c' > c := \inf_{(s,a)} \inf_{t \ge 1} \pi_{\theta_t}(a|s) > 0$, τ is the temperature for entropy regularization, π_τ^* is the softmax optimal policy, and c_τ, C_τ are problem-dependent constants.

Proof. According to the soft sub-optimality lemma of Lemma 15,

$$\begin{aligned}\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) &= \frac{1}{1-\gamma} \sum_s \left[d_\rho^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s) \| \pi_\tau^*(\cdot|s)) \right] \\ &= \frac{1}{1-\gamma} \sum_s \frac{d_\rho^{\pi_{\theta_t}}(s)}{d_\mu^{\pi_{\theta_t}}(s)} \cdot \left[d_\mu^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s) \| \pi_\tau^*(\cdot|s)) \right] \\ &\le \frac{1}{(1-\gamma)^2} \sum_s \frac{1}{\mu(s)} \cdot \left[d_\mu^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s) \| \pi_\tau^*(\cdot|s)) \right] \\ &\le \frac{1}{(1-\gamma)^2} \cdot \left\| \frac{1}{\mu} \right\|_\infty \sum_s \left[d_\mu^{\pi_{\theta_t}}(s) \cdot \tau \cdot D_{\mathrm{KL}}(\pi_{\theta_t}(\cdot|s) \| \pi_\tau^*(\cdot|s)) \right] \\ &= \frac{1}{1-\gamma} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot \left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu) \right] ,\end{aligned} \tag{93}$$

where the last equation is again by Lemma 15, and the first inequality is according to $d_\mu^{\pi_{\theta_t}}(s) \ge (1-\gamma) \cdot \mu(s)$ (cf. Eq. (73)).

According to Lemma 9, using $\frac{\partial \tilde{V}^{\pi_{\theta_t}}(\mu)}{\partial \theta_t}$ in Algorithm 1 with learning rate $\eta_t(s) = \eta \cdot \| \theta_t(s, \cdot) \|_p^2$ is equivalent to using $\frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}$ in Algorithm 2 with learning rate η. We have, in Algorithm 2,

$$\left| \tilde{V}^{\pi_{\tilde{\zeta}_{t+1}}}(\mu) - \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu) - \left\langle \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \tilde{\zeta}_{t+1} - \tilde{\theta}_t \right\rangle \right| \le \frac{4 \cdot p^2 \cdot A^{2/p} + c_\tau}{(1-\gamma)^3} \cdot \frac{\| \tilde{\zeta}_{t+1} - \tilde{\theta}_t \|_2^2}{\min_s \| \tilde{\theta}_{t,\xi}(s, \cdot) \|_p^2}, \tag{94}$$

where

$$\begin{aligned}\tilde{\theta}_{t,\xi} &:= \tilde{\theta}_t + \xi \cdot (\tilde{\zeta}_{t+1} - \tilde{\theta}_t) \\ &= \tilde{\theta}_t + \xi \cdot \eta \cdot \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \qquad \text{(Algorithm 2)}\end{aligned}$$

for some $\xi \in [0,1]$. Denote $s_\xi := \arg\min_s \| \tilde{\theta}_{t,\xi}(s, \cdot) \|_p^2$. We have,

$$\begin{aligned}\| \tilde{\theta}_{t,\xi}(s_\xi, \cdot) \|_p &\ge \| \tilde{\theta}_t(s_\xi, \cdot) \|_p - \xi \cdot \eta \cdot \left\| \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t(s_\xi, \cdot)} \right\|_p \qquad \text{(by triangle inequality)} \\ &\ge \min_s \| \tilde{\theta}_t(s, \cdot) \|_p - \xi \cdot \eta \cdot \left\| \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t(s_\xi, \cdot)} \right\|_p .\end{aligned} \tag{95}$$

The ℓ_p gradient norm can be upper bounded as,

$$\begin{aligned}
\left\| \frac{\partial \tilde{V}^{\pi_\theta}(\mu)}{\partial \theta(s,\cdot)} \right\|_p &= \left[\sum_a \left| \frac{1}{1-\gamma} \cdot d_\mu^{\pi_\theta}(s) \cdot p \cdot \frac{\pi_\theta(a|s)}{\theta(s,a)} \cdot \tilde{A}^{\pi_\theta}(s,a) \right|^p \right]^{\frac{1}{p}} && \text{(by Lemma 11)} \\
&\le \frac{p}{1-\gamma} \cdot \left[\sum_a \left| \frac{\pi_\theta(a|s)}{\theta(s,a)} \cdot \tilde{A}^{\pi_\theta}(s,a) \right|^p \right]^{\frac{1}{p}} \\
&= \frac{p}{1-\gamma} \cdot \frac{1}{\|\theta(s,\cdot)\|_p} \cdot \left[\sum_a \left(\pi_\theta(a|s)^{1-1/p} \cdot |\tilde{A}^{\pi_\theta}(s,a)| \right)^p \right]^{\frac{1}{p}} \\
&\le \frac{p}{1-\gamma} \cdot \frac{1}{\|\theta(s,\cdot)\|_p} \cdot \left[\sum_a \left(1 \cdot \frac{1+\tau \log A}{1-\gamma} \right)^p \right]^{\frac{1}{p}} \\
&\le \frac{p \cdot A^{1/p} \cdot (1+\tau \log A)}{(1-\gamma)^2} \cdot \max_s \frac{1}{\|\theta(s,\cdot)\|_p}. && (96)
\end{aligned}$$

Combining Eqs. (95) and (96), we have,

$$\begin{aligned}
\min_s \|\tilde{\theta}_{t,\xi}(s,\cdot)\|_p &\ge \min_s \|\tilde{\theta}_t(s,\cdot)\|_p - \xi \cdot \eta \cdot \frac{p \cdot A^{1/p} \cdot (1+\tau \log A)}{(1-\gamma)^2} \cdot \frac{1}{\min_s \|\tilde{\theta}_t(s,\cdot)\|_p} \\
&= 1 - \xi \cdot \eta \cdot \frac{p \cdot A^{1/p} \cdot (1+\tau \log A)}{(1-\gamma)^2} \qquad \left(\|\tilde{\theta}_t(s,\cdot)\|_p = 1, \text{ for all } s, \text{ Algorithm 2} \right) \qquad (97)
\end{aligned}$$

Note that $\eta = \frac{(1-\gamma)^3}{10 \cdot p^2 \cdot A^{2/p} \cdot (1+\tau \log A)}$. We have,

$$\begin{aligned}
\min_s \|\tilde{\theta}_{t,\xi}(s,\cdot)\|_p &\ge 1 - \xi \cdot \eta \cdot \frac{p \cdot A^{1/p} \cdot (1+\tau \log A)}{(1-\gamma)^2} && \text{(by Eq. (97))} \\
&= 1 - \xi \cdot \frac{(1-\gamma)^3}{10 \cdot p^2 \cdot A^{2/p}} \cdot \frac{p \cdot A^{1/p}}{(1-\gamma)^2} \\
&\ge 1 - \frac{1-\gamma}{10 \cdot p \cdot A^{1/p}} \\
&= \left(1 - \frac{2}{\sqrt{5}} \right) \cdot \left(1 - \frac{5+2\sqrt{5}}{10} \cdot \frac{1-\gamma}{p \cdot A^{1/p}} \right) + \frac{2}{\sqrt{5}} \\
&\ge \frac{2}{\sqrt{5}}. \qquad \left(p \ge 2, \ A^{1/p} \ge 1, \ 1-\gamma \in (0,1] \right) && (98)
\end{aligned}$$

Combining Eqs. (94) and (98), we have,

$$\begin{aligned}
&\left| \tilde{V}^{\pi_{\tilde{\zeta}_{t+1}}}(\mu) - \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu) - \left\langle \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \tilde{\zeta}_{t+1} - \tilde{\theta}_t \right\rangle \right| \le \frac{4 \cdot p^2 \cdot A^{2/p} + c_\tau}{(1-\gamma)^3} \cdot \frac{\|\tilde{\zeta}_{t+1} - \tilde{\theta}_t\|_2^2}{\min_s \|\tilde{\theta}_{t,\xi}(s,\cdot)\|_p^2} \\
&\le \frac{4 \cdot p^2 \cdot A^{2/p} + c_\tau}{(1-\gamma)^3} \cdot \frac{5}{4} \cdot \|\tilde{\zeta}_{t+1} - \tilde{\theta}_t\|_2^2 \\
&= \frac{5 \cdot p^2 \cdot A^{2/p} + c_\tau}{(1-\gamma)^3} \cdot \|\tilde{\zeta}_{t+1} - \tilde{\theta}_t\|_2^2, \qquad (99)
\end{aligned}$$

which implies,

$$\begin{aligned}
\tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu) - \tilde{V}^{\pi_{\tilde{\theta}_{t+1}}}(\mu) &= \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu) - \tilde{V}^{\pi_{\tilde{\zeta}_{t+1}}}(\mu) \qquad \left(\tilde{\theta}_{t+1}(s,a) = \frac{\tilde{\zeta}_{t+1}(s,a)}{\|\tilde{\zeta}_{t+1}(s,\cdot)\|_p}, \text{ Algorithm 2} \right) \\
&\le - \Big\langle \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \tilde{\zeta}_{t+1} - \tilde{\theta}_t \Big\rangle + \frac{5 \cdot p^2 \cdot A^{2/p} + c_\tau}{(1-\gamma)^3} \cdot \|\tilde{\zeta}_{t+1} - \tilde{\theta}_t\|_2^2 \qquad \text{(by Eq. (99))} \\
&= -\eta \cdot \left\| \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t} \right\|_2^2 + \frac{5 \cdot p^2 \cdot A^{2/p} + c_\tau}{(1-\gamma)^3} \cdot \eta^2 \cdot \left\| \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t} \right\|_2^2 \qquad \left(\tilde{\zeta}_{t+1} = \tilde{\theta}_t + \eta \cdot \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t}, \text{ Algorithm 2} \right) \\
&= - \frac{(1-\gamma)^3}{20 \cdot p^2 \cdot A^{2/p} + c_\tau} \cdot \left\| \frac{\partial \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu)}{\partial \tilde{\theta}_t} \right\|_2^2 \qquad \left(\eta = \frac{(1-\gamma)^3}{10 \cdot p^2 \cdot A^{2/p} + c_\tau} \right) \\
&\le - \frac{(1-\gamma)^3}{20 \cdot \cancel{p^2} \cdot A^{2/p} + c_\tau} \cdot \frac{2\tau}{S} \cdot \min_s \mu(s) \cdot \frac{\cancel{p^2} \cdot \min_{s,a} \pi_{\tilde{\theta}_t}(a|s)^{2-2/p}}{\max_s \|\tilde{\theta}_t(s,\cdot)\|_p^2} \cdot \left\| \frac{d_\mu^{\pi_\tau^*}}{d_\mu^{\pi_{\tilde{\theta}_t}}} \right\|_\infty^{-1} \cdot \left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu) \right] \qquad \text{(Lemma 12)} \\
&= - \frac{(1-\gamma)^3 \cdot \tau}{(10 \cdot A^{2/p} + c_\tau) \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{d_\mu^{\pi_{\tilde{\theta}_t}}} \right\|_\infty^{-1} \cdot \min_{s,a} \pi_{\tilde{\theta}_t}(a|s)^{2-2/p} \cdot \left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu) \right] \qquad \left(\|\tilde{\theta}_t(s,\cdot)\|_p = 1, \text{ for all } s \right) \\
&\le - \frac{(1-\gamma)^4 \cdot \tau}{(10 \cdot A^{2/p} + c_\tau) \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \cdot \min_{s,a} \pi_{\tilde{\theta}_t}(a|s)^{2-2/p} \cdot \left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu) \right], \qquad \text{(by Eq. (73))}
\end{aligned} \tag{100}$$

which implies,

$$\begin{aligned}
\tilde{V}^{\pi_{\theta_t}}(\mu) - \tilde{V}^{\pi_{\theta_{t+1}}}(\mu) &= \tilde{V}^{\pi_{\tilde{\theta}_t}}(\mu) - \tilde{V}^{\pi_{\tilde{\theta}_{t+1}}}(\mu) \qquad \text{(by Lemma 9)} \\
&\le - \frac{(1-\gamma)^4 \cdot \tau}{(10 \cdot A^{2/p} + c_\tau) \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \cdot \min_{s,a} \pi_{\theta_t}(a|s)^{2-2/p} \cdot \left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu) \right], \qquad \text{(by Eq. (100) and Lemma 9)} \\
&\le - \frac{(1-\gamma)^4 \cdot \tau}{(10 \cdot A^{2/p} + c_\tau) \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \cdot c^{2-2/p} \cdot \left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu) \right],
\end{aligned}$$

which is equivalent to,

$$\begin{aligned}
\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu) &\le \left(1 - \frac{(1-\gamma)^4 \cdot \tau}{(10 \cdot A^{2/p} + c_\tau) \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \cdot c^{2-2/p} \right) \cdot \left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_{t-1}}}(\mu) \right] \\
&\le \exp\left\{ - \frac{(1-\gamma)^4 \cdot \tau}{(10 \cdot A^{2/p} + c_\tau) \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \cdot c^{2-2/p} \right\} \cdot \left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_{t-1}}}(\mu) \right] \\
&\le \exp\left\{ - \frac{(1-\gamma)^4 \cdot \tau}{(10 \cdot A^{2/p} + c_\tau) \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \cdot c^{2-2/p} \cdot (t-1) \right\} \cdot \left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_1}}(\mu) \right] \\
&\le \exp\left\{ - \frac{(1-\gamma)^4 \cdot \tau}{(10 \cdot A^{2/p} + c_\tau) \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \cdot c^{2-2/p} \cdot (t-1) \right\} \cdot \frac{1 + \tau \log A}{1-\gamma},
\end{aligned}$$

which leads to the final result,

$$\begin{aligned}
\tilde{V}^{\pi_\tau^*}(\rho) - \tilde{V}^{\pi_{\theta_t}}(\rho) &\le \frac{1}{1-\gamma} \cdot \left\| \frac{1}{\mu} \right\|_\infty \cdot \left[\tilde{V}^{\pi_\tau^*}(\mu) - \tilde{V}^{\pi_{\theta_t}}(\mu) \right] \qquad \text{(by Eq. (93))} \\
&\le \left\| \frac{1}{\mu} \right\|_\infty \cdot \exp\left\{ - \frac{(1-\gamma)^4 \cdot \tau}{(10 \cdot A^{2/p} + c_\tau) \cdot S} \cdot \left\| \frac{d_\mu^{\pi^*}}{\mu} \right\|_\infty^{-1} \cdot {c'}^2 \cdot (t-1) \right\} \cdot \frac{1 + \tau \log A}{(1-\gamma)^2},
\end{aligned}$$

where $c' = c^{1-1/p} \ge c = \inf_{(s,a)} \inf_t \pi_{\theta_t}(a|s) > 0$. □

Lemma 15 (Soft sub-optimality lemma [14])**.** *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. See the proof in [14, Lemma 26]. □

D Proofs for Section 5 (Escort Cross Entropy)

Lemma 16 (Non-uniform smoothness)**.** *Let $\pi_\theta \coloneqq f_p(\theta)$, and $\pi_{\theta'} \coloneqq f_p(\theta')$. Denote $\theta_\xi \coloneqq \theta + \xi \cdot (\theta' - \theta)$ with some $\xi \in [0,1]$. Then for $p = 2$, we have $D_{\mathrm{KL}}(y \| \pi_\theta)$ is β-smooth, i.e.,*

$$\left| D_{\mathrm{KL}}(y \| \pi_{\theta'}) - D_{\mathrm{KL}}(y \| \pi_\theta) - \left\langle \frac{d\{D_{\mathrm{KL}}(y \| \pi_\theta)\}}{d\theta}, \theta' - \theta \right\rangle \right| \le \frac{\beta}{2} \cdot \|\theta' - \theta\|_2^2,$$

with $\beta = \frac{6}{\|\theta_\xi\|_p^2} + 2 \cdot \left(\max_i \frac{y(i)}{\theta(i)^2} \right)$.

Proof. The gradient of $D_{\mathrm{KL}}(y \| \pi_\theta)$ w.r.t. θ is

$$\begin{aligned} \frac{d\{D_{\mathrm{KL}}(y \| \pi_\theta)\}}{d\theta} &= \frac{d\{-y^\top \log \pi_\theta\}}{d\theta} \\ &= \left(\frac{d\pi_\theta}{d\theta} \right)^\top \left(\frac{d\{-y^\top \log \pi_\theta\}}{d\pi_\theta} \right) \\ &= p \cdot \mathrm{diag}\left(\frac{1}{\theta} \right) \left(\mathrm{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top \right) \mathrm{diag}\left(\frac{1}{\pi_\theta} \right) (-y) \\ &= p \cdot \mathrm{diag}\left(\frac{1}{\theta} \right) (\pi_\theta - y). \end{aligned}$$

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

$$\begin{aligned} K(y, \theta) &= \frac{d}{d\theta} \left\{ \frac{d\{D_{\mathrm{KL}}(y \| \pi_\theta)\}}{d\theta} \right\} \\ &= p \cdot \frac{d}{d\theta} \left\{ \mathrm{diag}\left(\frac{1}{\theta} \right) (\pi_\theta - y) \right\}. \end{aligned}$$

We have $K(y, \theta) \in \mathbb{R}^{K \times K}$, whose element at position $(i, j) \in [K]^2$ is

$$\begin{aligned} K_{i,j} &= p \cdot \frac{d\{\frac{\pi_\theta(i) - y(i)}{\theta(i)}\}}{d\theta(j)} \\ &= p \cdot \frac{\frac{p}{\theta(j)} \cdot [\delta_{ij} \pi_\theta(j) - \pi_\theta(i) \pi_\theta(j)] \cdot \theta(i) - (\pi_\theta(i) - y(i)) \cdot \delta_{ij}}{\theta(i)^2} \\ &= p \cdot (p-1) \cdot \delta_{ij} \cdot \frac{\pi_\theta(i)}{\theta(i)^2} - p^2 \cdot \frac{\pi_\theta(i)}{\theta(i)} \cdot \frac{\pi_\theta(j)}{\theta(j)} + \delta_{ij} \cdot p \cdot \frac{y(i)}{\theta(i)^2}, \end{aligned}$$

where the δ notation is defined in Eq. (23). For any $x \in \mathbb{R}^K$,

$$\begin{aligned} \left| x^\top K(y, \theta) x \right| &= \left| \sum_{i=1}^K \sum_{j=1}^K K_{i,j} x(i) x(j) \right| \\ &= \left| p \cdot (p-1) \sum_i \frac{\pi_\theta(i)}{\theta(i)^2} \cdot x(i)^2 - p^2 \sum_i \frac{\pi_\theta(i)}{\theta(i)} \cdot x(i) \sum_j \frac{\pi_\theta(j)}{\theta(j)} \cdot x(j) + p \sum_i \frac{y(i)}{\theta(i)^2} \cdot x(i)^2 \right| \\ &\le p \cdot (p-1) \cdot \left[\sum_i \frac{\pi_\theta(i)}{\theta(i)^2} \cdot x(i)^2 \right] + p^2 \cdot \left[\sum_i \frac{\pi_\theta(i)}{\theta(i)} \cdot x(i) \right]^2 + p \sum_i \frac{y(i)}{\theta(i)^2} \cdot x(i)^2, \quad (101) \end{aligned}$$

where the last inequality is by triangle inequality. The first term is upper bounded as,

$$\begin{aligned}\sum_i \frac{\pi_\theta(i)}{\theta(i)^2} \cdot x(i)^2 &= \frac{1}{\|\theta\|_p^2} \sum_i \pi_\theta(i)^{1-2/p} \cdot x(i)^2 \\ &\le \frac{1}{\|\theta\|_p^2} \sum_i 1 \cdot x(i)^2 \qquad (p=2) \\ &= \frac{1}{\|\theta\|_p^2} \cdot \|x\|_2^2. \qquad (102)\end{aligned}$$

The second term is upper bounded as,

$$\begin{aligned}\left[\sum_i \frac{\pi_\theta(i)}{\theta(i)} \cdot x(i)\right]^2 &\le \sum_i \left(\frac{\pi_\theta(i)}{\theta(i)}\right)^2 \cdot \|x\|_2^2 \qquad \text{(by Cauchy-Schwarz)} \\ &= \frac{1}{\|\theta\|_p^2} \cdot \sum_i \left(\pi_\theta(i)^{1-1/p}\right)^2 \cdot \|x\|_2^2 \\ &\le \frac{1}{\|\theta\|_p^2} \cdot \left[\sum_i \pi_\theta(i)\right] \cdot \|x\|_2^2 \qquad (p=2) \\ &= \frac{1}{\|\theta\|_p^2} \cdot \|x\|_2^2. \qquad (103)\end{aligned}$$

The last term is upper bounded as,

$$\sum_i \frac{y(i)}{\theta(i)^2} \cdot x(i)^2 \le \left(\max_i \frac{y(i)}{\theta(i)^2}\right) \cdot \|x\|_2^2. \qquad (104)$$

Combining Eqs. (101) to (104), for $p = 2$, for any $x \in \mathbb{R}^K$, we have,

$$\begin{aligned}\left|x^\top K(y,\theta)x\right| &\le p \cdot (p-1) \cdot \frac{1}{\|\theta\|_p^2} \cdot \|x\|_2^2 + p^2 \cdot \frac{1}{\|\theta\|_p^2} \cdot \|x\|_2^2 + p \cdot \left\|\frac{y}{\theta \odot \theta}\right\|_\infty \cdot \|x\|_2^2 \\ &= \frac{6}{\|\theta\|_p^2} \cdot \|x\|_2^2 + 2 \cdot \left(\max_i \frac{y(i)}{\theta(i)^2}\right) \cdot \|x\|_2^2.\end{aligned}$$

According to Taylor's theorem, we have,

$$\begin{aligned}&\left|D_{\mathrm{KL}}(y\|\pi_{\theta'}) - D_{\mathrm{KL}}(y\|\pi_\theta) - \left\langle \frac{d\{D_{\mathrm{KL}}(y\|\pi_\theta)\}}{d\theta}, \theta' - \theta\right\rangle\right| = \frac{1}{2} \cdot \left|(\theta'-\theta)^\top K(y,\theta_\xi)\,(\theta'-\theta)\right| \\ &\quad \le \left[\frac{3}{\|\theta_\xi\|_p^2} + \max_i \frac{y(i)}{\theta_\xi(i)^2}\right] \cdot \|\theta' - \theta\|_2^2. \qquad \square\end{aligned}$$

Lemma 17 (Non-uniform Łojasiewicz). *Let $\pi_\theta = f_p(\theta)$. For any $p \ge 2$, we have,*

$$\left\|\frac{d\{D_{\mathrm{KL}}(y\|\pi_\theta)\}}{d\theta}\right\|_2^2 \ge \frac{p^2}{\|\theta\|_p^2} \cdot \min_a \pi_\theta(a)^{1-2/p} \cdot D_{\mathrm{KL}}(y\|\pi_\theta).$$

Proof. According to the definition of KL-divergence, we have,

$$\begin{aligned}
D_{\mathrm{KL}}(y\|\pi_\theta) &= \sum_a y(a)\cdot\log\left(\frac{y(a)}{\pi_\theta(a)}\right)\\
&\le \sum_a y(a)\cdot\left(\frac{y(a)}{\pi_\theta(a)}-1\right) \qquad (\log x \le x-1)\\
&= \sum_a \left(y(a)-\pi_\theta(a)+\pi_\theta(a)\right)\cdot\frac{y(a)-\pi_\theta(a)}{\pi_\theta(a)}\\
&= \sum_a \frac{\left(y(a)-\pi_\theta(a)\right)^2}{\pi_\theta(a)}\\
&= \sum_a \frac{\left(y(a)-\pi_\theta(a)\right)^2}{\pi_\theta(a)^{2/p}}\cdot\frac{1}{\pi_\theta(a)^{1-2/p}}\\
&= \sum_a \frac{\left(y(a)-\pi_\theta(a)\right)^2}{\theta(a)^2}\cdot\|\theta\|_p^2\cdot\frac{1}{\pi_\theta(a)^{1-2/p}} \qquad \left(\pi_\theta(a)=\frac{|\theta(a)|^p}{\sum_{a'}|\theta(a')|^p}\right)\\
&\le \|\theta\|_p^2\cdot\frac{1}{\min_a \pi_\theta(a)^{1-2/p}}\cdot\sum_a\frac{\left(y(a)-\pi_\theta(a)\right)^2}{\theta(a)^2}\\
&= \|\theta\|_p^2\cdot\frac{1}{\min_a \pi_\theta(a)^{1-2/p}}\cdot\frac{1}{p^2}\cdot\left\|p\cdot\mathrm{diag}\left(\frac{1}{\theta}\right)(y-\pi_\theta)\right\|_2^2.
\end{aligned}$$

The proof is completed with the observation that

$$\frac{d\{D_{\mathrm{KL}}(y\|\pi_\theta)\}}{d\theta} = p\cdot\mathrm{diag}\left(\frac{1}{\theta}\right)(\pi_\theta - y). \qquad \square$$

Theorem 5. Using the escort transform on the cross entropy objective, we have, for all $t\ge 1$,

- (gradient flow) for $p=2$, with $\eta_t = \frac{\|\theta_t\|_p^2}{p^2}$,
$$D_{\mathrm{KL}}(y\|\pi_{\theta_t}) \le D_{\mathrm{KL}}(y\|\pi_{\theta_1})\cdot e^{-(t-1)};$$
- (gradient descent) for $p=2$, with $\eta_t = \frac{\|\theta_t\|_p^2}{4\cdot(3+c_1^2)}$,
$$D_{\mathrm{KL}}(y\|\pi_{\theta_t}) \le D_{\mathrm{KL}}(y\|\pi_{\theta_1})\cdot\exp\left\{-\frac{(t-1)}{2\cdot(3+c_1^2)}\right\},$$

where $\frac{1}{c_1^2} = \frac{|\theta_1(a_y)|^2}{\|\theta_1\|_2^2} = \pi_{\theta_1}(a_y)\in(0,1]$ only depends on initialization.

Proof. **First part.** For the gradient flow, we have the following update,

$$\frac{d\theta_t}{dt} = -\eta_t\cdot\frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t}. \tag{105}$$

Then we have,

$$\begin{aligned}
\frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{dt} &= \left(\frac{d\theta_t}{dt}\right)^\top\left(\frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t}\right)\\
&= -\eta_t\cdot\left\|\frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t}\right\|_2^2 \qquad \text{(by Eq. (105))}\\
&\le -\eta_t\cdot\frac{p^2}{\|\theta_t\|_p^2}\cdot\min_a \pi_{\theta_t}(a)^{1-2/p}\cdot D_{\mathrm{KL}}(y\|\pi_{\theta_t}) \qquad \text{(by Lemma 17)}\\
&= -\min_a \pi_{\theta_t}(a)^{1-2/p}\cdot D_{\mathrm{KL}}(y\|\pi_{\theta_t}) \qquad \left(\eta_t = \frac{\|\theta_t\|_p^2}{p^2}\right)\\
&= -D_{\mathrm{KL}}(y\|\pi_{\theta_t}), \qquad (p=2)
\end{aligned}$$

which implies,

$$\frac{d\left\{\log D_{\mathrm{KL}}(y\|\pi_{\theta_t})\right\}}{dt} = \frac{1}{D_{\mathrm{KL}}(y\|\pi_{\theta_t})} \cdot \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{dt} \leq -1.$$

Taking integral, we have,

$$\log D_{\mathrm{KL}}(y\|\pi_{\theta_t}) - \log D_{\mathrm{KL}}(y\|\pi_{\theta_1}) \leq -(t-1),$$

which is equivalent to

$$D_{\mathrm{KL}}(y\|\pi_{\theta_t}) \leq D_{\mathrm{KL}}(y\|\pi_{\theta_1}) \cdot e^{-(t-1)}.$$

Second part. For the gradient descent, according to Lemma 16, we have,

$$D_{\mathrm{KL}}(y\|\pi_{\theta_{t+1}}) - D_{\mathrm{KL}}(y\|\pi_{\theta_t}) - \left\langle \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \leq \frac{\beta}{2} \cdot \|\theta_{t+1} - \theta_t\|_2^2, \tag{106}$$

where

$$\begin{aligned}\beta &= \frac{6}{\|\theta_{t,\xi}\|_p^2} + 2 \cdot \left(\max_i \frac{y(i)}{\theta_{t,\xi}(i)^2} \right) \\ &= \frac{6}{\|\theta_{t,\xi}\|_p^2} + \frac{2}{\theta_{t,\xi}(a_y)^2}, \qquad (y \text{ is one-hot}) \end{aligned} \tag{107}$$

and

$$\theta_{t,\xi} \coloneqq \theta_t + \xi \cdot (\theta_{t+1} - \theta_t) = \theta_t - \xi \cdot \eta_t \cdot \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t}. \tag{108}$$

The ℓ_p gradient norm is upper bounded as,

$$\begin{aligned}\left\| \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t} \right\|_p &= \left[\sum_a \left| \frac{\pi_{\theta_t}(a) - y(a)}{\theta_t(a)} \right|^p \right]^{\frac{1}{p}} \\ &= \left[\sum_{a \neq a_y} \left| \frac{\pi_{\theta_t}(a)}{\theta_t(a)} \right|^p + \left| \frac{\pi_{\theta_t}(a_y) - 1}{\theta_t(a_y)} \right|^p \right]^{\frac{1}{p}} \qquad (y(a_y) = 1) \\ &\leq \left[\sum_{a \neq a_y} \left| \frac{\pi_{\theta_t}(a)}{\theta_t(a)} \right|^p + \frac{1}{|\theta_t(a_y)|^p} \right]^{\frac{1}{p}} \qquad (\pi_{\theta_t}(a_y) \in (0, 1]) \\ &= \left[\frac{1}{\|\theta_t\|_p^p} \sum_{a \neq a_y} \pi_{\theta_t}(a)^{p-1} + \frac{1}{|\theta_t(a_y)|^p} \right]^{\frac{1}{p}} \\ &\leq \left[\frac{1}{\|\theta_t\|_p^p} + \frac{1}{|\theta_t(a_y)|^p} \right]^{\frac{1}{p}} \qquad (p = 2) \\ &\leq \frac{1}{\|\theta_t\|_p} + \frac{1}{|\theta_t(a_y)|} \qquad \left(\sqrt{x + y} \leq \sqrt{x} + \sqrt{y}\right) \end{aligned} \tag{109}$$

Next, we have,

$$\theta_{t+1}(a_y) = \theta_t(a_y) - \eta_t \cdot \frac{p}{\theta_t(a_y)} \cdot (\pi_{\theta_t}(a_y) - 1) \begin{cases} \geq \theta_t(a_y), & \text{if } \theta_t(a_y) > 0, \\ \leq \theta_t(a_y), & \text{if } \theta_t(a_y) < 0. \end{cases}$$

Therefore we have $|\theta_{t+1}(a_y)| \geq |\theta_t(a_y)|$. On the other hand, for all $a \neq a_y$, we have,

$$\theta_{t+1}(a) = \theta_t(a) - \eta_t \cdot \frac{p}{\theta_t(a)} \cdot \pi_{\theta_t}(a) \begin{cases} \leq \theta_t(a_y), & \text{if } \theta_t(a_y) > 0, \\ \geq \theta_t(a_y), & \text{if } \theta_t(a_y) < 0. \end{cases}$$

Therefore we have for all $a \neq a_y$, $|\theta_{t+1}(a)| \le |\theta_t(a)|$. Denote $\frac{1}{c_1} = \frac{|\theta_1(a_y)|}{\|\theta_1\|_p}$. We have, for all $t \ge 1$,

$$\begin{aligned}
\frac{|\theta_t(a_y)|}{\|\theta_t\|_p} = \frac{|\theta_t(a_y)|}{\left(\sum_a |\theta_t(a)|^p\right)^{1/p}} &\ge \frac{|\theta_t(a_y)|}{\left(\sum_{a \neq a_y} |\theta_1(a)|^p + |\theta_t(a_y)|^p\right)^{1/p}} \\
&\ge \frac{|\theta_1(a_y)|}{\left(\sum_{a \neq a_y} |\theta_1(a)|^p + |\theta_1(a_y)|^p\right)^{1/p}} \\
&= \frac{|\theta_1(a_y)|}{\|\theta_1\|_p} = \frac{1}{c_1}.
\end{aligned} \tag{110}$$

Combining Eqs. (109) and (110), we have,

$$\left\| \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t} \right\|_p \le \frac{1}{\|\theta_t\|_p} + \frac{1}{|\theta_t(a_y)|} \le \frac{1}{\|\theta_t\|_p} \cdot (1 + c_1). \tag{111}$$

Then we have,

$$\begin{aligned}
\|\theta_{t,\xi}\|_p &= \left\| \theta_t - \xi \cdot \eta_t \cdot \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t} \right\|_p \qquad \text{(by Eq. (108))} \\
&\ge \|\theta_t\|_p - \xi \cdot \eta_t \cdot \left\| \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t} \right\|_p \qquad \text{(by triangle inequality)} \\
&\ge \|\theta_t\|_p - \xi \cdot \eta_t \cdot \frac{1}{\|\theta_t\|_p} \cdot (1 + c_1). \qquad \text{(by Eq. (111))} \\
&= \|\theta_t\|_p \cdot \left[1 - \xi \cdot \frac{1 + c_1}{4 \cdot (3 + c_1^2)} \right] \qquad \left(\eta_t = \frac{\|\theta_t\|_p^2}{4 \cdot (3 + c_1^2)} \right) \\
&\ge \|\theta_t\|_p \cdot \left[1 - \frac{1 + c_1}{4 \cdot (3 + c_1^2)} \right] \qquad (\xi \in [0, 1]) \\
&= \|\theta_t\|_p \cdot \left[\left(1 - \frac{1}{\sqrt{2}} \right) \cdot \left(1 - \frac{\sqrt{2} + 1}{2\sqrt{2}} \cdot \frac{1 + c_1}{3 + c_1^2} \right) + \frac{1}{\sqrt{2}} \right] \\
&\ge \frac{\|\theta_t\|_p}{\sqrt{2}} \qquad (1/c_1 \in (0, 1],\ c_1 \ge 1)
\end{aligned} \tag{112}$$

Similar to Eq. (110), we have,

$$\begin{aligned}
\beta &= \frac{6}{\|\theta_{t,\xi}\|_p^2} + \frac{2}{\theta_{t,\xi}(a_y)^2} \qquad \text{(by Eq. (107))} \\
&\le \frac{1}{\|\theta_{t,\xi}\|_p^2} \cdot \left(6 + 2 \cdot c_1^2 \right).
\end{aligned} \tag{113}$$

Combining the results, we have,

$$\begin{aligned}
&D_{\mathrm{KL}}(y\|\pi_{\theta_{t+1}}) - D_{\mathrm{KL}}(y\|\pi_{\theta_t}) - \left\langle \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \\
&\quad \le \frac{1}{2} \cdot \frac{1}{\|\theta_{t,\xi}\|_p^2} \cdot \left(6 + 2 \cdot c_1^2 \right) \cdot \|\theta_{t+1} - \theta_t\|_2^2 \qquad \text{(by Eqs. (106) and (113))} \\
&\quad \le \frac{2}{\|\theta_t\|_p^2} \cdot \left(3 + c_1^2 \right) \cdot \|\theta_{t+1} - \theta_t\|_2^2, \qquad \text{(by Eq. (112))}
\end{aligned}$$

which implies (using the update $\theta_{t+1} = \theta_t - \eta_t \cdot \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t}$),

$$
\begin{aligned}
D_{\mathrm{KL}}(y\|\pi_{\theta_{t+1}}) - D_{\mathrm{KL}}(y\|\pi_{\theta_t}) &\le -\eta_t \cdot \left\| \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t} \right\|_2^2 + \frac{2\cdot\left(3+c_1^2\right)}{\|\theta_t\|_p^2} \cdot \eta_t^2 \cdot \left\| \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t} \right\|_2^2 \\
&= -\frac{\|\theta_t\|_p^2}{8\cdot(3+c_1^2)} \cdot \left\| \frac{d\{D_{\mathrm{KL}}(y\|\pi_{\theta_t})\}}{d\theta_t} \right\|_2^2 \qquad \left(\eta_t = \frac{\|\theta_t\|_p^2}{4\cdot(3+c_1^2)} \right) \\
&\le -\frac{\cancel{\|\theta_t\|_p^2}}{8\cdot(3+c_1^2)} \cdot \frac{p^2}{\cancel{\|\theta_t\|_p^2}} \cdot \min_a \pi_{\theta_t}(a)^{1-2/p} \cdot D_{\mathrm{KL}}(y\|\pi_{\theta_t}) \qquad \text{(by Lemma 17)} \\
&= -\frac{1}{2\cdot(3+c_1^2)} \cdot D_{\mathrm{KL}}(y\|\pi_{\theta_t}), \qquad (p=2)
\end{aligned}
$$

which is equivalent to,

$$
\begin{aligned}
D_{\mathrm{KL}}(y\|\pi_{\theta_t}) &\le \left[1 - \frac{1}{2\cdot(3+c_1^2)} \right] \cdot D_{\mathrm{KL}}(y\|\pi_{\theta_{t-1}}) \\
&\le D_{\mathrm{KL}}(y\|\pi_{\theta_{t-1}}) \cdot \exp\left\{ -\frac{1}{2\cdot(3+c_1^2)} \right\} \\
&\le D_{\mathrm{KL}}(y\|\pi_{\theta_1}) \cdot \exp\left\{ -\frac{(t-1)}{2\cdot(3+c_1^2)} \right\},
\end{aligned}
$$

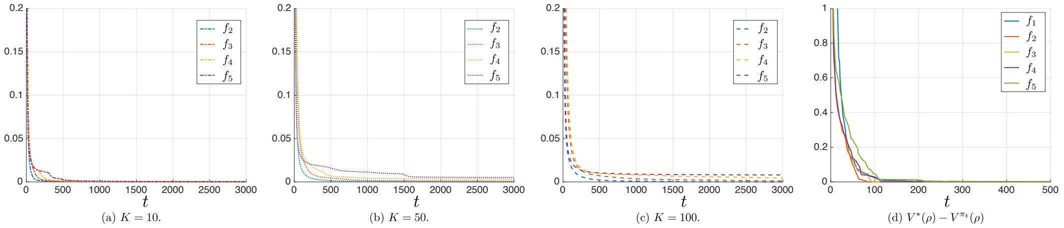
where $\frac{1}{c_1^2} = \frac{|\theta_1(a_y)|^2}{\|\theta_1\|_2^2} = \pi_{\theta_1}(a_y) \in (0,1]$. □

E Experimental Details and Additional Experiments

For the one-state MDPs, for each value of $K \in \{10, 50, 100\}$, the policy is parameterized by $\theta \in \mathbb{R}^K$. For SPG, $\pi_\theta = \mathrm{softmax}(\theta)$, and for EPG $\pi_\theta = f_p(\theta)$. The total number of runs for each algorithm under each K value is 20. In each run, we randomly generate the reward $r \in [0,1]^K$, and then randomly initialize π_{θ_1} within the $(K-1)$-dimensional probability simplex. SPG and EPG start from the same initial policy π_{θ_1}. The total number of iterations is $T = 5 \times 10^4$. Fig. 5 shows the results of SPG and EPG with $p = 2$. The learning rate of SPG is set to be $\eta = 0.4$ [14]. The learning rate of EPG is $\eta_t = 0.2 \cdot \|\theta_t\|_p^2$ (Theorem 2).

Additional experiments Fig. 7(a)-(c) show the results of EPG for $p \in \{2, 3, 4, 5\}$ in one-state MDPs, where each curve is the averaged result of 20 runs.

For the Four-room environment, the policy is $\pi_\theta = \mathrm{softmax}(\theta)$ for SPG, and $\pi_\theta = f_p(\theta)$ for EPG, and θ is the output of one hidden layer neural network with ReLU activation function and 64 hidden nodes. Fig. 5 shows the results of SPG and EPG with $p = 2$. The optimal value function V^* is approximately calculated using value iteration with threshold of two consecutive iterations $\|V_t - V_{t+1}\|_2^2 \le 1 \times 10^{-10}$. In each iteration, the true objective is used by calculating the stationary distribution $d^{\pi_{\theta_t}}$ and the state-action value $Q^{\pi_{\theta_t}}$. We use Adam optimizer [10] and the total number of iterations is 500. The total number of runs for each algorithm is 20. The p value for EPG is searched within $\{1, 2, 3, 4, 5\}$. The learning rates for SPG and EPG are searched within $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$ and 0.01 is used for both SPG and EPG.



(a) $K = 10$.

(b) $K = 50$.

(c) $K = 100$.

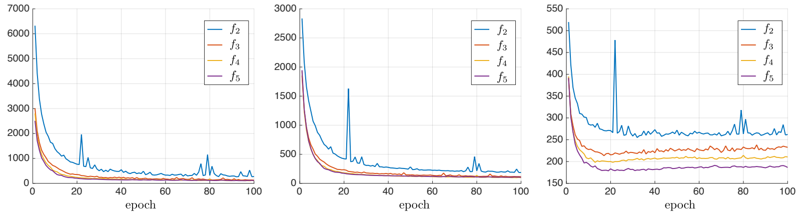
(d) $V^*(\rho) - V^{\pi_t}(\rho)$

Figure 7: *Results of EPG with different* p *values on one-state MDPs and Four-room.*

Additional experiments Fig. 7(d) shows the results of EPG with $p \in \{1, 2, 3, 4, 5\}$ on Four-room environment, where each curve is the averaged result of 20 runs.

For the MNIST dataset, the policy is $\pi_\theta = \mathrm{softmax}(\theta)$ for SPG and SCE, and $\pi_\theta = f_p(\theta)$ for EPG and ECE, where θ is the output of one hidden layer neural network with ReLU activation function and 512 hidden nodes. We use SGD with momentum 0.9 and the total number of epochs is 100. The total number of runs for each algorithm is 20. Fig. 6(a) and (b) show the results of SPG and EPG with $p = 4$. The learning rates for SPG and EPG are searched within $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$ and 0.05 is used for both SPG and EPG. The batchsize is searched within $\{10, 20, 50, 100, 200, 500\}$, and 20 is used for for SPG and 50 is used for EPG. The p value for EPG is searched within $\{1, 2, 3, 4, 5\}$. Fig. 6(c) and (d) show the results of SCE and ECE with $p = 2$. The learning rate and batchsized are searched within the same range as above, and we use the learning rate 0.01, and the batchsize 20 for both SCE and ECE.

Additional experiments Fig. 8 shows the results of EPG with $p \in \{2, 3, 4, 5\}$ on MNIST, where each curve is the averaged result of 10 runs. The best result in terms of the test error is with $p = 5$.

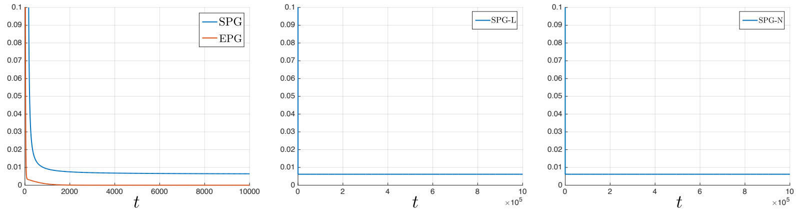


(a) Training objective.
(b) Training error.
(c) Test error.

Figure 8: *Results of EPG with different p values on MNIST.*

E.1 Failure of SPG Heuristics

According to Theorem 1, the SGW results from the SPG progresses being upper bounded by the small probabilities of the optimal action. Therefore, one may wonder if some simple heuristics would fix the SGW problem. We consider the following two natural heuristics.



(a) SPG and EPG with $p = 2$.
(b) SPG-L: SPG with large $\eta_t = 0.4/\pi_{\theta_t}(a^*)$.
(c) SPG-N: SPG with ℓ_2 normalized gradient.

Figure 9: *Sub-optimality results* $(\pi^* - \pi_{\theta_t})^\top r$ *on a single-state MDP with* $K = 50$.

SPG-L: Large learning rate Since the progress in each iteration of SPG is upper bounded by the order of $\pi_{\theta_t}(a^*)$, we could possibly use a large learning rate such as $\eta_t = 0.4/\pi_{\theta_t}(a^*)$ in SPG to compensate the slow progress around SGWs.

We run the SPG-L update on a constructed bandit problem with $K = 50$. Unfortunately, as shown in Fig. 9(b), after about 1×10^6 iterations, SPG-L still gets stuck on a sub-optimal plateau.

The intuitive explanation for the failure of SPG-L is that learning rate should be small enough to guarantee monotonic improvement [1, 14] for smooth functions, which means SPG with large (unbounded) learning rate is not even guaranteed to be asymptotically convergent.

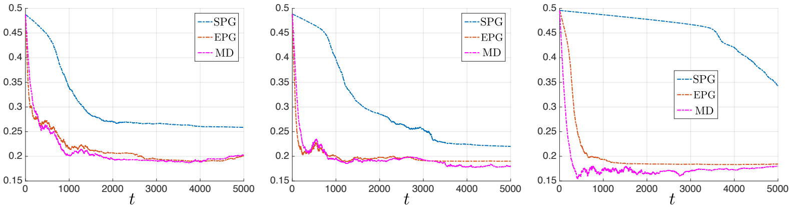
SPG-N: Normalized policy gradient Another heuristic is to update using normalized policy gradient. The intuition is that since the progress is upper bounded by the original SPG norm, which has small scale, we could hope the normalized gradient would provide nicer progress. In particular, we normalize the policy gradient by dividing its ℓ_2 norm, and then do update with $\eta = 0.4$.

We run the SPG-N update on the same problem as shown in Fig. 9(c), and it also failed. We do not have rigorous explanations for the failure of SPG-N. A speculation is that after ℓ_2 normalization, the normalized policy gradient still has small scale, since it is not necessary that the policy gradient norm is on the same scale as π_{θ_t}, which is the quantity that needs to be canceled.

EPG As shown in Fig. 9(a), EPG with $p = 2$ does not suffer from the plateau and works well.

E.2 Comparing SPG, EPG, and MD

As noted in Remark 3, EPG cannot be reduced to MD with any regularizer. Also as shown in Fig. 3(b), EPG and MD with KL regularization behave similarly in the 3-action case. We conduct experiments on bandit problems with $K = 50, 100, 500$ actions to compare EPG with MD. In each iteration, all the algorithms use the same stochastic gradient to do updates. Each curve is averaged over 20 runs.



(a) $K = 50$. (b) $K = 100$. (c) $K = 500$.

Figure 10: *Sub-optimality results* $(\pi^* - \pi_{\theta_t})^\top r$ *on single-state MDPs using stochastic gradients.*

As shown in Fig. 10, EPG and MD with KL regularization have comparable performances, significantly outperforming SPG. However, EPG in its nature is a policy gradient method, which has cheap update in each iteration, while MD needs to solve an optimization problem to do one update.